Statistical Learning for Public Policy I

Problem Set 1

Oct 6, 2022

In this problem set we use paper-and-pencil calculations to revisit some fundamental concepts in statistics and construct predictions using nearest neighbor averaging. The table below provides a mini data set containing six observations, three predictors, and an outcome variable.

Obs	X_1	X_2	X_3	Y
1	0	3	0	4.4
2	2	0	0	1.4
3	0	1	3	4.4
4	0	1	2	3.6
5	-1	0	1	0.4
6	1	1	1	2.9

Question Set 1

a) Provide an estimate for the expectation of Y ($\hat{\mu}$) and the standard error, $SE(\hat{\mu})$.

b) Construct a 95% confidence interval for the sample mean.

Question Set 2

Suppose we wish to use the mini data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

b) What is your prediction with K = 1?

c) What is your prediction with K = 3?

Question Set 3

The data generating process (DGP) for the mini data set is defined as:

$$Y \sim N(X_1 + X_2 + X_3, 1),$$

where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2.

a) What is $E(Y|X_1 = 0, X_2 = 0, X_3 = 0)$, i.e., the expected value of Y given $X_1 = X_2 = X_3 = 0$?

b) The table below provides some test data. Using this data, compute the mean squared prediction error for the predictions in Question 2b-c.

Obs	X_1	X_2	X_3	Y
7	0	0	0	1.5

Question Set 4

Let Y be a random variable with a probability density function (p.d.f.) such that:

$$f(Y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(Y-X^2)^2},$$

or $Y \sim N\big(X^2, 1\big)$. Let X be another random variable with a categorical probability mass function (p.m.f.) such that:

$$h(X) = \begin{cases} \text{1/3 for X= -1} \\ \text{1/3 for X= 0} \\ \text{1/3 for X= 1} \end{cases}$$

.

a) State the (piece-wise) conditional expectation function $E(Y|X)$.

b) State the unconditional expectation $E(Y)$.