1 Probability

1.1 Linearity of expectation

For any random variable $X$ and constants $a$ and $b$:

$$E[a + bX] = a + bE[X]$$

For any random variables of $X$ and $Y$, whether independent or not:

$$E[X + Y] = E[X] + E[Y]$$

Recall the definition of variance:

$$Var[X] = E \left[(X - E[X])^2\right]$$

Now let’s define $Y = a + bX$ and show that $Var[Y] = b^2 Var[X]$:

$$E[Y] = a + bE[X] \quad \text{by linearity of expectation}$$

Now we can derive the variance:

$$Var[Y] = E \left[(Y - E[Y])^2\right] \quad \text{definition of variance}$$

$$= E \left[((a + bX) - (a + bE[X]))^2\right]$$

$$= E \left[b^2(X - E[X])^2\right]$$

$$= b^2 E \left[(X - E[X])^2\right] \quad \text{linearity of expectation}$$

$$= b^2 Var[X] \quad \text{definition of variance}$$

This is why we often use the standard deviation (the square root of variance), because $StdDev[Y] = b StdDev[X]$, which is more intuitive.
1.2 Prediction, and expectation, and partial derivatives

Suppose we want to predict a random variable $Y$ simply using some constant $c$. What value of $c$ should we choose? Here we show that $E[Y]$ is a sensible choice.

But first, we need to decide what a good prediction should look like. A common choice is the mean-squared error, or MSE. We punish our prediction ever more harshly the further it gets from the observed $Y$.

\[ \text{MSE} = E \left[ (Y - c)^2 \right] \]

We now show that MSE is minimized at $E[Y]$. We set it up as an optimization problem:

\[
\begin{align*}
\min_c E \left[ (Y - c)^2 \right] \\
= \min_c E \left[ Y^2 - 2E[Y]c + c^2 \right] \\
= \min_c E[Y^2] - 2E[Y]c + c^2
\end{align*}
\]

This is a quadratic function of $c$. We can find the minimum of this quadratic by setting its partial derivative to 0, and solving for $c$:

\[
\frac{\partial}{\partial c} \left[ E[Y^2] - 2E[Y]c + c^2 \right] = 0 \\
-2E[Y] + 2c = 0 \\
\hspace{1cm} c = E[Y] \quad \text{This minimizes the MSE!}
\]

1.3 Sample mean and the Central Limit Theorem

Suppose we have $n$ random variables $X_1, \ldots, X_n$ that are independent and identically distributed (iid). Suppose we don’t know what the distribution is, but we do know their expectation and variance:

\[ E[X_i] = \mu \quad \text{and} \quad \text{Var}[X_i] = \sigma^2 \quad \text{for } i = 1, \ldots, n \]

A common way to estimate the unknown $\mu$ is to use the average (sample mean) of our data:

\[ \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \]

How does this estimate behave? We can characterize its behavior by deriving its expectation and variance.

\[
\begin{align*}
E[X_n] &= E \left[ \frac{X_1 + \cdots + X_n}{n} \right] \\
&= \frac{E[X_1] + \cdots + E[X_n]}{n} \quad \text{linearity of expectation} \\
&= \frac{n\mu}{n} = \mu
\end{align*}
\]
This tells us that $\bar{X}_n$ is “unbiased” - its expected value is the true mean.

$$\text{Var} [\bar{X}_n] = \text{Var} \left[ \frac{X_1 + \cdots + X_n}{n} \right]$$

$$= \frac{1}{n^2} \text{Var} \left[ X_1 + \cdots + X_n \right]$$

$$= \frac{1}{n^2} \left( \text{Var}[X_1] + \cdots + \text{Var}[X_n] \right)$$ only because $X_i$ are iid - variance isn’t linear!

$$= \frac{1}{n^2} (n \text{Var}[X_i]) = \frac{\sigma^2}{n}$$

This tells us that the variance of the average decreases as $n$ the number of samples increases.

But it turns out we know something more about the distribution of $\bar{X}_n$. It’s distribution actually converges to a Normal distribution as $n$ gets large. This is called the Central Limit Theorem:

$$\bar{X}_n \xrightarrow{} \mathcal{N}(\mu, \frac{\sigma^2}{n})$$

2 Linear Algebra

I discussed problems taken directly from Section 4 of Linear Algebra Review. Two other great online resources:

- YouTube tutorial on gradients
- Matrix Cookbook reference