Expectation and Prediction

Suppose you want to predict the value of a random variable X. What is the best predictor of X? To define "best" you must decide on a criterion and a class of predictors. The simplest prediction problem is to predict the value of X by a constant, say b. Think in terms of losing some amount $L(x, b)$ if you predict b and the value of X is actually x. The function $L(x, b)$ is called a *loss function* in decision theory. It seems reasonable to try to pick b so as to minimize the *expected loss*, or *risk* $r(b) = E[L(X, b)]$

Example 12. **Right or wrong.**

Suppose that $L(x, b) = 0$ if $x = b$, and 1 otherwise. So you are penalized nothing if you get the value of X right, and penalized by one unit if you get the value of X wrong.

Problem. What is the best predictor?

Solution. $E[L(X, b)] = 0P(X = b) + 1P(X \neq b) = 1 - P(X = b)$.
So choosing b to minimize expected loss for this loss function is the same as choosing b to maximize $P(X = b)$. That is to say, b should be a mode of the distribution of X. Many probability distributions have a unique mode. But every possible value of a uniformly distributed random variable is a mode.

Example 13. **Absolute error.**

Suppose $L(x, b) = |x - b|$. So the penalty is the absolute value of the difference between the actual value and the predicted value. Now there is a bigger penalty for bigger mistakes. The expected loss is

$$r(b) = E(|X - b|) = \sum_x |x - b| P(X = x)$$

by the formula for $E[g(X)]$ applied to $g(x) = |x - b|$ for fixed b.

Problem. Find b that minimizes $r(b)$.

Solution. This time the solution is the median. To see why, look for a fixed x at the derivative

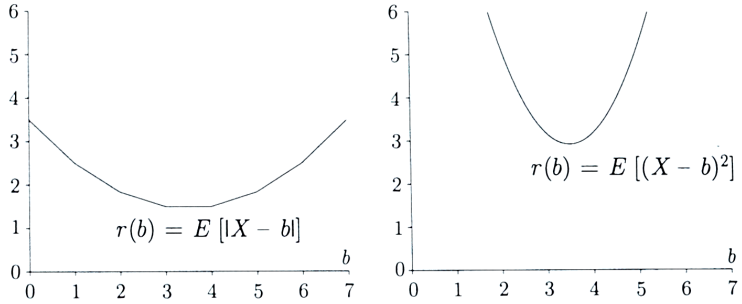
$$\frac{d}{db}|x - b| = \begin{cases} -1 & \text{if } b < x \\ 1 & \text{if } b > x \end{cases}$$

The sum defining $r(b)$ is over all possible values of X, say $x_1 < x_2 < \cdots < x_n$. So provided that $b \neq x_k$ for any k, the function $r(b)$ has the derivative

$$\begin{aligned} \frac{dr(b)}{db} &= \sum_{x<b} 1P(X = x) + \sum_{x>b} (-1)P(X = x) \\ &= P(X < b) - P(X > b) \\ &= 2P(X \leq x_k) - 1 \qquad \text{if } x_k < b < x_{k+1} \end{aligned}$$

So the function $r(b)$ is piecewise linear for b between x_k and x_{k+1}, decreasing if $P(X \leq x_k) < 1/2$, increasing if $P(X \leq x_k) > 1/2$, and flat if $P(X \leq x_k) = 1/2$. So a b is minimizing if and only if $P(X \leq b) \geq 1/2$ and $P(X \geq b) \geq 1/2$. Such a value b is a *median* of the distribution of X. A median always exists, but it may not be unique.

FIGURE 3. Risk functions for a die roll X with uniform distribution on $\{1, \ldots, 6\}$.
Left: Graph of the risk function $r(b) = E(|X - b|)$ for absolute error. (Refer to Example 13.) In this example, every number in the interval $[3, 4]$ is a median for X. Numbers in this interval are equally good as predictors of X according to the criterion of minimizing the expected absolute error, and better than any other number. Right: The risk function $r(b) = E[(X - b)^2]$ for quadratic loss function. (Refer to Example 14.) Now $E(X) = 3.5$ is the unique best predictor.



Example 14. **Squared error.**

Suppose now the penalty is *squared error*, using the *quadratic loss function* $L(x, b) = (x - b)^2$.

Problem. Find b that is the best constant predictor of X for this quadratic loss function.

Solution. This time the answer is just the mean. Now

$$\begin{aligned} r(b) &= E[(X - b)^2] = E(X^2) - 2bE(X) + b^2 \\ \frac{dr(b)}{db} &= -2E(X) + 2b \end{aligned}$$

so $b = E(X)$ gives the *unique* best predictor of X for the quadratic loss function.