Acercamiento cuantitativo y cualitativo a la teoría de la información y estimadores de la entropía

Proyecto de Grado

Laura Marcela Guzmán Rincón
Directores: Maricarmen Martinez
Adolfo Quiroz

Departamento de Matemáticas
Facultad de Ciencias
Universidad de Los Andes

Mayo 2013

Acercamiento cuantitativo y cualitativo a la teoría de la información y estimadores de la entropía

Proyecto de grado para optar al título de Matemáticas

Dirigida por: Adolfo Quiroz y Maricarmen Martinez

Departamento de Matemáticas
Facultad de Ciencias
Universidad de Los Andes

Mayo 2013

Agradecimientos

De corazón, agradezco a mis asesores Maricarmen Martínez y Adolfo Quiroz por asesorar este proyecto y ser parte sustancial de su inicio, desarrollo y culminación. Les agradezco por su comprensión, paciencia, disposición. Agradezco a la Universidad de Los Andes por permitirme acceder a la educación superior de calidad y brindarme los conocimientos necesarios para realizar mis proyectos y, a futuro, aportar al país. Agradezco a mi profesor Alonso Botero por ayudarme a resolver algunos interrogantes que surgieron en el desarrollo de este proyecto. Finalmente y no con menos importancia, agradezco a mi mamá por todo el apoyo incondicional que me ha brindado en toda la vida. Agradezco a mis hermanitos y a mi familia por estar presentes siempre.

Resumen

La teoría de la información se ha desarrollado desde perspectivas cuantitativas y cualitativas. Se presenta una descripción a los conceptos cuantitativos basados en la teoría de la probabilidad, estudiando el concepto de *entropía de Shannon* introducido por C. Shannon, demostrando teoremas asociados a dicha cantidad y analizando el comportamiento de su estimador cuando se desconoce la distribución asociada a una fuente. En el caso cualitativo, se introducen nuevos conceptos basados en el marco presentado por J. Barwise y J. Seligman, analizando el comportamiento de la lógica asociada a una clasificación basado en las técnicas de estimadores usadas en el estudio de la entropía. De igual forma se busca una conexión entre los marcos cualitativos y cuantitativos a partir del estudio de la entropía en ambos casos.

Índice

Índice de figuras

Capítulo 1

Introducción

La *Teoría de la Información* fue iniciada por Claude E. Shannon durante la segunda mitad del Siglo XX, con el objetivo de estudiar la información y lo concerniente a su almacenamiento y procesamiento. La formulación inicia con la definición del concepto de *Entropía*, a partir del cual se obtienen los primeros resultados y se derivan nuevas nociones. La descripción de la información proveniente de una fuente, el envío de datos a través de canales y la posibilidad de comprimir son algunos de los avances que la teoría de la información ha logrado durante su desarrollo.
Como base de la teoría, la entropía es introducida para medir tanto la incertidumbre de la información proveniente de una fuente descrita por una variable aleatoria, como el valor esperado de la información contenida en un mensaje. Sin embargo, bajo esa descripción la entropía carece de una interpretación operacional. Más aún no está definida en el caso en el que se desconoce la distribución que modela la fuente. El objetivo de este trabajo es, en primer lugar, presentar el *Teorema de la Fuente de Shannon*, que brinda una interpretación operacional a la entropía y caracteriza los mensajes provenientes de una fuente. Para ello se introduce previamente la *Ley de los Grandes Números* de la que se deduce el teorema. Adicionalmente, se define un *estimador* de la entropía que permite estimar dicha cantidad al desconocer la distribución de probabilidad que caracteriza a la fuente. En segundo lugar, se construye un análisis cualitativo de los conceptos desarrollados anteriormente, basado en los conceptos de clasificación probabilística y lógicas.

1.1. Descripción del Documento

En la primera parte del documento, se desarrollará una descripción cuantitativa de una fuente de información y de la entropía de Shannon. En el capítulo 2 se definirán los conceptos básicos de Espacios de Probabilidad y Fuente de Información necesarios para enunciar y demostrar los teoremas

posteriores. En el capítulo 3 se enunciará y demostrará la Ley Fuerte de los Grandes Números directamente. En el capítulo 4 se enunciará y demostrará el Teorema de la Codificación de la Fuente de Shannon. Posteriormente se definirá un estimador para la entropía y se calculará el error de dicho estimador respecto a la cantidad real. En la segunda parte del documento se realizará un análisis cualitativo de la entropía de información. Se definirá la noción de clasificación probabilística y lógica. Se hará un análisis de cómo estimar la lógica absoluta y la lógica probabilística haciendo uso de teoremas desarrollados previamente. Finalmente se harán las conclusiones respectivas.

1.2. Notación

A lo largo del documento se va a manejar la siguiente notación. Sea $A \subseteq \Omega$ un conjunto:

- $A^c = \{\omega \in \Omega : \omega \notin A\}$ es el complemento de A.
- $|A|$ es el cardinal de A.
- Sean $a_n \in A$. $(a_n)_{n\in\mathbb{N}} \subseteq A$ es una sucesión de elementos de A.

Si v es un vector en $\mathbb{R}^n$, v^t es la transpuesta de v.

La función $\log(x)$ es el logaritmo en base 2 de x.

Si x es un número real, $[x]$ es la parte entera superior.

La flecha $\xrightarrow{p}$ quiere decir convergencia en probabilidad, como se definirá en el siguiente capítulo. La p indica probabilidad y no debe confundirse con otra p que, en el mismo contexto, esté siendo usada con otro significado.

Capítulo 2

Preliminares

En este capítulo se introducirán los conceptos básicos que serán utilizados en el documento. Inicialmente se abordarán los espacios de medida, y los espacios de probabilidad y sus definiciones asociadas. Luego se definirá y describirá formalmente la noción de fuente de información.

2.1. Espacios de Medida

Las siguientes definiciones introducen los conceptos de espacio de medida y medida, fundamentales para las ideas desarrolladas en el capítulo.

Definición 2.1. Sea Ω un conjunto. El conjunto $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ es una *σ-álgebra* en Ω si:

- $\Omega \in \mathcal{A}$.
- Si $A \in \mathcal{A}$ entonces $A^c \in \mathcal{A}$.
- Si $(A_n)_{n\in\mathbb{N}} \in \mathcal{A}$, entonces $\bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$.

En ese caso, la tupla $(\Omega, \mathcal{A})$ se llama *espacio de medida* y $A \in \mathcal{A}$ *evento*. Si $\mu : \mathcal{A} \to [0, +\infty]$ es una función tal que $\mu(\emptyset) = 0$ y, para toda secuencia $(A_n)_{n\in\mathbb{N}}$ de conjuntos disjuntos de $\mathcal{A}$, $\mu(\cup_{n\in\mathbb{N}} A_n) = \sum_{n\in\mathbb{N}} \mu(A_n)$, entonces μ es una *medida* en $\mathcal{A}$.

Sean $(\Omega, \mathcal{A})$ y $(\Psi, \mathcal{B})$ espacios de probabilidad. Una función $f : \Omega \to \Psi$ se dice medible si para todo $B \in \mathcal{B}$, $f^{-1}(B) \in \mathcal{A}$.

2.2. Espacios de Probabilidad

Considere la situación en la que una fuente emite ciertos símbolos aleatoriamente. Para cuantificar las propiedades y predecir hechos asociados a

dicha fuente se puede asignar a cada símbolo una probabilidad de ocurrencia. La anterior situación motiva a la teoría de la probabilidad a ofrecer un formalismo para los problemas de la teoría de la información, basados en el concepto de *espacio de probabilidad*.

Definición 2.2. Sea Ω un conjunto, $\mathcal{A}$ una σ-álgebra de subconjuntos de Ω, y $P : \mathcal{A} \to [0, +\infty]$ una medida sobre $\mathcal{A}$ tal que $P(\Omega) = 1$.

- La tupla $(\Omega, \mathcal{A})$ es un *espacio medible*.
- La tupla $(\Omega, \mathcal{A}, P)$ es un *espacio de probabilidad*.
- La función P es una *medida de probabilidad*.
- Los elementos de $\mathcal{A}$ se llaman *conjuntos medibles* o *eventos*.

Nota 2.1. Si $a \in A$, la probabilidad $\mathcal{P}(\{a\})$ de dicho elemento se puede escribir como $\mathcal{P}(a)$ si su significado es claro en el contexto. Igualmente, $g(\{a\}) = g(a)$.

En un espacio de probabilidad $(\Omega, \mathcal{A}, P)$, si se realiza un experimento cuyas observaciones provienen de Ω, se puede asociar a cada resultado un número real o una cantidad más general. Esa asociación da lugar al concepto de *variable aleatoria*.

Definición 2.3. Sean $(\Omega, \mathcal{A}, P)$ un espacio de probabilidad, $(\mathcal{S}, \mathcal{B})$ un espacio de medida y $X, Y : \Omega \to \mathcal{S}$. Entonces,

- Si X es medible, X es una *variable aleatoria*.
- Sea $\mathcal{L}(X) : \mathcal{B} \to [0, +\infty]$ tal que $\mathcal{L}(X)(B) = (P \circ X^{-1})(B)$, para todo $B \in \mathcal{B}$. La función $\mathcal{L}(X)$ se denomina *ley de* X.

Si $S = \mathbb{R}$ y $\mathcal{B}$ es la σ-álgebra de Borel, entonces:

- El *valor esperado* de X se define como $EX = \int X \ dP$, si la integral existe.
- La *varianza* de X es

$$var(X) = \begin{cases} E((X - EX)^2), & \text{si } EX^2 < \infty \\ \infty, & \text{si } EX^2 = \infty. \end{cases}$$

- La *covarianza* de X viene dada por $cov(X, Y) = E((X - EX)(Y - EY))$.

El siguiente ejemplo reúne las ideas introducidas previamente, con elementos que serán utilizados posteriormente.

Ejemplo 2.1.

- La tupla $(\mathbb{R}, \mathcal{B})$, donde $\mathcal{B}$ es la σ-álgebra de Borel, es un espacio medible.
- La tupla $(\Omega, \mathcal{A}, P)$, en donde $\Omega = \{a_1, a_2\}$, $\mathcal{A} = \mathcal{P}(\Omega)$, $P(\emptyset) = 0$ y

$$P(\{\omega\}) = \begin{cases} p, & \text{si } \omega = a_1 \\ 1-p, & \text{si } \omega = a_2, \end{cases}$$

  con $0 \leq p \leq 1$, es un espacio de probabilidad.
- Con $(\mathbb{R}, \mathcal{B})$ y $(\Omega, \mathcal{A}, P)$ como antes, la función $X : \Omega \to \mathbb{R}$, tal que $X(\omega) = -\log(P(\omega))$ para $\omega \in \Omega$, es una variable aleatoria. Si $p = 1/2$, $X(\omega) = -\log(1/2) = 1$ para todo $\omega \in \Omega$. Además, en ese caso,

$$\mathcal{L}(B) = \begin{cases} 1, & \text{si } 1 \in B \\ 0, & \text{si } 1 \notin B. \end{cases}$$

La siguiente definición introduce las nociones de convergencia requeridas posteriormente.

Definición 2.4. Sean $(\Omega, \mathcal{A}, \mathcal{P})$ un espacio de probabilidad y $X_1, X_2, ...$ variables aleatorias.

- X_n converge *casi seguramente (c.s.)* a X si

$$\mathcal{P}(\{\omega \in \Omega : X_n(\omega) \to X(\omega)\}) = 1.$$

- X_n converge *en probabilidad* a X, $X_n \xrightarrow{p} X$, si, para todo $\epsilon > 0$,

$$\lim_{n\to\infty} \mathcal{P}(|X_n - X| > \epsilon) = 0.$$

- X_n converge *en distribución* a X, $X_n \xrightarrow{d} X$, si

$$\lim_{n\to\infty} \mathcal{P}(X_n) = \mathcal{P}(X).$$

Nota 2.2. De la definición se sigue que convergencia casi seguramente implica convergencia en probabilidad.

A las observaciones tomadas en un experimento se le pueden asociar distintas variables aleatorias. La relación que hay entre las cantidades cuantificadas por dichas variables está codificada en la noción de *independencia*.

Definición 2.5. Sea $(\Omega, \mathcal{A}, \mathcal{P})$ un espacio de probabilidad, $(\mathcal{S}, \mathcal{B})$ un espacio de medida. Sean $A_1, ..., A_n \in \mathcal{A}$ y sean $X_1, ..., X_n$ tales que X_i es una variable aleatoria de Ω en $\mathcal{S}$.

- Se dice que $A_1, ..., A_n$ son *independientes* si, para todo $i, j \in \{1, ..., n\}$

$$\mathcal{P}(A_i, A_j) = \mathcal{P}(A_i)\mathcal{P}(A_j).$$

- Se dice que $X_1, ..., X_n$ son *independientes* si

$$\mathcal{L}(\langle X_1, ..., X_n \rangle) = \mathcal{L}(X_1) \times ... \times \mathcal{L}(X_n).$$

- Se dice que $X_1, ..., X_n$ son *i.i.d.* si son independientes y $\mathcal{L}(X_i) = \mathcal{L}(X_1)$ para todo i.

Las definiciones incluidas en esta sección serán utilizadas a lo largo del documento, principalmente en el capítulo 3, para demostrar la Ley de Grandes Números.

2.3. Fuente de Información

En esta sección se formaliza la idea de fuente de información y algunos conceptos asociados a ésta, incluyendo la *Entropía de Shannon*.
Informalmente, una fuente emite una cadena de símbolos, cada uno con una probabilidad de ocurrencia asociada. Formalmente se presenta la siguiente definición.

Definición 2.6. Una fuente $\mathbf{A}$ es un espacio de probabilidad $(A, \mathcal{P}(A), P)$ donde el *alfabeto* A es un conjunto finito, y los elementos a que se observan provienen de A, cada uno con una probabilidad $P(\{a\})$.

- El *tamaño de información de* $\boldsymbol{A}$ es la cantidad mínima de bits que se requieren para codificar las observaciones provenientes de la fuente, sin pérdida de información, y viene dado por:

$$H_o(\mathbf{A}) := \log(|A|). \tag{2.1}$$

- Para $B \subseteq A$, $P(B) := \sum_{b \in B} P(\{b\})$ es la probabilidad de observar un elemento de B.

- $G(\{b\}) := -\log(P(\{b\}))$ es la *ganancia de información de* $b \in B$, es decir, la ganancia obtenida al observar b. Si $\mathcal{P}(\{b\}) = 0$, $G(\{b\}) := 0$.

En el ejemplo 2.1, la tupla $(\Omega, \mathcal{A}, P)$ puede representar una fuente $\mathbf{A}$ con un alfabeto de tamaño 2 y una distribución de probabilidad binomial. En ese caso la fuente requiere $H_o(\mathbf{A}) = 1$ bits para codificar la información observada.

Si $\mathbf{A}$ es una fuente, al observar un $a \in A$ se obtiene una ganancia de $g(a)$. Se puede calcular cuál es la ganancia de información media de la fuente, como

se introduce en la siguiente definición. Esta cantidad, conocida como *entropía de Shannon* fue introducida en 1948 por Claude Shannon en su artículo *A Mathematical Theory of Communication* [1].

Definición 2.7 (**Entropía de Shannon**)**.** Sea **A** una fuente. La *entropía de Shannon de* ***A*** es el valor esperado de la ganancia de información proveniente de la fuente, y viene dada por:

$$H(\mathbf{A}) := -\sum_{a \in A} P(a) \log(P(a)). \tag{2.2}$$

Si $P(a) = 0$, $P(a) \log P(a) := 0$.

La entropía de Shannon se puede interpretar como la incertidumbre asociada a **A**. Por ejemplo, si $\{a\}$ es un evento muy probable, su ganancia es muy baja y, así mismo, su incertidumbre también. Se considera que la entropía mide el nivel medio de sorpresa de los eventos generados por la fuente.

Otra interpretación de la entropía de una fuente está asociada con la noción de compresión, como se mostrará en el capítulo 4. En ese contexto, analizar la compresión aceptando un posible error δ requiere el siguiente concepto que será estudiado en ese mismo capítulo.

Definición 2.8. Sea **A** una fuente y $\delta > 0$. El *tamaño de información* con un error δ está dado por:

$$H_\delta(\mathbf{A}) := \inf_{\substack{B \subseteq A \\ P(B) \geq 1-\delta}} \log |B|. \tag{2.3}$$

Ejemplo 2.2. Sea **A** la fuente asociada a la tupla introducida en el ejemplo 2.1. La entropía asociada a ésta es:

$$H(\mathbf{A}) = -p \log p - (1-p) \log(1-p). \tag{2.4}$$

En este caso, $H(\mathbf{A})$ se conoce como *entropía binaria* y se denota como $H_2(p)$. En algunos ejemplos será de interés conocer el comportamiento de esta entro-

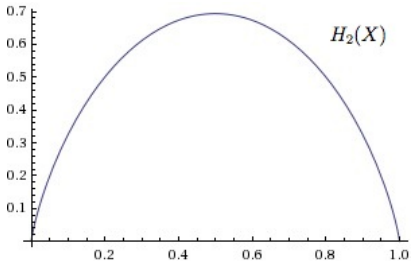

Figura 2.1: Entropía binaria en función de p.

pía respecto a p. $H_2(p)$ es máximo cuando $p = 1/2$, es decir, la incertidumbre se maximiza cuando a_1 y a_2 son igualmente probables. Por otro lado, la entropía se minimiza cuando $p(a_1) = 0$ o $p(a_2) = 0$, es decir, cuando hay un elemento de A que se observará con probabilidad 1. En la figura 2.1 se muestra el anterior resultado. Ahora suponga que $p = 1/4$ y tome $\delta = 1/2$. Los únicos subconjuntos B de A para los que $P(B) \geq 1 - \delta = 1/2$ son A y $\{a_2\}$. Como $\log|A| = 1 \geq \log|\{a_2\}| = 0$ entonces $H_\delta(\mathbf{A}) = 0$.

En el capítulo 4.1 se dará una nueva interpretación a la entropía de Shannon que explicará por qué el valor máximo de la entropía coincide con el tamaño de información de $\mathbf{A}$.

Capítulo 3

Ley de Grandes Números

En este capítulo se enunciará y demostrará la *ley fuerte de los grandes números* para el caso i.i.d, presentada en la sección 3.2. La primera sección presenta algunos lemas necesarios en el resto del capítulo. Las demostraciones son tomadas de [2].

3.1. Lemas auxiliares

Para el contenido de esta sección, sea $(\Omega, \mathcal{S}, \mathcal{P})$ un espacio de probabilidad.

Lema 3.1. *Sea $(p_n)_n$ una sucesión de reales tales que $0 \leq p_n < 1$ para todo n. Entonces $\prod_{n\in\mathbb{N}}(1-p_n) = 0$ si y sólo si $\sum_{n\in\mathbb{N}} p_n = \infty$.*

Demostración.

Caso 1: Suponga que $\limsup p_n > 0$. Luego $\sum_{n\in\mathbb{N}} p_n = \infty$, que implica que $\prod_{n\in\mathbb{N}}(1-p_n) = 0$ como se quería.

Caso 2: Suponga que $\limsup p_n = 0$. Como $p_n \geq 0$, $\lim_{n\to\infty} p_n = 0$. Para el caso $\Leftarrow$, suponga que $\sum_{n\in\mathbb{N}} = \infty$. Como $0 \leq p_n \leq 1$, $1 - p \leq e^{-p}$. Entonces $\prod_{n\in\mathbb{N}}(1-p_n) \leq \prod_{n\in\mathbb{N}} e^{-p} = e^{-\sum_{n\in\mathbb{N}} p_n}$. Luego, $\prod_{n\in\mathbb{N}}(1-p_n) = 0$. Para el caso $\Rightarrow$, suponga que $\prod_{n\in\mathbb{N}}(1-p_n) = 0$. Sabemos que existe $M \in \mathbb{N}$ tal que $p_M < 1/2$ para todo $n \geq M$. Como $1 - p_n \geq e^{-2p_n}$ si $0 \leq p_n \leq 1/2$, $\prod_{n\geq M}(1-p_n) \geq \prod_{n\geq M} e^{-2p_n} = e^{-2\sum_{n\geq M}} \geq 0$. Como $\prod_{n\in\mathbb{N}}(1-p_n) = 0$, $\prod_{n\geq M}(1-p_n)$. Luego, $-2\sum_{n\geq M} p_n = \infty$, obteniendo $\sum_{n\in\mathbb{N}} p_n = \infty$ como se quería. □

Definición 3.1. Si $(A_n) \subseteq \mathcal{S}$ es una sucesión de eventos, $\limsup A_n := \cap_{m\geq 1} \cup_{n\geq m} A_n \in \mathcal{S}$ (y está en el σ-álgebra por ser unión e intersección contable de eventos).

Teorema 3.1 (**Lema de Borel-Cantelli**)**.**

- *Si* $(A_n)_n \subseteq S$ *con* $\sum_{n\in\mathbb{N}} \mathcal{P}(A_n) < \infty$*, entonces* $P(\limsup A_n) = 0$.
- *Si* $(A_n)_n \subseteq S$ *eventos independientes con* $\sum_{n\in\mathbb{N}} \mathcal{P}(A_n) = \infty$*, entonces* $P(\limsup A_n) = 1$.

Demostración.

- Suponga que $\sum_{n\in\mathbb{N}} \mathcal{P}(A_n) < \infty$. Entonces,

$$\begin{aligned}\mathcal{P}(\limsup A_n) = \mathcal{P}(\cap_{m\geq 1} \cup_{n\geq m} A_n) &\leq \mathcal{P}(\cup_{n\geq m} A_n) \\ &\leq \sum_{n\geq m} \mathcal{P}(A_n) = \sum_{n\in\mathbb{N}} \mathcal{P}(A_n) - \sum_{n=1}^{m} \mathcal{P}(A_n).\end{aligned}$$

  A medida que m aumenta, $\sum_{n=1}^{m} \mathcal{P}(A_n)$ converge a $\sum_{n\in\mathbb{N}} \mathcal{P}(A_n)$, luego $\mathcal{P}(\limsup A_n) = 0$.

- Suponga que los A_n son independientes y que $\sum_{n\in\mathbb{N}} \mathcal{P}(A_n) = \infty$. Tome $m \in \mathbb{N}$. Como $\Omega \backslash \cup_{n\in\mathbb{N}} A_n = \cap_{n\in\mathbb{N}} (\Omega \backslash A_n)$,

$$\mathcal{P}(\Omega \backslash \cup_{n\geq m}) = \mathcal{P}(\cap_{n\geq m} (\Omega \backslash A_n)) = \prod_{n\in\mathbb{N}} \mathcal{P}(\Omega \backslash A_n) = \prod_{n\in\mathbb{N}} (1 - \mathcal{P}(A_n)).$$

  Esta última igualdad es cierta dado que, como los A_n son independientes, $\mathcal{P}(\cap_{n\in\mathbb{N}} A_n) = \prod_{n\in\mathbb{N}} \mathcal{P}(A_n)$ y $\mathcal{P}(\cap_{n\in\mathbb{N}} (\Omega \backslash A_n)) = \prod_{n\in\mathbb{N}} (\mathcal{P}(\Omega \backslash A_n)) = \prod_{n\in\mathbb{N}} (1 - \mathcal{P}(A_n))$.

  Si existe $n \geq m$ tal que $\mathcal{P}(A_n) = 1$, entonces $\prod_{n\geq m} (1 - \mathcal{P}(A_n)) = 0$. Si no es el caso, como $\sum_{n\geq m} \mathcal{P}(A_n) = \infty$, $\prod_{n\geq m} (1 - \mathcal{P}(A_n)) = 0$.

  Entonces para todo $m \in \mathbb{N}$, $\mathcal{P}(\Omega \backslash \cap_{n\geq m} A_n) = \mathcal{P}(\cup_{n\geq m} A_n) = 1$. Luego, como $\cup_{n\geq m+1} A_n \subseteq \cup_{n\geq m} A_n$, $\mathcal{P}(\cap_{m\geq 1} (\cup_{n\geq m} A_n)) = \lim_{m\to\infty} \mathcal{P}(\cup_{n\geq m} A_n)$. Por tanto $\mathcal{P}(\limsup A_n) = 1$.

□

Lema 3.2. *Sea* Y *una variable aleatoria no-negativa. Entonces* $EY < \infty$ *si y sólo si* $\sum_{n\in\mathbb{N}} \mathcal{P}(Y > n) < \infty$*, donde* $\mathcal{P}(Y > n) := \mathcal{P}(\{\omega \in \Omega : Y(\omega) > n\})$.

Demostración. Sea $A_k := \{\omega \in \Omega : k < Y(\omega) \leq k+1\}$ para $k \in \mathbb{N}$, $\mathcal{P}(Y > n) := \mathcal{P}(\{\omega \in \Omega : Y(\omega) > n\})$ y $U := \sum_{k\in\mathbb{N}}(k+1)1_{A_k}$ donde 1_{A_k} es la función característica de A_k. Entonces,

$$\sum_{n\in\mathbb{N}} \mathcal{P}(Y > n) = \sum_{n\in\mathbb{N}}(\sum_{k\geq n} \mathcal{P}(A_k)) = \sum_{n\in\mathbb{N}}((k+1)\mathcal{P}(A_k)).$$

Como $EU = \sum_{k\in\mathbb{N}}(k+1)\mathcal{P}(A_k)$ entonces $EU \leq EY$, pues $Y(\omega) < k+1$ para $\omega \in A_k$. Luego se concluye que $EY \leq \sum_{n\in\mathbb{N}} \mathcal{P}(Y > n)$ como se quería. □

Con los lemas y el teorema presentados se demostrará la ley de grandes números en la siguiente sección.

3.2. LFGN para variables i.i.d.

A continuación se enuncia y demuestra la ley fuerte de grandes números para el caso i.i.d.

Teorema 3.2 (**Ley Fuerte de Grandes Números**). *Sea $(\Omega, \mathcal{S}, \mathcal{P})$ un espacio de probabilidad y sean $X_1, X_2, ...$ variables aleatorias i.i.d. en Ω con $\mathrm{E}|X_1| < \infty$. Si $S_n := X_1 + ... + X_n$, entonces,*

$$\frac{S_n}{n} \to \mathrm{E}X_1 \ \ c.s.$$

Demostración. Sean $X_1, X_2, ...$ como en el enunciado.

El teorema se puede restringir al caso en el que $X_n \geq 0$ para todo n: defina $X_i^+ := \max(0, X_i)$ y $X_i^- := -\min(0, X_i)$. Como $|X_i|$ tiene esperanza finita, X_i^+ y X_i^- también. Como las funciones máx y mín son medibles, X_i^+ y X_i^- son independientes. Luego se distribuyen i.i.d y, por tanto, las funciones X_i^+ y X_i^- cumplen las hipótesis de teorema. Ahora, como $X_1 = X_i^+ - X_i^-$, $\mathrm{E}X_1 = \mathrm{E}X_i^+ - \mathrm{E}X_i^-$ y $S_1 = S_i^+ - S_i^-$, demostrar que

$$\frac{S_n^+}{n} \to \mathrm{E}X_1^+ \ \ c.s.$$

implica la conclusión del teorema. Luego la demostración se puede restringir a el caso $X_i \geq 0$.

Se definen $Y_i := \min(X_i, i)$, $T_n := \sum_{j=1}^n Y_i$. Se va a estudiar la convergencia de T_n/n restringido a una subsucesión $k(n) := [\alpha^n]$ definida para un $\alpha > 1$: restringirse a k y usar T_n como la suma truncada de Y_i facilita los cálculos y hace que la función S definida a continuación sea finita.

Sea $S := \sum_{n\geq 1} \mathcal{P}(A_n)$ con $A_n := \{|T_{k(n)} - \mathrm{E}T_{k(n)}|/k(n) > \epsilon\}$ y $\epsilon > 0$ fijo. Luego $S < \infty$. Aplicando el lema 3.1 de Borel-Cantelli a los A_n se obtiene que $\mathcal{P}(\limsup A_n) = 0$, es decir, $|T_{k(n)} - \mathrm{E}T_{k(n)}|/k(n) < \epsilon$ para algún n suficientemente grande. Luego $|T_{k(n)} - \mathrm{E}T_{k(n)}|/k(n)$ converge 0 c.s. Como

$$\mathrm{E}\frac{T_{k(n)}}{k(n)} - \frac{T_{k(n)}}{k(n)} = \mathrm{E}\frac{Y_1 + ... + Y_{k(n)}}{k(n)} - \frac{T_{k(n)}}{k(n)} = \frac{\mathrm{E}Y_1 + ... + Y_{k(n)}}{k(n)} - \frac{T_{k(n)}}{k(n)},$$

converge a 0 c.s. y $\mathrm{E}Y_n$ converge a $\mathrm{E}X_1$ cuando n tiende a ∞, entonces

$$\frac{T_{k(n)}}{k(n)} \text{ converge a } \mathrm{E}X_1 \text{ c.s.} \tag{3.1}$$

Ahora se quiere demostrar ese mismo resultado pero no para Y_i sino para X_i en $X_1, ..., X_{k(n)}$. Se puede usar el lema 3.2 y el hecho de que $\mathrm{E}X_1 = \mathrm{E}|X_1| < \infty$ para afirmar que $\sum_{j=1}^{k(n)} \mathcal{P}(X_j \neq Y_j) = \sum_{j=1}^{k(n)} \mathcal{P}(X_j > j) < \infty$. Luego $X_j \neq Y_j$ ocurre sólo finitas veces c.s. y, por (3.1), se concluye que

$$\frac{S_{k(n)}}{k(n)} := \frac{X_1 + ... + X_{k(n)}}{k(n)} \text{ converge a } \mathrm{E}X_1 \text{ c.s.,} \tag{3.2}$$

encontrando la convergencia para X_i restringido a $k(n)$.

Por último, se quiere probar la convergencia de $S_n = X_1 + ... + X_n$ para cualquier n: como $k(n+1)/k(n)$ converge a a, existe $n \in \mathbb{N}$ tal que $1 \leq k(n+1)/k(n) < \alpha^2$, i.e. $k(n) \leq k(n+1) < \alpha^2$. Luego, para algún $k(n) < j \leq k(n+1)$, $k(n+1) < \alpha^2 k(n) < \alpha^2 j \leq \alpha^2 k(n+1) < \alpha^4 k(n)$. Luego $S_j/\alpha^4 k(n) < S_j/\alpha^2 j \leq S_j/k(n+1)$. Como $S_{k(n)} \leq S_j \leq S_{k(n+1)}$ (por ser $X_i > 0$),

$$\frac{S_{k(n)}}{\alpha^2 k(n)} \leq \frac{S_j}{j} \leq \frac{\alpha^2 S_{k(n+1)}}{k(n+1)}.$$

Luego, por (3.2),

$$\frac{\mathrm{E}X_1}{\alpha^2} \leq \liminf_{j\to\infty} \frac{S_j}{j} \leq \limsup_{j\to\infty} \frac{S_j}{j} \leq \alpha^2 \mathrm{E}X_1.$$

Como α es arbitrario, se puede hacer que α sea tan próximo a 1 como se desee, obteniendo finalmente el resultado del teorema:

$$\frac{S_n}{n} \to \mathrm{E}X_1 \ \ c.s.$$

□

La ley fuerte de los grandes números es cierta con convergencia c.s. El caso particular de esta ley con convergencia en probabilidad se conoce como *ley débil de los grandes números* y se sigue como consecuencia de la nota 2.2.

Teorema 3.3 (**Ley Débil de Grandes Números**). *Sea* $(\Omega, \mathcal{S}, \mathcal{P})$ *un espacio de probabilidad y sean* $X_1, X_2, ...$ *variables aleatorias i.i.d. en* Ω *con* $\mathrm{E}|X_1| < \infty$. *Si* $S_n := \frac{X_1,...,X_n}{n}$, *entonces,*

$$\frac{S_n}{n} \xrightarrow{p} \mathrm{E}X_1.$$

Capítulo 4

Entropía de Shannon

En el capítulo 2 se definió la entropía de Shannon de una fuente **A**. Este capítulo está enfocado en analizar dicho concepto desde dos situaciones diferentes. En la primera sección se estudiará el teorema de la fuente de Shannon, que asigna a $H(\mathbf{A})$ una interpretación en términos de compresión [3]. En la segunda sección se trabajará la entropía cuando se desconoce la distribución de probabilidad asociada a la fuente.

4.1. Teorema de Codificación de la Fuente

Si se quiere codificar la información proveniente de un conjunto A, se requieren $\log|A|$ bits. Sin embargo, se puede admitir un error δ para hacer la codificación, es decir, se pueden omitir los símbolos menos probables y así se requieren menos bits para codificar la información observada. La cantidad H_δ introducida en la ecuación (2.8) cuantifica el número mínimo de bits necesarios para hacer dicha codificación. El teorema de la fuente de Shannon dice que a medida que aumenta el número de observaciones (que se puede asociar a la longitud de las palabras a ser transmitidas), el promedio de bits por carácter mínimo se acerca a la entropía.

Teorema 4.1 (**Teorema de la Fuente de Shannon**). *Sea* $\boldsymbol{A} = (A, \mathcal{P}(A), p)$ *una fuente. Para todo* $\delta > 0$,

$$\lim_{n\to\infty} \frac{1}{n} H_\delta(\boldsymbol{A}^n) = H(\boldsymbol{A}). \tag{4.1}$$

Demostración. Sean $\epsilon > 0$, $a_1, ..., a_n \in A$ n observaciones de A, y $X_1, ..., X_n$ una muestra aleatoria i.i.d. tal que $X_i(a_i) = X(a_i)$, donde $X(a) := -\log(p(a))$ para todo $i \in \{1, ..., n\}$, $a \in A$. Defina

$$\bar{X}_n := \frac{X_1 + ... + X_n}{n} = -\frac{1}{n}\sum_{i=1}^{n} \log(p(a_i))$$

como la *media muestral*. El valor esperado de X viene dado por

$$EX = \sum_{a \in A} p(a)X(a) = -\sum_{a \in A} p(a)\log(p(a)) = H(\mathbf{A}).$$

Sea $\delta > 0$. Se quiere encontrar una cota superior y una cota inferior para $\frac{1}{n}H_\delta(\mathbf{A}^n)$:

- Para encontrar una cota superior, defina

$$A_{n,\epsilon} := \{(a_1, ..., a_n) \in A^n : |\bar{X}_n - H(\mathbf{A})| \leq \epsilon\}.$$

Para todo $(a_1, ..., a_n) \in A_{n,\epsilon}$ tenemos que:

$$\begin{aligned}\epsilon &\geq |\bar{X}_n - H(\mathbf{A})| \\ &= \left| -\frac{1}{n}\sum_{i=1}^{n} \log(p(a_i)) - H(\mathbf{A}) \right| \\ &= \left| -\frac{1}{n}\log(\prod_{i=1}^{n} p(a_i)) - H(\mathbf{A}) \right|.\end{aligned}$$

Como la muestra es i.i.d.,

$$= \left| -\frac{1}{n}\log(p(a_1, ..., a_n)) - H(\mathbf{A}) \right|.$$

Entonces,

$$\begin{aligned}&-\epsilon \leq -\frac{1}{n}\log(p(a_1, ..., a_n)) - H(\mathbf{A}) \leq \epsilon \\ \iff &-\epsilon + H(\mathbf{A}) \leq -\frac{1}{n}\log(p(a_1, ..., a_n)) \leq \epsilon + H(\mathbf{A}) \\ \iff &n(\epsilon - H(\mathbf{A})) \geq \log(p(a_1, ..., a_n)) \geq n(-\epsilon - H(\mathbf{A})) \\ \iff &2^{n(\epsilon - H(\mathbf{A}))} \geq p(a_1, ..., a_n) \geq 2^{n(-\epsilon - H(\mathbf{A}))} \geq 0.\end{aligned}$$

Como la anterior expresión se cumple para todo $\tilde{a} \in A_{n,\epsilon}$,

$$\sum_{\tilde{a} \in A_{n,\epsilon}} 2^{n(\epsilon - H(\mathbf{A}))} \geq \sum_{\tilde{a} \in A_{n,\epsilon}} p(a) \geq \sum_{\tilde{a} \in A_{n,\epsilon}} 2^{n(-\epsilon - H(\mathbf{A}))},$$

obteniendo

$$|A_{n,\epsilon}|2^{n(\epsilon - H(\mathbf{A}))} \geq p(A_{n,\epsilon}) \geq |A_{n,\epsilon}|2^{n(-\epsilon - H(\mathbf{A}))}. \tag{4.2}$$

Como $1 \geq p(A_{n,\epsilon})$, de la anterior expresión obtenemos una cota para $\log(|A_{n,\epsilon}|)$:

$$\begin{aligned}1 \geq |A_{n,\epsilon}|2^{n(-\epsilon - H(\mathbf{A}))} &\iff 2^{n(\epsilon + H(\mathbf{A}))} \geq |A_{n,\epsilon}| \\ &\iff n(\epsilon + H(\mathbf{A})) \geq \log(|A_{n,\epsilon}|).\end{aligned} \tag{4.3}$$

Ahora, queremos probar que $\log(|A_{n,\epsilon}|) \geq H_\delta(\mathbf{A}^n)$ para obtener la cota de H_δ deseada.

Por la ley débil de grandes números (capítulo 3), tenemos que $\bar{X}_n$ converge a EX en probabilidad, i.e. $\Pr(|\bar{X}_n - EX| \leq \epsilon) \to 1$ cuando $n \to \infty$. Entonces, por definición de $A_{n,\epsilon}$, existe N tal que para todo $n \geq N$,

$$\Pr(|\bar{X}_n - EX| \leq \epsilon) \geq 1 - \delta.$$

Luego, $A_{n,\epsilon} \in \{B \subseteq A^n : \Pr(B) \geq 1 - \delta\}$. De la ecuación (2.3) se deduce que:

$$\log(|A_{n,\epsilon}|) \geq H_\delta(\mathbf{A}^n).$$

Por la ecuación (4.3) obtenemos finalmente la cota:

$$\epsilon + H(\mathbf{A}) \geq \frac{1}{n} H_\delta(\mathbf{A}^n). \tag{4.4}$$

Luego, se obtiene una cota para el limsup de $H_\delta(\mathbf{A}^n)$:

$$H(\mathbf{A}) \geq \lim_{n\to\infty} \sup \frac{1}{n} H_\delta(\mathbf{A}^n). \tag{4.5}$$

- Ahora, para encontrar una cota inferior de $H_\delta(\mathbf{A}^n)$, se define $B_{n,\delta} \subseteq A^n$ como el conjunto que minimiza $H_\delta(A^n)$. Luego $H_\delta(\mathbf{A}^n) = \log(|B_{n,\delta}|)$ y,

  $$P(B_{n,\delta}) \geq 1 - \delta$$

  Queremos estimar $\log(|A_{n,\epsilon} \cap B_{n,\delta}|)$. Por lo anterior, tenemos que:

  $$P(B_{n,\delta}) = P(A_{n,\epsilon} \cap B_{n,\delta}) + P(A^c_{n,\delta} \cap B_{n,\delta}) \geq 1 - \delta. \tag{4.6}$$

  Luego, como $|A_{n,\epsilon} \cap B_{n,\delta}| \leq |A_{n,\epsilon}|$, $\mathcal{P}(A_{n,\epsilon} \cap B_{n,\delta}) \leq \mathcal{P}(A_{n,\epsilon})$, por (4.2) se obtiene que

  $$P(A_{n,\epsilon} \cap B_{n,\delta}) \leq |A_{n,\epsilon} \cap B_{n,\delta}| 2^{n(\epsilon - H(\mathbf{A}))}.$$

  Además, como $P(A_{n,\epsilon}) \to 1$ cuando $n \to \infty$, existe $N \in \mathbb{N}$ tal que para todo $n \geq N$,

  $$P(A_{n,\epsilon}) \geq 1 - \delta,$$

  entonces,

  $$P(A^c_{n,\epsilon}) = 1 - P(A_{n,\epsilon}) \leq 1 - 1 + \delta = \delta. \tag{4.7}$$

Luego, de (4.6) y (4.7) se deduce que:

$$|A_{n,\epsilon} \cap B_{n,\delta}|2^{-n(\epsilon - H(\mathbf{A}))} + \delta \geq 1 - \delta,$$

que se puede reescribir como:

$$|A_{n,\epsilon} \cap B_{n,\delta}| \geq (1 - 2\delta)2^{-n(\epsilon - H(\mathbf{A}))}.$$

Tenemos además,

$$H_\delta(\mathbf{A}^n) = \log(|B_{n,\delta}|) \geq \log(|A_{n,\epsilon} \cap B_{n,\delta}|).$$

De lo anterior se deduce que:

$$\frac{1}{n}H_\delta(\mathbf{A}^n) \geq \frac{1}{n}\log(1 - 2\delta) + H(\mathbf{A}) - \epsilon.$$

Como el lado derecho de la desigualdad tiende a $H(\mathbf{A}) - \epsilon$ cuando n tiende a ∞, obtenemos finalmente la cota para $H_\delta(\mathbf{A}^n)$:

$$\liminf_{n\to\infty} \frac{1}{n}H_\delta(\mathbf{A}^n) \geq H(\mathbf{A}). \tag{4.8}$$

Con los resultados en (4.5) y (4.8) se demuestra el teorema:

$$\lim_{n\to\infty} \frac{1}{n}H_\delta(\mathbf{A}^n) = H(\mathbf{A}).$$

□

Ejemplo 4.1. Sea $\mathbf{A}$ una fuente como se trabajó en los ejemplos 2.1 y 2.2 (caso binario). Para $p = 1/4$, la entropía es:

$$H(\mathbf{A}) = -\frac{1}{4}\log\frac{1}{4} - (\frac{3}{4})\log(\frac{3}{4}) \approx 0{,}81.$$

Sin importar el valor de δ, eventualmente el número mínimo promedio de bits por carácter se acercará a la entropía de Shannon. En la figura 4.1 se muestra el comportamiento de $H_\delta(\mathbf{A}^n)$ para $\delta = 1/20 < p$ para los valores $n = 1, ..., 25$. Como el error es pequeño y menor que p, los conjuntos que satisfacen $P(B) \geq 1 - \delta$ son menos y grandes y, por tanto, la curva es decreciente. Se exige que sea menor que p para que inicialmente $H_\delta(\mathbf{A}^n) > nH(\mathbf{A})$.

Por el contrario, si $\delta = 1/3 > p$, el error es grande y los conjuntos son más y pueden ser muy pequeños. Luego la curva va creciendo asintóticamente hacía $H(A)$, que se muestra horizontalmente en la figura 4.2.

Como el teorema de Shannon indica, a medida que aumenta n, el tamaño requerido para codificar la información es el mismo sin importar el error. Una consecuencia de esto se conoce como *principio de equipartición asintótica*. A medida que n aumenta la probabilidad de las secuencias de observaciones en A^n más probables se hace mucho mayor que la probabilidad de las menos probables.

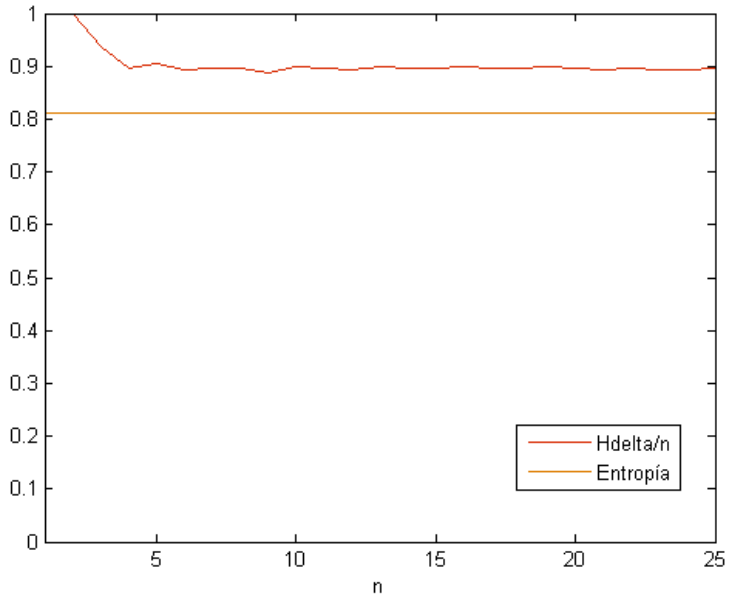

Figura 4.1: n vs. $\frac{1}{n}H_\delta(\mathbf{A}^n)$ para $\delta = 1/20$.

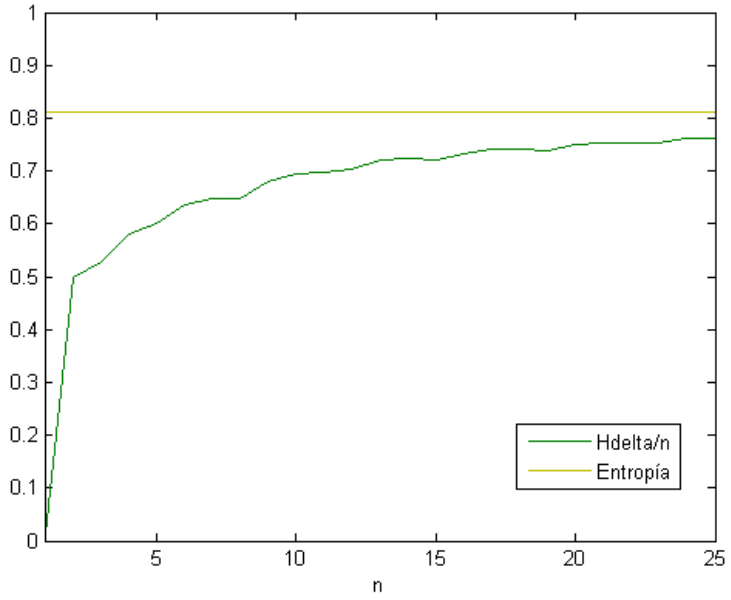

Figura 4.2: n vs. $\frac{1}{n}H_\delta(\mathbf{A}^n)$ para $\delta = 1/3$.

4.2. Estimación de la Entropía

En capítulos anteriores se introdujo el concepto de entropía y se interpretó dicha cantidad. A su vez se demostró que la entropía de una fuente establece asintóticamente una cota máxima de compresión (dado que en el caso de δ pequeño, $\frac{1}{n}H_\delta(A^n)$ decrece). Para concluir estas observaciones, se asume que la distribución de cada variable es conocida. Sin embargo, cuando se desconoce dicha información, no se puede determinar la probabilidad de obtener una secuencia específica de caracteres ni se puede calcular la entropía de una fuente. En la presente sección se presentará un estimador para la entropía, basándose en la estimación de las probabilidades, así como el error que se obtiene al hacer dicha estimación.

A lo largo del capítulo se hará uso del Teorema del Límite Central, del Método Delta Multivariado y del Teorema del Mapa Continuo. Estos serán enunciados pero su demostración no hace parte de los objetivos de este documento. Los enunciados y sus respectivas pruebas pueden ser consultadas en [4], [5], [6] respectivamente.

Sea $\mathbf{A} = (A, \mathcal{P}(A), p)$ un espacio de probabilidad con $A = \{a_1, ..., a_M\}$ finito (M letras), asumiendo por razones técnicas que $0 < p(a_i) < 1$ para todo $i \in \{1, ..., M\}$. Por notación, $p_i := p(a_i)$. Sean $X_1, ..., X_n$ vectores aleatorios tal que si la i-ésima letra que se observa es a_j, X_i es el j-ésimo vector de la base estándar de $\mathbb{R}^M$, i.e. $X_i = \hat{e_j} \in \mathbb{R}^n$. Luego $x_1, ..., x_n$ se distribuyen i.i.d. como una multinomial con parámetros $(1, p_1, ..., p_M)$, con media $\bar{p} = (p_1, ..., p_M)$. Defina el número de ocurrencías de a_j como $\hat{p}_{n,j} := |\{i : X_i = \hat{e}_j\}|/n$. Observe que:

$$[\frac{1}{n}(X_1 + ... + X_n)]_j = \hat{p}_{n,j},$$

i.e. $\frac{1}{n}\bar{X}_n = \hat{\bar{p}}_n$. Luego, por la ley débil de grandes números (teorema 3.3), $\hat{\bar{p}} \xrightarrow{p} \bar{p}$.

Ahora, como la función $f(x) = -x \log x$ es continua para todo $x \in (0, \infty)$,

$$-\hat{\bar{p}}_n \log \hat{\bar{p}}_n \xrightarrow{p} -\bar{p} \log \bar{p}.$$

Tomando como motivación la anterior ecuación, se define el *estimador de entropía* como:

$$\hat{H}_n := -\sum_{j=1}^{M} \hat{p}_{n,j} \log \hat{p}_{n,j}. \tag{4.9}$$

Para el caso en el que $\hat{p}_{n,j} = 0$, $\hat{p}_{n,j} \log \hat{p}_{n,j} := 0$.

El objetivo ahora no es sólo demostrar que $\hat{H}_n \xrightarrow{d} H$, sino encontrar el error al hacer dicha estimación. Para ello iniciamos con el siguiente teorema.

Teorema del Límite Central. *Sean $X_1, ..., X_n$ vectores aleatorios i.i.d. con media μ y matriz de covarianzas C definida positiva. Entonces,*

$$Y_n = \sqrt{n}\sum_{i=1}^{n}(X_i - \mu) \xrightarrow{d} N(0, C),$$

donde $N(0, C)$ es la distribución normal multivariada M-dimensional.

Para los vectores $X_1, ..., X_n$ introducidos previamente, el Teorema del Límite Central implica que:

$$\sqrt{n}\begin{pmatrix} \hat{p}_{n,1} - p_1 \\ \vdots \\ \hat{p}_{n,M} - p_M \end{pmatrix} = \sqrt{n}\sum_{1}^{n}(\hat{\bar{p}}_n - \bar{p}) \xrightarrow{d} N(0, C), \tag{4.10}$$

donde C es la matriz de covarianzas de la distribución multinomial con parámetro $(1, \bar{p})$.

Teniendo en cuenta que el parámetro de la distribución es 1,

$$Cov(X_i, X_j) = \begin{cases} p_i(1 - p_i), & \text{si } i = j \\ -p_i p_j, & \text{si } i \neq j \end{cases},$$

luego C viene dada por:

$$\begin{aligned} C &= \begin{pmatrix} Cov(X_1, X_1) & Cov(X_1, X_2) & \dots \\ Cov(X_2, X_1) & Cov(X_2, X_2) & \dots \\ \vdots & \vdots & \ddots \end{pmatrix} \\ &= \begin{pmatrix} p_1(1 - p_1) & -p_1 p_2 & \dots \\ -p_2 p_1 & p_2(1 - p_2) & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}. \end{aligned} \tag{4.11}$$

Ahora, para poder analizar la convergencia deseada, se hace uso del siguiente método, que se puede consultar en detalle en [5], y cuya prueba será omitida.

Método Delta Multivariado. *Sea $G : \mathbb{R}^k \to \mathbb{R}^l$ un mapa medible diferenciable en un punto $x_o \in \mathbb{R}^k$. Suponga que $\sqrt{n}(x_n - x_o) \xrightarrow{d} Z$. Entonces:*

$$\sqrt{n}(G(x_n) - G(x_o)) \xrightarrow{d} LZ,$$

donde $L : \mathbb{R}^k \to \mathbb{R}^l$ es el mapa lineal tal que:

$$G(x) = G(x_o) + L(x - x_o) + \sigma(|x - x_o|^2),$$

para alguna vecindad de x_o (que existe dada la diferenciabilidad de G en x_o).

Para analizar la convergencia del vector $(-\hat{p}_{n,i}\log\hat{p}_{n,i})_i$ haciendo uso del método Método Delta, se define $G:\mathbb{R}^M\to\mathbb{R}^M$ como:

$$G\begin{pmatrix} x_1 \\ \vdots \\ x_M \end{pmatrix} = \begin{pmatrix} -x_1\log x_1 \\ \vdots \\ -x_M\log x_M \end{pmatrix}.$$

El mapa $L:\mathbb{R}^M\to\mathbb{R}^M$ correspondiente viene dado por la matriz Jacobiana de G:

$$L = \begin{pmatrix} -\log p_1 - 1 & & & \\ & -\log p_2 - 1 & & 0 \\ 0 & & \ddots & \\ & & & -\log p_M - 1 \end{pmatrix}.$$

Entonces, como G es diferenciable en $\bar{p}$ (dado que se estableció que $p_i\neq 0$ para todo i), aplicamos el método a partir del resultado obtenido en la ecuación (4.10), obteniendo:

$$\sqrt{n}\begin{pmatrix} \hat{p}_{n,1}\log\hat{p}_{n,1} - p_1\log p_1 \\ \vdots \\ \hat{p}_{n,M}\log\hat{p}_{n,M} - p_M\log p_M \end{pmatrix} \xrightarrow{d} LZ,$$

donde Z es la distribución $N(0,C)$. Como LZ tiene matriz de covarianza LCL^t, se obtiene finalmente que:

$$\sqrt{n}\begin{pmatrix} \hat{p}_{n,1}\log\hat{p}_{n,1} - p_1\log p_1 \\ \vdots \\ \hat{p}_{n,M}\log\hat{p}_{n,M} - p_M\log p_M \end{pmatrix} \xrightarrow{d} N(0,D), \tag{4.12}$$

donde la matriz D viene dada por:

$$D = LCL^t = \begin{pmatrix} s(p_1)^2p_1(1-p_1) & -s(p_1)s(p_2)p_1p_2 & \dots \\ -s(p_2)s(p_1)p_1p_2 & s(p_2)^2p_2(1-p_2) & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}, \tag{4.13}$$

donde $s(p)=\log p+1$.

Hasta ahora se ha estudiado la convergencia asintótica del vector en $\mathbb{R}^M$ dado por $-(\hat{\hat{p}}_{n,i}\log\hat{\hat{p}}_{n,i})_i$. El objetivo final es encontrar la convergencia de $\hat{H}_n = -\sum_i \hat{\hat{p}}_{n,i}\log\hat{\hat{p}}_{n,i}$. Gracias al siguiente teorema, se puede determinar el error en la convergencia de $\sqrt{n}(\hat{H}_n - H)$.

Teorema del Mapa Continuo. *Sea $X_1,...,X_n$ una secuencia de vectores aleatorios tales que $X_n\xrightarrow{d}X$ para algún vector aleatorio X, y sea g una función continua en el soporte de X. Entonces,*

$$g(X_n)\xrightarrow{d}g(X).$$

A partir del resultado obtenido en la ecuación (4.12), se necesita una función $F : \mathbb{R}^M \to \mathbb{R}$ que satisfaga las condiciones del Teorema del Mapa Continuo, tal que $F((-(\hat{\bar{p}}_{n,i} \log \hat{\bar{p}}_{n,i}))_i) = \hat{H}_n$, que es la variable que se desea analizar. Para ello, se define $F(\bar{x}) = \mathbb{1}_M^t \cdot \bar{x}$ con $\mathbb{1}_M := (1 \dots 1) \in \mathbb{R}^M$. Luego,

$$\begin{aligned} &\sqrt{n}(F(\hat{\bar{p}}_n) - F(\bar{p})) \\ &\quad = \sqrt{n}(\hat{H}_n - H) \xrightarrow{d} N(0, \mathbb{1}_M^t D \mathbb{1}_M). \end{aligned} \tag{4.14}$$

Algoritmo

Para simular el comportamiento asintótico de $\hat{H}_n$ de una fuente con alfabeto finito y probabilidades $\bar{p}$ (con $0 < p_i < 1$), se sigue el siguiente algoritmo:

- Para n fijo, se generan muestras de n vectores aleatorios con distribución multivariada $(1, \bar{p})$.
- Para cada muestra se calculan los estimadores de cada p_i, y se computa $\hat{H}_n$ según la ecuación (4.9).
- El anterior procedimiento se realiza l veces y con los $\hat{H}_n$ obtenidos se construye un histograma normalizado.

Ejemplo 4.2. Considere el caso en el que $A = \{a_1, a_2, a_3, a_4\}$ y la distribución es $p(a_1) = p(a_2) = 1/8$, $p(a_3) = 1/4$, $p(a_4) = 1/2$. A continuación se presenta el histograma obtenido en base al algoritmo previamente presentado, para los valores $n = 1000$ y $l = 10000$.

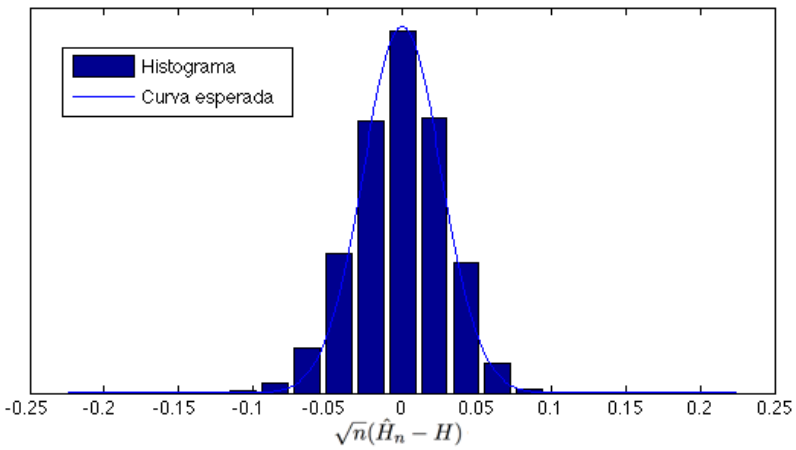

Figura 4.3: Histograma para $n = 1000$ y $l = 10000$.

El histograma anterior está centrado en 0, como se esperaba, y coincide con la curva normal cuya varianza es $\mathbb{1}_M^t D \mathbb{1}_M$.

La varianza calculada por el método y la varianza de la muestra simulada se muestra en la siguiente tabla, para diferentes valores de n y l. A medida que n aumenta, ambos valores se hacen más cercanos.

n	Var. Teórica	Var. Muestra
100	0.6875	0.7179
1000	0.6875	0.6967
10000	0.6875	0.6910
100000	0.6875	0.6896

Tabla 4.1: Comparación de los valores de la varianza para diferentes valores de n y para $l = 10000$.

Caso binomial

En esta sección se analizará el caso en el que $A = \{a_1, a_2\}$, en donde la distribución de las letras sigue una distribución binomial con parámetro p (donde $p = P(a_1)$. La matriz D correspondiente (de acuerdo con la ecuación (4.13) es:

$$D = \begin{pmatrix} (\log p + 1)^2 p(1-p) & -(\log p + 1)(\log(1-p) + 1)p(1-p) \\ -(\log p + 1)(\log(1-p) + 1)p(1-p) & (\log(1-p) + 1)^2 p(1-p) \end{pmatrix}$$

$$= p(1-p) \begin{pmatrix} (\log p + 1)^2 & -(\log p + 1)(\log(1-p) + 1) \\ -(\log p + 1)(\log(1-p) + 1) & (\log(1-p) + 1)^2 \end{pmatrix}.$$

Luego,

$$\mathbb{1}_2 D \mathbb{1}_2 = p(1-p)\left(\log \frac{p}{1-p}\right)^2. \tag{4.15}$$

Del resultado anterior se puede construir la siguiente gráfica, que representa el comportamiento de (4.15) en función de $p \in (0, 1)$:

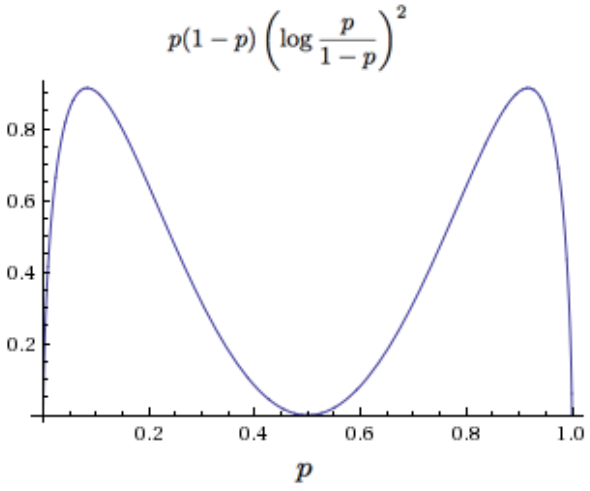

Figura 4.4: Varianza asintótica teórica de $\hat{H}_n$ en función de p.

Para $p = 1/2$, el valor de la varianza asintótica es 0. ¿Qué significa ese resultado? Una varianza nula significaría que para esa probabilidad, no hay error en la estimación. Para entender ese resultado se debe notar que, en este caso, la entropía es máxima, como se mostró en la figura 2.1. En general, a medida que $\hat{p}_n$ oscila alrededor de p, $\hat{H}_n$ oscila alrededor de H como una normal. Sin embargo, al ser H máxima en este caso, $\hat{H}$ no pude tomar valores superiores al máximo, por tanto la estimación no está centrada en 0 y el resultado mostrado en la figura 4.4 para $p = 1/2$ no sería válido. Esta situación, en la que no se puede asumir una distribución normal, ocurre además cuando $p = 0$ y $p = 1$. Para visualizar este análisis, en la figura 4.5 se muestra el histograma para $n = 1000$, $l = 10000$, de las observaciones cuando $p = 1/2$.

Note que el histograma no está centrado en 0 como se discutió y que la varianza es muy pequeña. Para representar esa proporción y, además, para estudiar la validez de (4.15), se realizó una simulación para $n = 2000$ y $l = 50000$ de la varianza asintótica para 98 valores de p entre 0,1 y 0,99. Los resultados se muestran en la figura 4.6. Comparado con la figura 4.4, las simulaciones observadas coinciden con las predicciones hechas en esta sección, confirmando cómo es el comportamiento del error cuando se aproxima la entropía con el estimador $\bar{H}_n$.

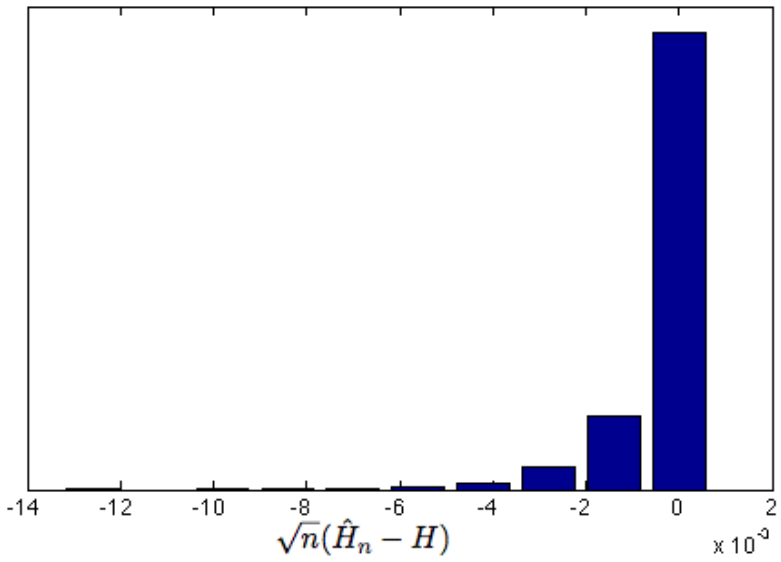

Figura 4.5: Histograma para $p = 1/2$, $n = 1000$ y $l = 10000$.

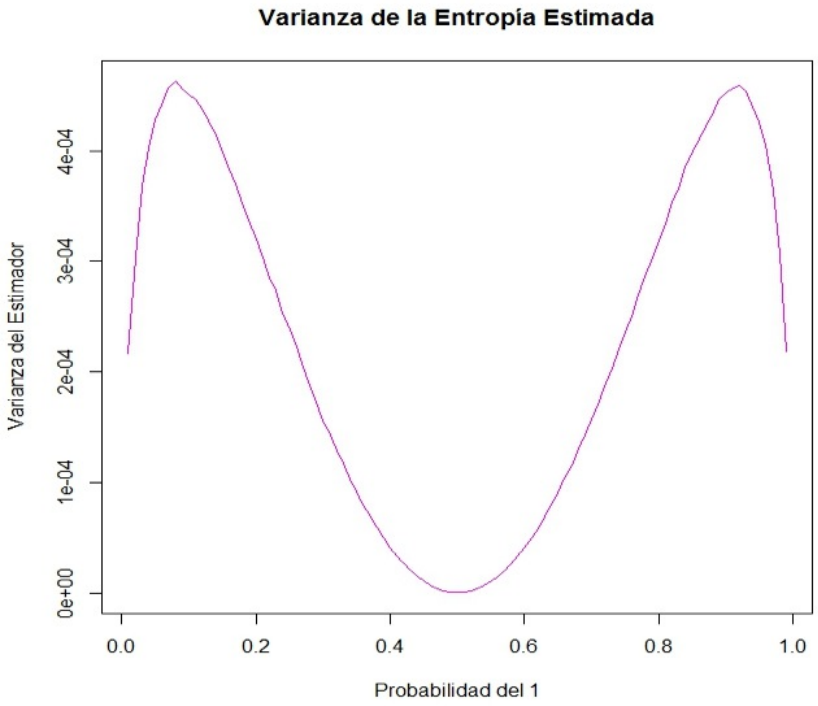

Figura 4.6: Varianza de la Entropía Estimada en el caso binario.

Capítulo 5

Clasificaciones y Lógicas

Como se trabajó en capítulos previos, el concepto de información puede extenderse y estudiarse desde un marco probabilístico, y con ayuda de herramientas cuantitativas, como la estadística o la teoría de la medida, se pueden predecir resultados numéricos. Sin embargo se han desarrollado formalismos cualitativos que buscan entender el concepto de información y transferencia haciendo uso de la lógica formal. Como ejemplo de estos formalismos, en el libro *Information Flow: The Logic of Distributed Systems* [7], Jon Barwise y Jerry Seligman construyen un modelo de flujo de información usando un modelo de canal más abstracto que el trabajado por Shannon, donde el flujo de información es posible gracias a las regularidades presentes en el sistema. En ese desarrollo, la descripción de los canales de información se basa en los conceptos de información e infomorfismos. En este capítulo se presentarán los conceptos básicos de esa teoría y, omitiendo el modelo de canales de Barwise y Seligman, se estudiará la conexión entre el concepto de clasificación y las nociones de fuente y entropía trabajadas en capítulos anteriores. De esa forma se construirá una forma de integrar los marcos cuantitativos y cualitativos de la Teoría de la Información.

Una clasificación es una tupla de dos conjuntos y una relación, como se introduce a continuación.

Definición 5.1 (**Clasificación**)**.**
Una *clasificación* es una tupla $\mathbb{A} = \langle A, T_A, \models_A \rangle$, donde S es el conjunto de *instancias*, T_A es el conjunto de *tipos* y $\models_A \subseteq A \times T_A$ es una relación binaria entre ambos conjuntos.

Las instancias son objetos que van a ser clasificados, mientras que los tipos son objetos para clasificar instancias. La relación $\models_A$ indica cómo ocurre la clasificación. A una clasificación A se le puede asociar una *teoría*, que consiste en todos las parejas $\langle \Gamma, \Delta \rangle$ de subconjuntos de T_A tales que si una instancia satisface todos los tipos en Γ, satisface alguno en Δ. Esa restricción

impone reglas al construir una clasificación. Si un secuente $\Gamma \vdash \Delta$ está en la teoría, ninguna instancia se puede clasificar en todos los tipos de Γ y ninguno de Δ. Esta colección de parejas representa todas las regularidades del sistema modelado por A [7] y con la cual se puede modelar la transferencia de información.

La clasificación y la teoría de ésta no tienen ningún carácter aleatorio. Sin embargo, para buscar una conexión con el concepto de fuente y entropía de la teoría de Shannon, es necesario asignarle una distribución de probabilidad al conjunto de instancias, como se presenta a continuación.

Definición 5.2 (**Clasificación Probabilística**)**.** Una *clasificación probabilística* es una tupla $\mathbb{C} = \langle Inst_c, Tipos_c, \models_c \rangle$ donde:

- $Inst_c = \langle I_c, \mathcal{B}_c, \mathcal{P}_c \rangle$ es un espacio de probabilidad.
- $Tipos_c = \{T_1, ..., T_N\}$ es un conjunto no vacío y finito de *tipos*.
- $\models_c \subseteq I_c \times Tipos_c$ es una relación binaria tal que para todo $T_i \in Tipos_c$, la *extensión de* T_i $[\![T_i]\!] = \{a \in I_c : a \models_c T_i\}$ es medible, es decir, $[\![T_i]\!] \in \mathcal{B}_c$.

Cada instancia $c \in I_c$ de una clasificación satisface un conjunto de tipos de $\mathbb{C}$. Por ello se define el *estado de* c como la tupla $\langle \bar{T}_1, ..., \bar{T}_N \rangle$, donde

$$\bar{T}_i = \begin{cases} 0, & \text{si } c \not\models_c T_i \\ 1, & \text{si } c \models_c T_i \end{cases}$$

para todo $i = 1, ..., N$. Se define la *extensión de* $\langle \bar{T}_1, ..., \bar{T}_N \rangle$ como

$$[\![\bar{T}_1, ..., \bar{T}_N]\!] := \{c \in I_c : \text{ estado de c } = \langle \bar{T}_1, ..., \bar{T}_N \rangle\}.$$

Note que el conjunto es medible, pues es intersección de conjuntos medibles:

$$[\![\bar{T}_1, ..., \bar{T}_N]\!] = (\bigcap \{[\![\bar{T}_i]\!] : \bar{T}_i = 1\}) \cap (\bigcap \{[\![\bar{T}_i]\!]^c : \bar{T}_i = 0\}) \in \mathcal{B}.$$

A partir de una clasificación probabilística se puede definir una fuente en el sentido de Shannon, haciendo uso de las anteriores definiciones.

Definición 5.3 (**Fuente**)**.** Una *fuente* $\mathcal{F}$ inducida por una clasificación probabilística $\mathbb{C}$ es una tupla $\langle A, \mathscr{P}(A), \mathcal{P}_{alf} \rangle$ donde:

- $A = \{\langle \bar{T}_1, ..., \bar{T}_N \rangle : \bar{T}_i \in \{0, 1\}\} = \{0, 1\}^N$ es el *universo* de la fuente.
- $\mathcal{P}_{alf}(\langle \bar{T}_1, ..., \bar{T}_N \rangle) = \mathcal{P}_c([\![\bar{T}_1, ..., \bar{T}_N]\!])$ es la probabilidad heredada de $\mathbb{C}$.

Note que $\mathcal{F}$ es un espacio de probabilidad.

Nota: Como se pide en la Definición 5.2, el conjunto de tipos es finito. Por esa razón, aunque el conjunto de instancias puede ser infinito, el universo de la fuente siempre será finito como los alfabetos trabajados en el capítulo 4.

El alfabeto de la fuente cuantifica las observaciones que provienen del conjunto de instancias y hereda la distribución de probabilidad de $\mathbb{C}$ de forma única. El siguiente ejemplo muestra que, en cambio, una fuente en el sentido de 5.2 puede corresponder a dos clasificaciones probabilísticas distintas.

Ejemplo 5.1. Sean $\mathbb{C} = \langle I_C, T, \models_C \rangle$ y $\mathbb{D} = \langle I_D, T, \models_D \rangle$ clasificaciones tales que $T = \{T_1, T_2\}$ y:

a. Instancias $I_C = \{c_{11}, c_{10}, c_{01}\}$,

$\mathcal{B}_C = \mathscr{P}(I_C)$,

$\mathcal{P}_C(c_{11}) = \mathcal{P}_C(c_{10}) = 1/4$, $\mathcal{P}_C(c_{01}) = 1/2$, y

la relación es $\models_C = \{(c_{10}, T_1), (c_{01}, T_2), (c_{11}, T_1), (c_{11}, T_2)\}$.

b. Instancias $I_D = \{c_{11}, c_{10}, c_{01}, \widetilde{c_{01}}\}$,

$\mathcal{B}_D = \mathscr{P}(I_D)$,

la probabilidad de las instancias es $\mathcal{P}_D(c) = 1/4$ para todo $c \in I_C$,

la relación es $\models_D = \{(c_{10}, T_1), (c_{01}, T_2), (\widetilde{c_{01}}, T_2), (c_{11}, T_1), (c_{11}, T_2)\}$.

Para ambas clasificaciones el alfabeto de la fuente es

$$A = \{\langle 0,0\rangle, \langle 0,1\rangle, \langle 1,0\rangle, \langle 1,1\rangle\}$$

y la probabilidad es $\mathcal{P}_{Alf}(\langle 0,0\rangle) = 0$, $\mathcal{P}_{Alf}(\langle 0,1\rangle) = 1/2$ y $\mathcal{P}_{Alf}(\langle 1,0\rangle) = \mathcal{P}_{Alf}(\langle 1,1\rangle) = 1/4$. Por tanto, las fuentes de ambas clasificaciones son iguales.

Como se mencionó antes, la teoría de una clasificación contiene las relaciones entre tipos que se cumplen para todas las instancias, que es necesario conocerlas en el contexto de la teoría cualitativa. Con el fin de extenderla al caso de clasificaciones probabilísticas, se definen la lógica absoluta y la lógica probabilística. La primera reúne las mismas ideas que la teoría. La segunda está relacionada con el carácter probabilístico de las instancias. Para entender estas lógicas es necesario introducir primero la noción de secuente.

Definición 5.4 (**Secuente**)**.** Sea $\mathbb{C}$ una clasificación y $c \in I_C$ una instancia. Si Γ, Δ son subconjuntos de $Tipos_C$, a $\Gamma \vdash \Delta$ se le denomina *secuente*.
Se dice que c *satisface* $\Gamma \vdash \Delta$ si

$$(\forall \gamma \in \Gamma : c \vdash \gamma) \Rightarrow (\exists \delta \in \Delta : c \vdash \delta).$$

Un secuente es *trivial* si es de la forma $\Gamma \vdash \Delta$ donde $\Gamma \cap \Delta$ es no vacío. Es decir, todas las instancias lo satisfacen sin importar la estructura de la clasificación.

Definición 5.5 (**Lógica de una clasificación**). Sea $\mathbb{C}$ una clasificación.

- La *lógica absoluta* de $\mathbb{C}$ es el conjunto

$$\mathcal{L}_{\mathbb{C}} := \{\Gamma \vdash \Delta : \forall c \in I_c \ \ (c \text{ satisface } \Gamma \vdash \Delta)\}.$$

- La *lógica probabilística* de $\mathbb{C}$ es el conjunto

$$\mathcal{L}^{Pr}_{\mathbb{C}} := \{\Gamma \vdash_p \Delta : \mathcal{P}_c(\{c : c \text{ satisface } \Gamma \vdash \Delta\}) = p\}.$$

Dado que el conjunto

$$\{c : c \text{ satisface } \Gamma \vdash \Delta\} = (\bigcup_{T_i \in \Gamma} \{[\![\bar{T}_i]\!]^c : \bar{T}_i = 1\}) \cup (\bigcup_{T_j \in \Delta} \{[\![\bar{T}_j]\!] : \bar{T}_j = 1\}) \in \mathcal{B}_C$$

es medible, la definición de $\mathcal{L}^{Pr}_{\mathbb{C}}$ tiene sentido.

Mientras que la lógica abstracta contiene los secuentes que son satisfechos por todas las instancias, la lógica probabilística contiene todos los secuentes (así no exista una instancia que los satisfaga), cada uno con una probabilidad asociada. Como para la fuente, dos clasificaciones diferentes pueden tener las mismas lógicas:

Ejemplo 5.2. Sean $\mathbb{C}$ y $\mathbb{D}$ las clasificaciones definidas en el ejemplo 5.1. El único secuente no trivial satisfecho por toda instancia de $\mathbb{C}$ es $\vdash \{T_1, T_2\}$. De igual forma, el único secuente no trivial satisfecho por toda instancia de $\mathbb{D}$ es $\vdash \{T_1, T_2\}$. Luego $\mathcal{L}_{\mathbb{C}} = \mathcal{L}_{\mathbb{D}}$.
Como para todo secuente los conjuntos $\{c : c \text{ satisface } \Gamma \vdash \Delta\}$ tienen la misma distribución de probabilidad para ambas clasificaciones, la lógica probabilística coincide.

En general, si las fuentes asociadas a dos clasificaciones son iguales, lo son también sus lógicas correspondientes. Gracias a ello, las lógicas se pueden caracterizar en términos de la fuente:

Definición 5.6. Sea $\mathbb{C}$ una clasificación.

- $E_{\models} := \{\langle \bar{T}_1, ..., \bar{T}_N \rangle : [\![\bar{T}_1, ..., \bar{T}_N]\!] \neq \emptyset\} \subseteq A$ es el conjunto de los elementos del alfabeto que tienen asociada al menos una instancia.

- $B_{\Gamma \vdash \Delta} :=$

$$\{\langle \bar{T}_1, ..., \bar{T}_N \rangle \in E_{\models} : (\forall T_i \in \Gamma : \bar{T}_i = 1) \Rightarrow (\exists T_j \in \Delta : \bar{T}_j = 1)\}.$$

Proposición 5.1. *Sea $\mathbb{C}$ una clasificación y $\Gamma \vdash \Delta$ un secuente.*

1. *$\Gamma \vdash \Delta \in \mathcal{L}_{\mathbb{C}}$ si y sólo si $E_{\models} \subseteq B_{\Gamma \vdash \Delta}$.*

2. $\Gamma \vdash_p \Delta \in \mathcal{L}_{\mathbb{C}}^{Pr}$ *si y sólo si* $\mathcal{P}(B_{\Gamma\vdash\Delta}) = p$.

Demostración. Se fija el secuente $\Gamma \vdash \Delta$.

1. Suponga que $\Gamma \vdash \Delta \in \mathcal{L}_{\mathbb{C}}$. Luego para todo $c \in I_c$, c satisface $\Gamma \vdash \Delta$. Tome $\langle \bar{T}_1, ..., \bar{T}_N \rangle \in E_{\models}$. Existe $c \in I_c$ tal que $c \in [\![\bar{T}_1, ..., \bar{T}_N]\!]$. Como c satisface $\Gamma \vdash \Delta$ entonces $\langle \bar{T}_1, ..., \bar{T}_N \rangle \in B_{\Gamma\vdash\Delta}$.

   Ahora suponga que $E_{\models} \subseteq B_{\Gamma\vdash\Delta}$. Tome $c \in I_c$ y sea $\langle \bar{T}_1, ..., \bar{T}_N \rangle$ el estado de c. Entonces $\langle \bar{T}_1, ..., \bar{T}_N \rangle \in E_{\models}$. Luego $\langle \bar{T}_1, ..., \bar{T}_N \rangle \in B_{\Gamma\vdash\Delta}$. Por tanto c satisface $\Gamma \vdash \Delta$.

2. $\Gamma \vdash_p \Delta \in \mathcal{L}_{\mathbb{C}}^{Pr}$ si y sólo si $p = \mathcal{P}_c(\{c \in I_c : c \text{ satisface } \Gamma \vdash \Delta\}) = \mathcal{P}_{alf}(\{B_{\Gamma\vdash\Delta} \cap E_{\models}\}) = \mathcal{P}_{alf}(\{B_{\Gamma\vdash\Delta}\})$. Esta última igualdad es cierta ya que $\mathcal{P}_{alf}(e) = 0$ para todo $e \in E_{\models}$.

□

Nota: Con la proposición se puede notar la diferencia entre los secuentes $\Gamma \vdash \Delta$ y $\Gamma \vdash_1 \Delta$: En la lógica absoluta, todas las instancias satisfacen el secuente $\Gamma \vdash \Delta$, mientras que en la lógica probabilística, el conjunto de instancias que satisfacen $\Gamma \vdash_1 \Delta$ tiene probabilidad 1 pero puede ser un subconjunto propio del conjunto completo de todas las instancias I_C.

Con estas caracterizaciones se puede estudiar cómo estimar las lógicas si se desconoce $\mathcal{P}_c$, como se explicará en la siguiente sección.

5.1. Estimación de la Lógica

Suponga que se tiene una fuente $\mathcal{F}$ asociada a una clasificación $\mathbb{C}$, y que independientemente se observan instancias $c_1, c_2, ... \in I_c$, cada uno con una letra del alfabeto asociada respectivamente $b_1, b_2, ..., b_n \in A = \{a_1, ..., a_N\}$, y su respectiva probabilidad $p_1 = \mathcal{P}_{alf}(\{a_1\})$,..., $p_N = \mathcal{P}_{alf}(\{a_N\})$. (Tratar a $\mathcal{F}$ como una fuente i.i.d. es natural en este contexto, pues no es necesario para las estimaciones que se realizarán, tener en cuenta efectos de memoria en las observaciones). Ahora suponga que la distribución $\mathcal{P}_{alf}$ se desconoce y se quiere estimar el contenido de las lógicas correspondientes a $\mathbb{C}$. En esta sección se explorarán algunos métodos para estimar la lógica probabilística, se analizará la velocidad de convergencia de un secuente en ésta, y finalmente se detallará la estimación de la lógica absoluta.

Lógica probabilística

Para estimar el contenido de $\mathcal{L}_{\mathbb{C}}^{Pr}$, se determinan previamente los estimadores de cada $p_i = \mathcal{P}_{alf}(\{a_i\})$, $i = 1, ..., N$, análogo a como se establecieron

en el capitulo 4,

$$\hat{p}_i := \frac{|\{j : b_j = a_i\}|}{n}, i = 1, ..., N \tag{5.1}$$

para n observaciones $b_1, b_2, ..., b_N \in A$. Sea $\Gamma \vdash \Delta$ un secuente de $\mathbb{C}$. Se busca estimar la probabilidad asociada al secuente en $\mathcal{L}_{\mathbb{C}}^{Pr}$. De acuerdo con la Proposición 5.1 numeral 1, sea

$$q := \mathcal{P}(B_{\Gamma\vdash\Delta}) = \sum_{b\in B_{\Gamma\vdash\Delta}} \mathcal{P}_{alf}(b).$$

Entonces, la probabilidad estimada $\hat{q}_n$ que $\Gamma \vdash \Delta$ va a tener asociada en $\mathcal{L}_{\mathbb{C}}^{Pr}(n)$ será:

$$\hat{q}_n := \sum_{b\in B_{\Gamma\vdash\Delta}} \hat{\mathcal{P}}_{alf}(b),$$

donde $\mathcal{L}_{\mathbb{C}}^{Pr}(n)$ es la lógica probabilística construida al realizar n observaciones. Como la convergencia en probabilidad es cerrada bajo suma [4], y $\hat{p}_i \xrightarrow{p} p_i$ para todo $i = 1, ..., N$, luego,

$$\hat{q}_n \xrightarrow{p} q. \tag{5.2}$$

Es decir, para cada secuente, su probabilidad $\hat{q}_n$ asociada a las observaciones converge a la probabilidad real q cuando n tiende a ∞.

Ahora se estudiará la convergencia de $\mathcal{L}_{\mathbb{C}}^{Pr}(n)$.

Convergencia de $\mathcal{L}_{\mathbb{C}}^{Pr}(n)$

Como una noción de convergencia de $\mathcal{L}_{\mathbb{C}}^{Pr}(n)$ no tendría sentido desde el punto de vista de la convergencia de conjuntos, se analizará el comportamiento asintótico de la probabilidad de que un secuente $\Gamma \vdash_p \Delta$ esté en $\mathcal{L}_{\mathbb{C}}^{Pr}(n)$. Por la ecuación (5.2), se deduce que, a medida que n aumenta, la probabilidad del secuente se aproxima a su valor real. Para analizar la rapidez de convergencia se estudiará la varianza asintótica con la cual se distribuye $\sqrt{n}(\hat{q}_n - q)$, haciendo uso del Teorema del Límite Central (TMC) presentado en el teorema 4.2.

Sea $\Gamma \vdash \Delta$ un secuente de $\mathbb{C}$. Se define el conjunto

$$I := \{i \in \{1, .., N\} : a_i \in B_{\Gamma\vdash\Delta}\},$$

y el vector de unos en las entradas de I, como

$$\mathbb{1}_I := \sum_{i\in I} \hat{e}_i,$$

que es un elemento de $\mathbb{R}^N$. Se obtiene el siguiente resultado.

Teorema 5.1.

$$\sqrt{n}(\hat{q}_n - q) \xrightarrow{d} N(0, \mathbb{1}_I^t C \mathbb{1}_I). \tag{5.3}$$

donde C es a matriz de covarianza de la distribución multinomial con parámetros $(1, \bar{p})$.

Demostración. El estimador de q se puede escribir como $\hat{q}_n = \mathbb{1}_I^t \hat{\bar{p}}$ donde $\hat{\bar{p}} = (\hat{p}_1, ..., \hat{p}_N)$ es el vector obtenido en (5.1). Como $\hat{\bar{p}}$ se distribuye como una multinomial con parámetros $(1, \bar{p})$, se tiene la misma convergencia que en (4.10). Ahora, como la función $H : \mathbb{R}^n \to \mathbb{R}$, $H(\bar{x}) = \mathbb{1}_I^t \bar{x}$ es continua en $\mathbb{R}$, del Teorema del Mapa Continuo se sigue el teorema, con la matriz C como en (4.11). □

Con el comportamiento de la varianza asintótica de $\hat{q}_n$ se estudia la velocidad de convergencia, como se muestra en los siguientes ejemplos.

Ejemplo 5.3. Se construye la clasificación $\mathbb{C}$ de siguiente forma y con distribución de probabilidad $\mathcal{P}_c$ desconocida:

- $I_C = \{c_{00}, c_{01}, c_{10}, c_{11}\}$, $\mathcal{B}_C = \mathscr{P}(I_C)$.
- $T = \{T_1, T_2\}$.
- $\models_C = \{(c_{10}, T_1), (c_{01}, T_2), (c_{11}, T_1), (c_{11}, T_2)\}$.

El alfabeto de $\mathbb{C}$ es entonces $A = \{\langle 0,0\rangle, \langle 0,1\rangle, \langle 1,0\rangle, \langle 1,1\rangle\}$ y $\mathcal{P}_{alf}$ es desconocido. Denote $p_{ij} := \mathcal{P}_{alf}(\langle i,j\rangle)$ para $i, j \in \{0,1\}$. Se desea analizar la velocidad de convergencia de $T_1 \vdash_{\hat{q}_n} T_2$ al único secuente probabilístico $T_1 \vdash_q T_2$ que pertenece a $\mathcal{L}_{\mathbb{C}}^{Pr}$. Como $B_{T_1 \vdash T_2} = \{\langle 0,0\rangle, \langle 0,1\rangle, \langle 1,1\rangle\}$ entonces

$$\begin{aligned}\mathbb{1}_I^t C \mathbb{1}_I = {} & p_{00}(1-p_{00}) + p_{01}(1-p_{01}) \\ & + p_{11}(1-p_{11}) - 2p_{00}p_{01} - 2p_{00}p_{11} - 2p_{01}p_{11}.\end{aligned}$$

Como se demostrará más adelante, la máxima varianza asintótica se obtiene cuando

$$\begin{aligned}1/2 &= p_{00} + p_{01} + p_{11} \\ &= \mathcal{P}_{alf}(B_{T_1 \vdash T_2}).\end{aligned} \tag{5.4}$$

El mínimo se obtiene cuando $p_{00} + p_{01} + p_{11} = 0$ ó 1.

Ejemplo 5.4. Con la misma clasificación y notación del ejemplo anterior, se desea analizar la convergencia del secuente $T_1 \vdash_{\hat{q}_n}$, conociendo previamente que $p_{10} = 0$. Como $B_{T_1 \vdash} = \{\langle 0,0\rangle, \langle 0,1\rangle\}$ entonces

$$\mathbb{1}_I^t C \mathbb{1}_I = p_{00}(1-p_{00}) + p_{01}(1-p_{01}) - 2p_{00}p_{01}.$$

El máximo valor se obtiene cuando

$$\begin{aligned} 1/2 &= p_{00} + p_{01} \\ &= \mathcal{P}_{alf}(B_{T_1\vdash}). \end{aligned} \tag{5.5}$$

El mínimo se obtiene cuando $p_{00} + p_{01} = 0$ ó 1.

En ambos ejemplos, como muestra 5.5 y 5.4, la máxima varianza se alcanza cuando la probabilidad del conjunto $B_{\Gamma\vdash\Delta}$ es $1/2$. Así mismo el mínimo se obtiene cuando la probabilidad del conjunto $B_{\Gamma\vdash\Delta}$ es 0 ó 1. Ese resultado, que sugiere que existe una relación entre esta convergencia y la entropía binaria asociada a $\mathcal{P}(B_{\Gamma\vdash\Delta})$, se muestra en general en el siguiente teorema.

Teorema 5.2. *El problema de maximización de la varianza asintótica de p_n para $\Gamma \vdash_{p_n} \Delta$ se puede replantear como:*

$$\begin{aligned} p_{max} = \ arg\ \max_{x\in\mathcal{R}^{|I|}} \quad & f(x) := \sum_{i=1}^{|I|} x_i - \sum_{i=1}^{|I|}\sum_{j=1}^{|I|} x_i x_j \\ sujeto\ a \quad & b^t x \leq 1 \\ & 0 \leq x_1, ..., x_I \leq 1. \end{aligned} \tag{5.6}$$

donde $b \in \mathbb{R}^{|I|}$ es el vector de unos. El conjunto de soluciones del problema (5.6) es:

$$\{x \in \mathbb{R}^I : b^t x = 1/2, 0 \leq x_1, ..., x_I \leq 1\}.$$

Por el contrario, el conjunto de soluciones del problema de minimizar la varianza asintótica, es decir, minimizar (5.6) es:

$$\{x \in \mathbb{R}^I : b^t x = 0, 0 \leq x_1, ..., x_I \leq 1\} \cup \{x \in \mathbb{R}^I : b^t x = 1, 0 \leq x_1, ..., x_I \leq 1\}.$$

Demostración. Para $\Gamma \vdash_{p_n} \Delta$, la varianza asintótica de p_n que se quiere maximizar viene dada por:

$$\begin{aligned} \mathbb{1}_I^t C \mathbb{1}_I &= \sum_{i\in I}\sum_{i\in I} C_{ij} \\ &= \sum_{i\in I} p_i(1-p_i) - \sum_{i\in I}\sum_{\substack{j\in I \\ i\neq j}} p_i p_j \\ &= \sum_{i\in I} p_i - \sum_{i\in I} p_i^2 - \sum_{i\in I}\sum_{\substack{j\in I \\ i\neq j}} p_i p_j \\ &= \sum_{i\in I} p_i - \sum_{i\in I}\sum_{j\in I} p_i p_j. \end{aligned}$$

donde C_{ij} es la entrada (i, j) de la matriz de covarianzas C encontrada en la ecuación (4.11), y los p_i están sujetos a:

$$\sum_{i\in I} p_i \leq \sum_{i=1}^{N} p_i \leq 1, \text{ y } 0 \leq p_i \leq 1 \text{ para } i \in \{1, ..., N\}.$$

El problema de minimizar respecto a p_i, $i \in I$ es equivalente a minimizar respecto a un vector $x \in \mathbb{R}^{|I|}$ con restricciones

$$\sum_{i=1}^{|I|} x_i \leq 1, \text{ y } 0 \leq x_i \leq 1 \text{ para } i \in \{1, ..., N\},$$

obteniendo (5.6).

Ahora se resuelve el problema de la ecuación (5.6). Sea $g(x) := -f(x)$. Como la matriz hessiana $\nabla^2 g(x)$ es la matriz con todas las entradas igual a 2, entonces es positiva semidefinida. Luego, $g(x)$ es convexa y, por tanto, el problema de maximización de $f(x)$ es equivalente a resolver la ecuación $\nabla g(x) = 0$, [8]. Es decir,

$$\frac{\partial}{\partial x_i} g(x) = 2 \sum_{i=1}^{|I|} x_i - 1 = 0,$$

cuya solución es $b^t x = 1/2$, obteniendo que el conjunto de soluciones que maximiza el problema (5.6) es

$$\{x \in \mathbb{R}^I : b^t x = 1/2, 0 \leq x_1, ..., x_I \leq 1\}$$

como se quería. (Note que, como existen máximos en la *región factible*, es decir la región de $\mathcal{R}^{|I|}$ donde se cumplen las restricciones, el problema se pudo resolver buscando los máximos globales de $f(x)$ sin restricciones).

Por otro lado, como $f(x)$ es cóncava, el problema de minimizar (5.6) es equivalente a buscar el mínimo de $f(x)$ en los extremos de la región factible. Como la dirección de decrecimiento de la función es $\nabla f(x)$, cuyas entradas son todas iguales, el máximo se encuentra en el hiperplano $\{x \in \mathbb{R}^I : b^t x = 1\}$ o en el extremo (0,...,0). Como la función $f(x)$ es igual ambos casos, el máximo se obtiene en los conjuntos:

$$\{x \in \mathbb{R}^I : b^t x = 0, 0 \leq x_1, ..., x_I \leq 1\} \cup \{x \in \mathbb{R}^I : b^t x = 1, 0 \leq x_1, ..., x_I \leq 1\}.$$

□

Sea $\mathcal{P}_{\Gamma \vdash \Delta} := \mathcal{P}_{alf}(B_{\Gamma \vdash \Delta})$ y $\mathcal{H}_{\Gamma \vdash \Delta} := H_2(\mathcal{P}_{\Gamma \vdash \Delta})$, donde H_2 es la entropía binaria definida en (2.4). El teorema sugiere que, para un secuente $\Gamma \vdash \Delta$, la velocidad de convergencia es máxima cuando $\mathcal{P}_{\Gamma \vdash \Delta} = 1/2$, es decir, cuando $\mathcal{H}_{\Gamma \vdash \Delta}$ es máxima. Además, la velocidad es mínima cuando $\mathcal{P}_{\Gamma \vdash \Delta} = 0$ ó 1, es decir, cuando $\mathcal{H}_{\Gamma \vdash \Delta}$ es mínima. En el primer caso, como $\mathcal{H}_{\Gamma \vdash \Delta}$ es máxima, se dificulta determinar si la siguiente observación satisface $\Gamma \vdash \Delta$ o no, razón por la cual la velocidad de convergencia es lenta. En el segundo caso, como $\mathcal{H}_{\Gamma \vdash \Delta}$ es mínima, la incertidumbre de determinar si la siguiente

observación satisface $\Gamma \vdash \Delta$ o no se minimiza (si es 0 es fácil determinar que no lo satisface, y si es 1, que sí lo satisface). De esta forma se encuentra una primera relación entre las caracterizaciones cuantitativas de una fuente y las clasificaciones probabilísticas.

Lógica absoluta

Para estimar la lógica absoluta se puede utilizar el resultado obtenido para el caso probabilístico. Pero dado que puede haber instancias con probabilidad nula de ser observadas, la lógica absoluta se haría imposible de predecir mediante observaciones. Este análisis se resume en el siguiente teorema.

Teorema 5.3. *Suponga que* $\mathcal{P}_{alf}(\{e\}) > 0$ *para todo* $e \in E_{\models}$. *Entonces,*

$$\Gamma \vdash \Delta \in \mathcal{L}_{\mathbb{C}} \text{ si y sólo si } \Gamma \vdash_1 \Delta \in \mathcal{L}_{\mathbb{C}}^{Pr}.$$

Demostración. Sea $\Gamma \vdash \Delta$ un secuente. Primero suponga que $\Gamma \vdash \Delta \in \mathcal{L}_{\mathbb{C}}$. Entonces toda observación satisface $\Gamma \vdash \Delta$ y por tanto $\Gamma \vdash_1 \Delta \in \mathcal{L}_{\mathbb{C}}^{Pr}$. (Note que esta dirección no necesita el supuesto del teorema).

Ahora suponga que $\Gamma \vdash_1 \Delta \in \mathcal{L}_{\mathbb{C}}^{Pr}$. Luego, $\mathcal{P}_{alf}(B_{\Gamma\vdash\Delta}) = 1 = \mathcal{P}_{alf}(E_{\models})$. Como $\mathcal{P}_{alf}(\{e\}) > 0$ para todo $e \in E_{\models}$, entonces $B_{\Gamma\vdash\Delta} = E_{\models}$. Luego, toda observación satisface $\Gamma \vdash \Delta$. Por tanto $\Gamma \vdash \Delta \in \mathcal{L}_{\mathbb{C}}$. □

Con el teorema es posible aproximar $\mathcal{L}_{\mathbb{C}}$ mediante observaciones, tomando los secuentes $\Gamma \vdash_1 \Delta \in \mathcal{L}_{\mathbb{C}}^{Pr}$. Sin embargo, ¿qué pasa si no se asume que $\mathcal{P}_{alf}(\{e\}) > 0$ para todo $e \in E_{\models}$? En ese caso, hay secuentes que, con casi toda seguridad, se agregan a la lógica estimada $\mathcal{L}_{\mathbb{C}}$ que no son satisfechos por todos los elementos de I_C. Si $\Gamma \vdash \Delta$ es uno de estos secuentes,

$$\mathcal{P}_{alf}(\{e \in E_{\models} : e \notin B_{\Gamma\vdash\Delta}\}) = 0,$$

donde $\{e \in E_{\models} : e \notin B_{\Gamma\vdash\Delta}\} \neq \emptyset$. Es decir, las excepciones de $\Gamma \vdash \Delta$ tienen probabilidad cero de ocurrir, por tanto casi con toda seguridad no se observarán. Para este caso, no hay cómo estimar la lógica $\mathcal{L}_{\mathbb{C}}$ de forma exacta, pero la lógica encontrada va a coincidir confiablemente con las observaciones.

5.2. Entropía de la Fuente

Al igual que con una fuente en el sentido de Shannon, a la fuente de una clasificación se le puede asociar la noción de entropía, pues de hecho es un tipo especial de fuente en el sentido de Shannon.

Definición 5.7. Sea $\mathbb{C}$ una clasificación y A el universo de la fuente de $\mathbb{C}$. La *entropía de la fuente* asociada a $\mathbb{C}$ viene dada por

$$H(\mathbb{C}) = -\sum_{a \in A} \mathcal{P}_{alf}(a) \log \mathcal{P}_{alf}(a). \tag{5.7}$$

Nota: Recuerde que $0 \log 0 := 0$ como en el capítulo 2.

Como esta definición coincide con la entropía de Shannon (Definición 2.7), la entropía de una fuente adquiere las interpretaciones discutidas en el capítulo 4. En particular, que la entropía cuantifica la incertidumbre de la fuente. A continuación se discutirá una relación entre la lógica absoluta de una clasificación y la entropía máxima de la fuente asociada.

Sean $\mathbb{C}_1$ y $\mathbb{C}_2$ dos clasificaciones con el mismo conjunto $T = \langle T_1, ..., T_N \rangle$ de tipos (y mismo alfabeto A) y cuya distribución de probabilidad se desconoce. Suponga además que las lógicas absolutas $\mathcal{L}_1$ y $\mathcal{L}_2$ de $\mathbb{C}_1$ y $\mathbb{C}_2$ respectivamente, cumplen la condición $\mathcal{L}_1 \subseteq \mathcal{L}_2$. El hecho de que $\mathcal{L}_1$ tenga menos secuentes indica que hay menos restricciones que las instancias de $\mathbb{C}_1$ deban cumplir y, por lo tanto, el número de elementos de A que se pueden asociar a las observaciones es mayor. De esta discusión se puede encontrar una relación entre las entropías máximas de cada fuente, como lo presenta el teorema 5.4.

Lema 5.1. *Si $x_1, ..., x_n$ son variables aleatorias con valores en el intervalo $(0, 1]$ tales que $\sum_i x_i = 1$, la función*

$$f(x_1, ..., x_n) = -\sum_{i=1}^{n} x_i \log x_i.$$

tiene un único máximo cuando $x_i = 1/n$ para todo $i \in \{1, .., n\}$, y su valor máximo es

$$f_{max} = -\log x_1.$$

Demostración. (Idea general tomada de [3]). Si $A = \{a_1, ..., a_n\}$ y $p_i := p(a_i)$, la función f se puede expresar como $f = f(p_1, ..., p_n)$. Primero se prueba que f es estrictamente cóncava. Como la matriz Hessiana de f es

$$\mathcal{H} = \begin{cases} -1/p_i, & \text{si } i = j \\ 0, & \text{si } i \neq j \end{cases}$$

entonces $\mathcal{H}$ es definida negativa. Luego f es estrictamente cóncava y tiene un único máximo global.
Ahora, para encontrar el máximo, se define la función ϕ sobre el intervalo (0,1) tal que $\phi(x) := -\log x$ y se define la variable aleatoria X en A tal que

$X(a) = 1/p(a)$. Luego,

$$\begin{aligned} EX &= \sum_{a\in A} X(a)p(a) = \sum_{a\in A} \frac{1}{p(a)} p(a) = |A|. \\ \phi(EX) &= \phi(|A|) = -\log|A|. \\ \phi(X(a)) &= \phi(\frac{1}{p(a)}) = \log(\frac{1}{p(a)}). \\ E(\phi(X(a))) &= \sum_{a\in A} p(a) \log \frac{1}{p(a)} = -f(p_1, ..., p_n). \end{aligned}$$

Como ϕ es cóncava, por la desigualdad de Jensen[1] se obtiene que

$$\phi(EX) \leq E\phi(X).$$

Luego $f(p_1, ..., p_n) \leq \log|A|$ para cualesquiera $p_1, ..., p_n$, implicando que f es máxima en $\hat{p}_1, ..., \hat{p}_n$ si y sólo si $f(\hat{p}_1, ..., \hat{p}_n) = \log|A|$.
Si $p(a) = 1/|A|$ para todo $a \in A$,

$$f(p_1, ..., p_n) = -\sum_{i=1}^{n} p_i \log|A| = \log|A| \sum_{i=1}^{n} p_1 = \log|A|.$$

Luego el máximo de f se obtiene cuando $p(a) = 1/|A|$ para todo $a \in A$. □

Teorema 5.4. *Sean $\mathbb{C}_1$ y $\mathbb{C}_2$ clasificaciones con distribución de probabilidad desconocida, sean $\mathcal{L}_1$ y $\mathcal{L}_2$ sus lógicas absolutas respectivas y suponga que $\mathcal{L}_1 \subseteq \mathcal{L}_2$. Entonces el máximo valor $\mathcal{E}_2$ que la entropía $H(\mathbb{C}_2)$ puede tomar es menor o igual que el máximo valor $\mathcal{E}_1$ que la entropía $H(\mathbb{C}_1)$ puede tomar, i.e.*

$$\mathcal{E}_2 \leq \mathcal{E}_1.$$

Demostración. Sea E_1 el conjunto $E_{\models}$ asociado a $\mathbb{C}_1$ y E_2 el conjunto $E_{\models}$ asociado a $\mathbb{C}_2$. Primero se prueba que $E_2 \subseteq E_1$. Suponga, por contradicción, que no es cierto, i.e. existe $\langle \bar{T}_1, ..., \bar{T}_N \rangle \in E_2$ tal que $\langle \bar{T}_1, ..., \bar{T}_N \rangle \notin E_1$. Se definen los siguientes subconjuntos de tipos:

$$\Gamma := \{T_i \in T : \bar{T}_i = 1\}, \quad \Delta := \{T_j \in T : \bar{T}_j = 0\}.$$

$\langle \bar{T}_1, ..., \bar{T}_N \rangle$ es el único elemento de A tal que las instancias en $[\![\bar{T}_1, ..., \bar{T}_N]\!]$ no satisfacen $\Gamma \vdash \Delta$, razón por la cual $\Gamma \vdash \Delta \notin \mathcal{L}_2$ pero $\Gamma \vdash \Delta \in \mathcal{L}_1$, contradiciendo el supuesto $\mathcal{L}_1 \subseteq \mathcal{L}_2$. Por tanto $E_2 \subseteq E_1$ y $|E_2| \leq |E_1|$.

[1] Si ϕ es convexa en un intervalo abierto I y X es una variable aleatoria tal que $\text{supp}(\phi) \subseteq I$ y $EX < \infty$, entonces $\phi(EX) \leq E\phi(X)$ [4].

Ahora se desea probar que $\mathcal{E}_2 \leq \mathcal{E}_1$. Sean $\mathcal{P}_1$ y $\mathcal{P}_2$ las distribuciones de probabilidad de A respecto a $\mathbb{C}_1$ y $\mathbb{C}_2$ respectivamente. Aunque $\mathcal{P}_1$ y $\mathcal{P}_2$ son desconocidas, $\mathcal{P}_1(e) = 0$ si $e \notin E_1$ y $\mathcal{P}_2(e) = 0$ si $e \notin E_2$. Entonces

$$H(\mathbb{C}_1) = -\sum_{a \in A} \mathcal{P}_1(a) \log \mathcal{P}_1(a) = -\sum_{a \in E_1} \mathcal{P}_1(a) \log \mathcal{P}_1(a),$$

y similarmente

$$H(\mathbb{C}_2) = -\sum_{a \in A} \mathcal{P}_2(a) \log \mathcal{P}_2(a) = -\sum_{a \in E_2} \mathcal{P}_2(a) \log \mathcal{P}_2(a).$$

Por el lema 5.1, $H(\mathbb{C}_1)$ es máxima si y sólo si $\mathcal{P}_1(a) = 1/|E_1|$ para todo $a \in E_1$ y, en ese caso, $\mathcal{E}_1 = \log |E_1|$. De igual forma $\mathcal{E}_2 = \log |E_2|$. Por tanto $\mathcal{E}_2 \leq \mathcal{E}_1$. □

Del teorema se puede concluir que entre más fuerte sea la lógica de una clasificación, es menor la incertidumbre máxima sobre las observaciones emitidas por la fuente. Cada secuente perteneciente a la lógica contiene información sobre qué instancias pueden ser observadas, pues impone una restricción sobre los secuentes que cada instancia debe satisfacer. Esta idea se puede entender en términos de cuántos elementos del alfabeto se observan: si hay más secuentes son menos los elementos del alfabeto tales que su extensión sea no vacía.

Capítulo 6

Conclusiones

El desarrollo de la teoría de información desde los marcos cuantitativo y cualitativo ha sido objeto de estudio en los últimos años y constituye una herramienta muy útil en el almacenamiento y procesamiento de información. Dentro de este contexto, se realizó un análisis de los conceptos básicos de ambos esquemas y se determinó el comportamiento de la entropía.

En primer lugar, se hizo una breve introducción a los conceptos básicos de teoría de la probabilidad necesarios para abordar las herramientas utilizadas a lo largo del documento. A partir de estas nociones se demostró la ley de grandes números para el caso i.i.d. haciendo uso de la teoría de la medida. Con estas herramientas se hizo un análisis de la entropía obteniendo los siguientes resultados:

- La entropía de Shannon se puede interpretar como la incertidumbre asociada a fuente.
- A medida que el número de observaciones aumenta, los bits por carácter requeridos para codificar la información de esas observaciones proveniente de una fuente se aproxima a la entropía.
- Si se desconoce la entropía de una fuente, ésta se puede evaluar estimando la probabilidad de cada elemento del alfabeto.
- Este estimador de la entropía converge al valor real a medida que aumentan las observaciones.
- El error de esta aproximación está descrito por una curva normal, cuya varianza se pudo calcular.
- En el caso binario, las predicciones se pudieron comprobar mediante simulaciones, analizando el caso especial de $p = 1/2$.

En segundo lugar, basado en el análisis cualitativo de Barwise y Seligman, se introdujeron los conceptos de clasificación probabilística, fuente asociada a una clasificación, lógica absoluta y la lógica probabilística. Cuando

se desconoce la distribución de la fuente, se estudió cómo estimar las lógicas asociadas, obteniendo los siguientes resultados:

- La probabilidad estimada de un secuente de estar en la lógica probabilística converge a la probabilidad real y su error se comporta como una normal con varianza conocida.
- La velocidad de convergencia es mínima cuando la entropía asociada al conjunto B asociado al secuente es máxima.
- La velocidad de convergencia es máxima cuando la entropía asociada al conjunto B asociado al secuente es mínima.
- Los anteriores hechos brindan una relación entre la estimación de las lógicas y la entropía.
- La lógica absoluta se puede estimar a partir de la lógica probabilística si todos los conjuntos de instancias medibles tienen probabilidad no nula.
- Existe una relación entre el tamaño de la lógica y la entropía máxima.

La teoría de la información puede ser abordada desde dos perspectivas distintas (cualitativa y cuantitativa), haciendo posible encontrar una correlación entre los conceptos trabajados por ambas. Esta relación puede ser el punto de partida para construir una teoría que unifique ambos esquemas. En el futuro se puede continuar el estudio de estimadores para casos más generales, por ejemplo, sin suponer que la fuente se distribuye i.i.d., abarcando casos reales. Asi mismo, todavía hace falta analizar las consecuencias de introducir ideas cuantitativas en el marco de Barwise y Seligman.

Bibliografía

[1] C. E. Shannon. *Bell Syst. Tech. J.*, **27**: 379, (1948).

[2] R.M. Dudley. *Real Analysis and Probability.* (Cambridge University Press, 2002).

[3] D.J.C. Mackay. *Information Theory, Inference, and Learning Algorithms, 2nd edition.* (Cambridge University Press, 2003).

[4] R. Hogg, J. McKean, and A. Craig. *Introduction to Mathematical Statistics, 6th edition.* (Pearson, 2004).

[5] D. Pollard. *Convergence of Stochastic Processes.* (Springer, 1984).

[6] P. Billingsley. *Convergence of Probability Measures, 2th edition.* (John Wiley & Sons, 1999).

[7] J. Barwise and J. Seligman. *Information Flow: The Logic of Distributed Systems.* (Cambridge University Press, 1997).

[8] S. Boyd and Lieven Vandenberghe. *Convex Optimization.* (Cambridge University Press, 2004).