Statistical Learning

Homework 1

Part 1: Conceptual Questions

Problem 1. Let (X, Y) be a pair of random variables. Show that if $\mathrm{E}[Y \mid X] = \mathrm{E}[Y]$, then $\mathrm{Cov}[X, Y] = 0$.

Solution. By law of iterated expectations (LIE), $\mathrm{E}[YX] = \mathrm{E}[\mathrm{E}[YX \mid X]] = \mathrm{E}[X \cdot \mathrm{E}[Y \mid X]] = \mathrm{E}[X \cdot \mathrm{E}[Y]] = \mathrm{E}[X]\,\mathrm{E}[Y]$.

Problem 2. Let (X, Y) be a pair of random variables. Denote $f(X) = \mathrm{E}[Y \mid X]$. Show that for any function g,

$$\mathrm{E}\left[(Y - f(X))^2 \mid X\right] \leq \mathrm{E}\left[(Y - g(X))^2 \mid X\right].$$

Hint: write

$$\mathrm{E}\left[(Y - g(X))^2 \mid X\right] = \mathrm{E}\left[(Y - f(X) + f(X) - g(X))^2 \mid X\right]$$

and use the law of iterated expectations (LIE).

Solution. By LIE,

$$\begin{aligned}
\mathrm{E}\left[(Y - g(X))^2 \mid X\right] &= \mathrm{E}\left[(Y - f(X) + f(X) - g(X))^2 \mid X\right] \\
&= \mathrm{E}\left[(Y - f(X))^2 \mid X\right] + (f(X) - g(X))^2 \\
&\quad +2 \cdot \mathrm{E}\left[(Y - f(X))(f(X) - g(X)) \mid X\right].
\end{aligned}$$

Note that

$$\begin{aligned}
\mathrm{E}\left[(Y - f(X))(f(X) - g(X)) \mid X\right] &= (f(X) - g(X))\,\mathrm{E}[Y - f(X) \mid X] \\
&= (f(X) - g(X))(\mathrm{E}[Y \mid X] - f(X)) \\
&= 0.
\end{aligned}$$

Then,

$$\mathrm{E}\left[(Y - g(X))^2 \mid X\right] = \mathrm{E}\left[(Y - f(X))^2 \mid X\right] + (f(X) - g(X))^2 \geq \mathrm{E}\left[(Y - f(X))^2 \mid X\right].$$

Problem 3. Given training data $\mathsf{Tr} = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ and a predictor $\hat{f}(x)$ which depends on Tr for any x, we have a new observation (X_0, Y_0) that is independent from Tr. Suppose that (X_0, Y_0) is generated by the model $Y_0 = f(X_0) + \epsilon_0$ with ϵ_0 being a new error term that is independent from (X_0, Tr). Show that the conditional expected test MSE can be decomposed into

$$\mathrm{E}\left[\left(Y_0 - \hat{f}(X_0)\right)^2 \mid X_0\right] = \mathrm{Var}[\epsilon] + \mathrm{Bias}(X_0)^2 + \mathrm{Variance}(X_0)$$

where $\text{Bias}(X_0) = \mathrm{E}\left[\hat{f}(X_0) \mid X_0\right] - f(X_0)$ and

$$\text{Variance}(X_0) = \text{Var}\left[\hat{f}(X_0) \mid X_0\right] = \mathrm{E}\left[\left(\hat{f}(X_0) - \mathrm{E}\left[\hat{f}(X_0) \mid X_0\right]\right)^2 \mid X_0\right].$$

Hint: by LIE, write

$$\begin{aligned} \mathrm{E}\left[\left(Y_0 - \hat{f}(X_0)\right)^2 \mid X_0\right] &= \mathrm{E}\left[\mathrm{E}\left[\left(Y_0 - \hat{f}(X_0)\right)^2 \mid X_0, \mathsf{Tr}\right] \mid X_0\right] \\ &= \mathrm{E}\left[\mathrm{E}\left[\left(Y_0 - f(X_0) + f(X_0) - \hat{f}(X_0)\right)^2 \mid X_0, \mathsf{Tr}\right] \mid X_0\right]. \end{aligned}$$

You may use the result $\mathrm{E}[Y_0 \mid \mathsf{Tr}, X_0] = \mathrm{E}[Y_0 \mid X_0]$ without proving it.

Solution. By LIE and simple algebra,

$$\begin{aligned} \mathrm{E}\left[\left(Y_0 - \hat{f}(X_0)\right)^2 \mid X_0\right] &= \mathrm{E}\left[\mathrm{E}\left[(Y_0 - f(X_0))^2 \mid X_0, \mathsf{Tr}\right] \mid X_0\right] \\ &\quad + \mathrm{E}\left[\mathrm{E}\left[\left(f(X_0) - \hat{f}(X_0)\right)^2 \mid X_0, \mathsf{Tr}\right] \mid X_0\right] \\ &\quad + 2 \cdot \mathrm{E}\left[\mathrm{E}\left[(Y_0 - f(X_0))\left(f(X_0) - \hat{f}(X_0)\right) \mid X_0, \mathsf{Tr}\right] \mid X_0\right]. \end{aligned}$$

Then,

$$\mathrm{E}\left[\mathrm{E}\left[(Y_0 - f(X_0))^2 \mid X_0, \mathsf{Tr}\right] \mid X_0\right] = \mathrm{E}\left[\mathrm{E}\left[\epsilon_0^2 \mid X_0, \mathsf{Tr}\right] \mid X_0\right] = \mathrm{E}\left[\epsilon_0^2\right] = \text{Var}[\epsilon],$$

since ϵ_0 is independent from (X_0, Tr). Note that this implies that ϵ_0^2 is also independent from (X_0, Tr) and therefore, ϵ_0^2 is mean independent from (X_0, Tr): $\mathrm{E}[\epsilon_0^2 \mid X_0, \mathsf{Tr}] = \mathrm{E}[\epsilon_0^2] = \text{Var}[\epsilon]$. For the third term,

$$\begin{aligned} &\mathrm{E}\left[\mathrm{E}\left[(Y_0 - f(X_0))\left(f(X_0) - \hat{f}(X_0)\right) \mid X_0, \mathsf{Tr}\right] \mid X_0\right] = \\ &\qquad \mathrm{E}\left[\left(f(X_0) - \hat{f}(X_0)\right)\mathrm{E}\left[(Y_0 - f(X_0)) \mid X_0, \mathsf{Tr}\right] \mid X_0\right] = \\ &\qquad\qquad \mathrm{E}\left[\left(f(X_0) - \hat{f}(X_0)\right)\mathrm{E}\left[\epsilon_0 \mid X_0, \mathsf{Tr}\right] \mid X_0\right] = 0, \end{aligned}$$

since $\mathrm{E}[\epsilon_0 \mid X_0, \mathsf{Tr}] = \mathrm{E}[\epsilon_0] = 0$. For the second term,

$$\begin{aligned} &\mathrm{E}\left[\mathrm{E}\left[\left(f(X_0) - \hat{f}(X_0)\right)^2 \mid X_0, \mathsf{Tr}\right] \mid X_0\right] = \mathrm{E}\left[\left(f(X_0) - \hat{f}(X_0)\right)^2 \mid X_0\right] \\ &\quad = \mathrm{E}\left[\left(f(X_0) - \mathrm{E}\left[\hat{f}(X_0) \mid X_0\right] + \mathrm{E}\left[\hat{f}(X_0) \mid X_0\right] - \hat{f}(X_0)\right)^2 \mid X_0\right] \\ &= \text{Bias}(X_0)^2 + \text{Variance}(X_0) + 2 \cdot \mathrm{E}\left[\left(f(X_0) - \mathrm{E}\left[\hat{f}(X_0) \mid X_0\right]\right)\left(\mathrm{E}\left[\hat{f}(X_0) \mid X_0\right] - \hat{f}(X_0)\right) \mid X_0\right]. \end{aligned}$$

The conclusion follows from

$$\begin{aligned} &\mathrm{E}\left[\left(f(X_0) - \mathrm{E}\left[\hat{f}(X_0) \mid X_0\right]\right)\left(\mathrm{E}\left[\hat{f}(X_0) \mid X_0\right] - \hat{f}(X_0)\right) \mid X_0\right] = \\ &\qquad \left(f(X_0) - \mathrm{E}\left[\hat{f}(X_0) \mid X_0\right]\right)\mathrm{E}\left[\mathrm{E}\left[\hat{f}(X_0) \mid X_0\right] - \hat{f}(X_0) \mid X_0\right] = \\ &\qquad\qquad \left(f(X_0) - \mathrm{E}\left[\hat{f}(X_0) \mid X_0\right]\right)\left(\mathrm{E}\left[\hat{f}(X_0) \mid X_0\right] - \mathrm{E}\left[\hat{f}(X_0) \mid X_0\right]\right) = 0. \end{aligned}$$

Problem 4. Suppose that Y is a binary response variable. The range of values taken by Y is $\{0,1\}$. The goal is to predict Y given another random variable X. When we observe a new X, we predict Y to be $h(X)$, where $h:\mathbb{R}\to\{0,1\}$ is a function that takes 0 or 1. We call h a classification rule. The "classification risk" of h is

$$R(h) = \Pr(Y \neq h(X)).$$

Let $m(x) = \mathrm{E}[Y \mid X = x]$. Since Y is binary,

$$\mathrm{E}[Y \mid X = x] = 1 \times \Pr(Y = 1 \mid X = x) + 0 \times \Pr(Y = 0 \mid X = x) = \Pr(Y = 1 \mid X = x).$$

(You may assume X is discrete if you have difficulty making sense of "$\Pr(Y = 1 \mid X = x)$". This is like $\Pr(A \mid B)$ with A being the event "$Y = 1$" and B being the event $X = x$). Show that the rule that minimizes $R(h)$ is

$$h^*(x) = \begin{cases} 1 & \text{if } m(x) > \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}$$

Hint: Note that

$$R(h) = \Pr(Y \neq h(X)) = \int \Pr(Y \neq h(x) \mid X = x) f_X(x)\, dx,$$

where the second equality follows from LIE. It suffices to show that

$$\Pr(Y \neq h(x) \mid X = x) - \Pr(Y \neq h^*(x) \mid X = x) \geq 0 \text{ for all } x.$$

Use $\Pr(Y \neq h(x) \mid X = x) = 1 - \Pr(Y = h(x) \mid X = x)$ and

$$\Pr(Y = h(x) \mid X = x) = h(x) \Pr(Y = 1 \mid X = x) + (1 - h(x)) \Pr(Y = 0 \mid X = x).$$

Solution. Note:

$$\begin{aligned} \Pr(Y \neq h(x) | X = x) &= 1 - \Pr(Y = h(x) | X = x) \\ &= 1 - [h(x) \Pr(Y = 1 | X = x) + (1 - h(x)) \Pr(Y = 0 | X = x)] \\ &= 1 - [h(x) m(x) + (1 - h(x))(1 - m(x))]. \end{aligned}$$

Therefore,

$$\begin{aligned} &\Pr(Y \neq h(x) | X = x) - \Pr(Y \neq h^*(x) | X = x) \\ &= [h^*(x) m(x) + (1 - h^*(x))(1 - m(x))] - [h(x) m(x) + (1 - h(x))(1 - m(x))] \\ &= (2m(x) - 1)(h^*(x) - h(x)) \\ &= 2\left(m(x) - \frac{1}{2}\right)(h^*(x) - h(x)). \end{aligned}$$

When $m(x) \geq 1/2$ and $h^*(x) = 1$, $\left(m(x) - \frac{1}{2}\right)(h^*(x) - h(x))$ must be non-negative, since $h(x) = 1$ or $h(x) = 0$. When $m(x) < 1/2$ and $h^*(x) = 0$, $\left(m(x) - \frac{1}{2}\right)(h^*(x) - h(x))$ is again non-negative.

Problem 5. Let $\{x_i : i = 1, \ldots, n\}$ and $\{y_i : i = 1, \ldots, n\}$ be two sequences. Define the averages

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$
$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

1. Show that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.
2. Using the result in part (1), show that

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i (x_i - \bar{x}), \text{ and}$$
$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} y_i (x_i - \bar{x}) = \sum_{i=1}^{n} x_i (y_i - \bar{y}).$$

Solution. (a)

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n \cdot \bar{x} - n \cdot \bar{x} = 0,$$

because $\sum_{i=1}^{n} x_i = n \cdot \bar{x}$. (b)

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 - \sum_{i=1}^{n} x_i (x_i - \bar{x}) &= \sum_{i=1}^{n}\left[(x_i - \bar{x})^2 - x_i (x_i - \bar{x})\right] \\ &= \sum_{i=1}^{n}\left[\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) - \left(x_i^2 - x_i\bar{x}\right)\right] \\ &= \sum_{i=1}^{n}\left(\bar{x}^2 - x_i\bar{x}\right) \\ &= \bar{x}\sum_{i=1}^{n}(\bar{x} - x_i) \\ &= 0,\end{aligned}$$

where the last equality follows from $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ proved in (a).

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^{n}(x_i - \bar{x})\, y_i - \sum_{i=1}^{n}(x_i - \bar{x})\, \bar{y} \\ &= \sum_{i=1}^{n}(x_i - \bar{x})\, y_i - \bar{y}\sum_{i=1}^{n}(x_i - \bar{x}) \\ &= \sum_{i=1}^{n}(x_i - \bar{x})\, y_i,\end{aligned}$$

where the last equality follows from $\sum_{i=1}^{n}(x_i-\bar{x})=0$ proved in (a). The proof of

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}(y_i-\bar{y})x_i$$

is similar.

Problem 6. Given training data $\mathsf{Tr}=\{(X_1,Y_1),...,(X_n,Y_n)\}$, suppose that $Y_i=\beta_0+\beta_1 X_i+\epsilon_i$ where ϵ_i is the error term. The simple regression coefficient presented in class is

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)\left(X_i-\overline{X}\right)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}.$$

Denote $X_1^n=(X_1,...,X_n)$ for notational simplicity. Assume that $\mathrm{E}\left[\epsilon_i \mid X_1^n\right]=0$, $\mathrm{E}\left[\epsilon_i^2 \mid X_1^n\right]=\sigma^2$ (for some $\sigma^2>0$) and $\mathrm{E}\left[\epsilon_i\epsilon_j \mid X_1^n\right]=0$, $\forall i$ and $\forall j\neq i$.

1. Use the result in the last problem, show that

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)Y_i}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}.$$

2. Show that $\mathrm{E}\left[\hat{\beta}_1 \mid X_1^n\right]=\beta_1$ and $\mathrm{Var}\left[\hat{\beta}_1 \mid X_1^n\right]=\sigma^2/\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$, where $\overline{X}=n^{-1}\sum_{i=1}^{n}X_i$.

3. Assume that the conditional distribution of ϵ_i given X_1^n is $\mathrm{N}(0,\sigma^2)$. What is the conditional distribution of Y_i given X_1^n?

4. What is the conditional distribution of $\hat{\beta}_1$ given X_1^n?

5. What is the unconditional distribution of the z-statistic:

$$\frac{\hat{\beta}_1-\beta_1}{\sqrt{\sigma^2/\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}}?$$

Solution. For 1, use $\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}y_i(x_i-\bar{x})$. For 2, use

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)Y_i}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\\ &= \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)(\beta_0+\beta_1X_i+\epsilon_i)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\\ &= \beta_1+\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\epsilon_i}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}.\end{aligned}$$

Then,

$$\mathrm{E}\left[\hat{\beta}_1 \mid X_1^n\right]=\beta_1+\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\mathrm{E}\left[\epsilon_i \mid X_1^n\right]}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}=\beta_1.$$

Also,

$$\mathrm{Var}\left[\hat{\beta}_1 \mid X_1^n\right] = \mathrm{E}\left[\left(\hat{\beta}_1 - \mathrm{E}\left[\hat{\beta}_1 \mid X_1^n\right]\right)^2 \mid X_1^n\right] = \mathrm{E}\left[\left\{\frac{\sum_{i=1}^n \left(X_i - \overline{X}\right)\epsilon_i}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}\right\}^2 \mid X_1^n\right]$$

$$= \frac{1}{\left\{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2\right\}^2}\mathrm{E}\left[\left\{\sum_{i=1}^n \left(X_i - \overline{X}\right)\epsilon_i\right\}^2 \mid X_1^n\right]$$

$$= \frac{1}{\left\{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2\right\}^2}\mathrm{E}\left[\sum_{i=1}^n\sum_{j=1}^n \left(X_i - \overline{X}\right)\left(X_j - \overline{X}\right)\epsilon_i\epsilon_j \mid X_1^n\right]$$

$$= \frac{1}{\left\{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2\right\}^2}\left\{\mathrm{E}\left[\sum_{i=1}^n \left(X_i - \overline{X}\right)^2\epsilon_i^2 \mid X_1^n\right] + \mathrm{E}\left[\sum_{i=1}^n\sum_{j\neq i} \left(X_i - \overline{X}\right)\left(X_j - \overline{X}\right)\epsilon_i\epsilon_j \mid X_1^n\right]\right\}.$$

Now $\mathrm{Var}\left[\hat{\beta}_1 \mid X_1^n\right] = \sigma^2/\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$ follows from

$$\mathrm{E}\left[\sum_{i=1}^n \left(X_i - \overline{X}\right)^2\epsilon_i^2 \mid X_1^n\right] = \sum_{i=1}^n \left(X_i - \overline{X}\right)^2 \mathrm{E}\left[\epsilon_i^2 \mid X_1^n\right] = \sigma^2\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$$

and

$$\mathrm{E}\left[\sum_{i=1}^n\sum_{j\neq i} \left(X_i - \overline{X}\right)\left(X_j - \overline{X}\right)\epsilon_i\epsilon_j \mid X_1^n\right] = \sum_{i=1}^n\sum_{j\neq i} \left(X_i - \overline{X}\right)\left(X_j - \overline{X}\right)\mathrm{E}\left[\epsilon_i\epsilon_j \mid X_1^n\right] = 0.$$

3. $Y_i \mid X_1^n \sim \mathrm{N}\left(\beta_0 + \beta_1 X_i, \sigma^2\right)$. 4. $\hat{\beta}_1$ given X_1^n is normal, since conditional on X_1^n, $\hat{\beta}_1$ is a linear function of $Y_1, ..., Y_n$, which are jointly normal. And,

$$\hat{\beta}_1 \mid X_1^n \sim \mathrm{N}\left(\mathrm{E}\left[\hat{\beta}_1 \mid X_1^n\right], \mathrm{Var}\left[\hat{\beta}_1 \mid X_1^n\right]\right) \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}\right).$$

5. By 4,

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/\sum_{i=1}^n \left(X_i - \overline{X}\right)^2}} \mid X_1^n \sim \mathrm{N}\left(0, 1\right).$$

The conditional distribution given X_1^n of the z-statistic is independent from X_1^n (standard normal), therefore, the unconditional distribution of it is also standard normal. (Why?)

Problem 7. ISL (2nd edition) Question 7 on Page 54.

Solution. (a) The distances are 3, 2, 3.16, 2.23, 1.41 and 1.73 (obs 1 to 6, respectively). (b) The fifth observation is in the nearest neighbor. The prediction is Green, since the KNN estimate of the conditional probability of Red is 0 and the estimated probability of Green is 1. (c) The second, fifth and sixth are in the 3-nearest neighbor. KNN estimate of the conditional probability of Red is 2/3 and the estimated probability of Green is 1/3. The prediction is Red. (d) As K becomes larger, the KNN boundary becomes inflexible (linear). So in this case we would expect that the optimal K should be small so that the KNN boundary is flexible enough to approximate the Bayes decision boundary.

Part 2: Applied Questions

Write your answer in an RMarkdown report, print your report and hand in.

Problem 8. ISL (2nd edition) Question 8. Give answers to Parts a, b and c(i-iv).

Problem 9. ISL (2nd edition) Question 9. Give answers to Parts a, b, c and d.