Unit 5 Homework: Learning from Random Samples

w203: Statistics for Data Science

Applied Practice

1. Safety First through Statistics

Suppose the strength of a particular metal beam is given by,

$$S = 5 + 2T^2 \cdot P$$

Where T is a random variable representing the forging temperature and P is a random variable representing purity. Suppose the following statements are true about these random variables:

- T has a uniform distribution on $[0, 2]$.
- Conditional on a value for T, P is has a normal distribution with mean $T/2$ and standard deviation $T/12$.

For example, if $T = 1$, then,

$$E[P|T = 1] = \frac{1}{2}, \text{and,}$$
$$\sigma[P|T = 1] = \sqrt{V[P|T = 1]} = \frac{1}{12}$$

1. (6 points) Compute the expectation of S.

2. Comparing Estimators

Say that $\{X_1, ..., X_n\}$ is an i.i.d. sample from an exponential distribution, with common density,

$$f_X(x) = \lambda e^{-\lambda x}$$

You don't have to derive (we'll tell you) that, $E[X] = \int x \cdot f_X(x)dx$, the expectation of X is $1/\lambda$ and variance $1/\lambda^2$. You are considering the following estimators.

1. $\hat{\theta}_1 = \frac{X_1+X_2+...+X_n}{n}$
2. $\hat{\theta}_2 = \frac{X_1+X_2}{2}$
3. $\hat{\theta}_3 = \frac{X_1+X_2+...+X_n}{n+1}$

a. (3 points) Compute the bias of each estimator, $E[\hat{\theta}_i] - E[X]$.
b. (3 points) Compute the sampling variance of each estimator.
c. (3 points) Compute the MSE of each estimator.

d. (3 points) Explain in your own words, why estimator 3 has the highest bias, but the lowest MSE.

Proof Practice

3. Best Linear Predictor of a Constrained Outcome Space

Suppose that discrete random variables X and Y have joint probability mass function given by:

$$f(x,y) = \begin{cases} 1/2, & (x,y) \in \{(0,0),(2,1)\} \\ 0, & \text{otherwise} \end{cases}$$

(This means that there is equal probability that the points $(0,0)$ and $(2,1)$ are drawn; there is zero probability that any other point is drawn.)

Let $g(x) = \beta_0 + \beta_1 x$ be a predictor for y that is a function of x, and define the error, ϵ, to be the difference between the true value of y and the prediction $g(x)$, $\epsilon = Y - g(X)$.

1. (1 point) If you impose the moment condition (that is, you *require* that), $E[\epsilon] = 0$, what one point in the plane must the predictor pass through? (In some places, this point is referred to as the *grand mean.*)
2. (2 points) Because we have defined $\epsilon = Y - g(X)$, we can ask the question, "*What is the covariance between X and ϵ?*"

Because how how we have defined ϵ, we can know that the answer probably starts with a substitution:

$$Cov[X,\epsilon] = Cov[X, Y - g(x)]$$

Assume (or you might say, "require") that the expected value ϵ is zero, $E[\epsilon] = 0$. Then, prove that $cov[X,\epsilon]$ has the form $a + b\beta_1$.

Given the constraints of the *pdf*, $f(x,y)$, you have been provided, what is the specific value of b?

3. (2 points) How is the sign of $cov[X,\epsilon]$ is related to the angle of the line.
4. (2 points) Compute the BLP in this way:

a. Assume (or you might say require) that $E[\epsilon] = 0$.

b. Then, set $Cov[X,\epsilon] = 0$ and solve for β_1.

What is the value of β_1?

4. Think of a Friendly Type of Function

Let T_i be a sequence of discrete random variables for $i \in \{1, 2, 3, ...\}$. Suppose that T_i has the pmf,

$$f_i(t) = \begin{cases} 1/2, & t = \frac{1}{i} \\ 1/2, & t = -\frac{1}{i} \\ 0, & \text{otherwise} \end{cases}$$

Define $g : \mathbb{R} \to \mathbb{R}$ by $g(t) = t^2 + e^t$.

(4 points total) a. Prove that $\text{plim}_{n\to\infty} T_n = 0$ b. Prove that $\text{plim}_{n\to\infty} g(T_n) = 1$, *without computing the distribution of $g(T_n)$.*

Note: Maximum score on any homework is 100%