

STA 35C Statistical Data Science III

Practice midterm exam 1 solutions

Instructor: Dogyoon Song

Problem 1

(a) **Classification**, because 1, 2, 3 are labels for product categories rather than quantitative responses.

(b) By the law of total probability,

$$P(B) = P(B | A)P(A) + P(B | A^c)P(A^c) = \frac{1}{2} \frac{1}{5} + \frac{1}{8} \frac{4}{5} = \boxed{\frac{1}{5}}.$$

(c) Since $X \in \{0, 1\}$, we have $X^2 = X$, so $\boxed{\mathbb{E}[X^2] = \mathbb{E}[X] = 0.3}$.

(d) Independence gives $\text{Cov}(X, Y) = 0$, hence

$$\text{Var}(2X - Y) = 2^2 \text{Var}(X) + (-1)^2 \text{Var}(Y) = 4(4) + 9 = \boxed{25}.$$

(e) The usage difference is $\Delta x = 200/100 = 2$, so the predicted monthly bill is $\boxed{\$40}$ higher: $\Delta \hat{y} = 20(2) = 40$ dollars.

(f) **No**; the confidence interval concerns the population mean monthly bill at 200 kWh, while individual bills have additional variability around that mean.

(g) **B**. The fit accounts for 64% of the training-response variation about the sample mean.

(h) **No**; $p(x+1) - p(x)$ is positive for every finite x but tends to zero as $x \rightarrow \pm\infty$, so this increase cannot be constant.

(i) The residual standard error is measured in **centimeters**, and 5 estimates the standard deviation of the regression errors, describing typical vertical scatter around the fitted relationship.

(j) Here $FN = 10$, $FP = 0$, and $TP = 0$, so

$$\begin{aligned} \text{Overall error rate} &= \frac{FP + FN}{100} = \frac{10}{100} = \boxed{10\%}, \\ \text{Fraction of actual positives missed} &= \frac{FN}{TP + FN} = \frac{10}{10} = \boxed{100\%}. \end{aligned}$$

Problem 2

(a) Using the positive-probability cells of the joint distribution,

$$\begin{aligned}\mathbb{E}[X] &= 0\left(\frac{1}{4}\right) + 1\left(\frac{1}{2}\right) + 2\left(\frac{1}{4}\right) = \boxed{1}, \\ \mathbb{E}[Y] &= \frac{1}{2}, \quad \mathbb{E}[XY] = 1 \cdot 1 \cdot \frac{1}{2} = \frac{1}{2}, \\ \text{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \frac{1}{2} - 1\left(\frac{1}{2}\right) = \boxed{0}.\end{aligned}$$

(b) **Yes.** For the possible values of X , use

$$\hat{Y}(x) = \begin{cases} 1, & x = 1, \\ 0, & x = 0 \text{ or } 2. \end{cases} \quad P\{\hat{Y}(X) \neq Y\} = 0.$$

Zero covariance coexists here with a nonlinear deterministic relationship, so it does not imply independence or absence of predictive information.

Problem 3

(a) The test statistic is

$$t = \frac{24 - 30}{4} = \boxed{-1.5}, \quad |t| = 1.5 < 2.10.$$

Thus, using the supplied critical value, we **do not reject** H_0 at the 5% level. The data do not provide sufficient evidence that the population slope relating monthly bill to usage differs from \$30 per additional 100 kWh.

(b) Favor **Model B** based on its smallest observed test MSE, 31, compared with 33 for C and 36 for A. Model C's higher training R^2 describes in-sample fit and does not ensure better prediction for new households.

(c) The two fitted bills and their difference are

$$\begin{aligned}\hat{y}_0(x) &= 50 + 30x, & \hat{y}_1(x) &= 110 + 15x, \\ \hat{y}_1(x) - \hat{y}_0(x) &= 60 - 15x = 0 & \iff & x = 4 \quad (400 \text{ kWh}).\end{aligned}$$

No; the coefficient 60 gives the fitted plan difference only at $x = 0$, and the interaction makes that difference depend on usage. The fitted bills are equal at **400 kWh**, with **Plan 0 lower below 400 kWh** and **Plan 1 lower above 400 kWh**.

(d*) Write $f_2(x) = x + c x(x-1)$, which preserves the fitted values at $x = 0, 1$. The condition $f_2(2) = 2 + 2c = 10$ gives $c = 4$, hence

$$\boxed{f_2(x) = 4x^2 - 3x}.$$

Indeed, $f_2(0) = 0$, $f_2(1) = 1$, and $f_2(2) = 10$; therefore its training RSS is zero.

Problem 4

(a) **Linear discriminant analysis (LDA)**, because it combines Gaussian class-conditional models with a common variance and class priors through Bayes' rule.

(b) Under the fitted model, Bayes' rule gives the posterior odds

$$\begin{aligned}\frac{P(Y = 1 \mid Z = z)}{P(Y = 0 \mid Z = z)} &= \frac{\hat{\pi}_1 f_1(z)}{\hat{\pi}_0 f_0(z)} \\ &= \frac{1}{3} \exp\left(\frac{z^2 - (z - 2)^2}{4}\right) = \frac{1}{3} e^{z-1}.\end{aligned}$$

Class 1 is chosen when these odds are at least 1, equivalently

$$z - 1 - \log 3 \geq 0 \quad \Longleftrightarrow \quad \boxed{z \geq 1 + \log 3}.$$

The cutoff is $\boxed{1 + \log 3 \approx 2.099}$; scores at or above the exact cutoff are classified as defective, including the tie.

(c) The total costs on the stated test set are

$$C_{0.5} = 30r + 10, \quad C_{0.8} = 10r + 35, \quad C_{0.8} - C_{0.5} = 25 - 20r.$$

Therefore,

$$r > 5/4 : \quad \text{favor threshold 0.8,}$$

$$r = 5/4 : \quad \text{the thresholds have equal total cost,}$$

$$0 < r < 5/4 : \quad \text{favor threshold 0.5.}$$

The higher threshold avoids 20 false positives at the expense of 25 additional false negatives, so it has lower cost precisely when $20r > 25$.

(d*) **Yes:** choose $(\mu_0, \mu_1) = (-1, 1)$ for Model I and $(-2, 2)$ for Model II, each with $\pi_0 = \pi_1 = 1/2$ and common variance 1, and write $p_m(x) = P_m(Y = 1 \mid X = x)$ for $m \in \{\text{I, II}\}$.

$$\begin{aligned}\log \frac{p_{\text{I}}(x)}{1 - p_{\text{I}}(x)} &= \frac{(x+1)^2 - (x-1)^2}{2} = 2x, \\ \log \frac{p_{\text{II}}(x)}{1 - p_{\text{II}}(x)} &= \frac{(x+2)^2 - (x-2)^2}{2} = 4x.\end{aligned}$$

Both models choose class 1 exactly when $x \geq 0$, including the tie at zero, but their posterior probabilities conditional on $X = 1$ differ:

$$P_{\text{I}}(Y = 1 \mid X = 1) = \frac{e^2}{1 + e^2} < \frac{e^4}{1 + e^4} = P_{\text{II}}(Y = 1 \mid X = 1).$$