 December 20, 2022

# Exam 3

| Last Name | First Name | Student ID # |
|---|---|---|
| | | |

Honor Code: This exam represents only my own work. I did not give or receive help.

Signature: ______________________________

**Partial Credit:** The most important issue is knowing how to approach a particular problem. Therefore, there will be partial credit for good solution outlines even if not all the mathematical manipulations are completed correctly. Be sure to attempt every problem!

- You have exactly **2 hours** to complete this exam.
- **No devices are allowed** - including no phones and no calculators.
- Unless indicated otherwise, you only need to setup up integrals correctly for full credit, **which includes the correct limits and case-by-case conditions.**
- You can use the provided formula sheet handouts - no extra materials are allowed.
- No form of collaboration is allowed.
- There are 5 problems in total, each worth 20 points.

*** Good Luck! ***

| Problem | Points earned | out of | Problem | Points earned | out of |
|---|---|---|---|---|---|
| Problem 1 | | 20 | Problem 4 | | 20 |
| Problem 2 | | 20 | Problem 5 | | 20 |
| Problem 3 | | 20 | | | |
| | | | | | |
| | | | Total | | 100 |

**Problem 1** (Detection) *20 points*

(*This is a two-page problem with a single scenario.*)
Consider the following detection problem. $\mathbb{P}[H_0] = 4/5$ and $\mathbb{P}[H_1] = 1/5$.
Under $H_0$, $Y$ is Exponential(1). Under $H_1$, $Y$ is Uniform(0, 4).

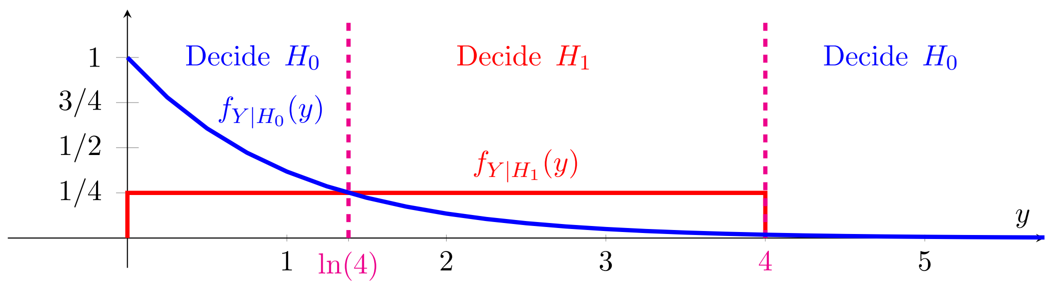
(a) Determine the ML rule. Simplify your expression as much as you can.

**Solution:**

We know that $f_{Y|H_0}(y) = \begin{cases} e^{-y} & y \geq 0 \\ 0 & y < 0 \end{cases}$ and $f_{Y|H_1}(y) = \begin{cases} \frac{1}{4} & 0 \leq y \leq 4 \\ 0 & \text{otherwise.} \end{cases}$ The likelihood ratio is $\mathcal{L}(y) = \frac{1}{4}e^y$. We see that $\mathcal{L}(y) \geq 1$ if $e^y \geq 4 \implies y \geq \ln(4)$. Therefore, the ML rule is

$$D^{\mathrm{ML}}(y) = \begin{cases} 1 & \ln(4) \leq y \leq 4 \\ 0 & 0 \leq y < \ln(4) \text{ or } y > 4 \end{cases}$$

(b) Sketch the conditional PDFs $f_{Y|H_0}(y)$ and $f_{Y|H_1}(y)$ below. Clearly indicate the regions where the ML rule will decide 0 and where it will decide 1.




(c) Determine the probability of error for the ML rule.

**Solution:**

$$P_{\mathrm{FA}} = \mathbb{P}[D^{\mathrm{ML}}(Y) = 1 | H_0] = \int_{\ln(4)}^{4} e^{-y}\,dy = \frac{1}{4} - e^{-4}$$

$$P_{\mathrm{MD}} = \mathbb{P}[D^{\mathrm{ML}}(Y) = 0 | H_1] = \int_{0}^{\ln(4)} \frac{1}{4}\,dy = \frac{\ln(4)}{4}$$

$$\mathbb{P}[\mathrm{error}_{\mathrm{ML}}] = P_{\mathrm{FA}}\,\mathbb{P}[H_0] + P_{\mathrm{MD}}\,\mathbb{P}[H_1] = \frac{4}{5} \cdot \left(\frac{1}{4} - e^{-4}\right) + \frac{1}{5} \cdot \frac{\ln(4)}{4} = \frac{4 - 16e^{-4} + \ln(4)}{20}.$$

**(CONTINUED ON NEXT PAGE)**

(d) Determine the MAP rule. Simplify your expression as much as you can.

**Solution:**

We see that $\mathcal{L}(y) \geq \frac{\mathbb{P}[H_0]}{\mathbb{P}[H_1]} = 4$ if $e^y \geq 16 \implies y \geq \ln(16)$. Therefore, the MAP rule is

$$D^{\mathrm{MAP}}(y) = \begin{cases} 1 & \ln(16) \leq y \leq 4 \\ 0 & 0 \leq y < \ln(16) \text{ or } y > 4 \end{cases}$$

(e) Determine the probability of error for the MAP rule.

**Solution:**

$$P_{\mathrm{FA}} = \mathbb{P}[D^{\mathrm{MAP}}(Y) = 1 | H_0] = \int_{\ln(16)}^{4} e^{-y}\, dy = \frac{1}{16} - e^{-4}$$

$$P_{\mathrm{MD}} = \mathbb{P}[D^{\mathrm{MAP}}(Y) = 0 | H_1] = \int_{0}^{\ln(16)} \frac{1}{4}\, dy = \frac{\ln(16)}{4}$$

$$\mathbb{P}[\mathrm{error}_{\mathrm{MAP}}] = P_{\mathrm{FA}}\, \mathbb{P}[H_0] + P_{\mathrm{MD}}\, \mathbb{P}[H_1] = \frac{4}{5} \cdot \left(\frac{1}{16} - e^{-4}\right) + \frac{1}{5} \cdot \frac{\ln(16)}{4} = \frac{1 - 16e^{-4} + \ln(16)}{20}.$$

**Problem 2** (Estimation) *20 points*

(*This is a two-page problem with a two different scenarios.*)

**Scenario 1:** $Y = X + Z$ where $X$ and $Z$ are independent random variables with

$$\mathbb{E}[X] = 3 \qquad \mathbb{E}[Z] = 0 \qquad \mathsf{Var}[X] = 4 \qquad \mathsf{Var}[Z] = 2$$

(a) Determine $\mathbb{E}[Y]$, $\mathsf{Var}[Y]$, and $\mathsf{Cov}[X, Y]$.

**Solution:**

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}[X] + \mathbb{E}[Z] = 3 + 0 = 3 \\ \mathsf{Var}[Y] &= \mathsf{Var}[X] + \mathsf{Var}[Z] + 2\mathsf{Cov}[X, Z] = 4 + 2 + 0 = 6 \\ \mathsf{Cov}[X, Y] &= \mathsf{Cov}[X, X + Z] = \mathsf{Var}[X] + \mathsf{Cov}[X, Z] = 4 + 0 = 4\end{aligned}$$

(b) We would like to find a linear estimator of $X$ of the form $aY + b$ that minimizes the mean-squared error. Determine the optimal values of $a$ and $b$.

**Solution:**

The optimal estimator is the LLSE estimator

$$\begin{aligned}\hat{x}_{\text{LLSE}}(Y) &= \mathbb{E}[X] + \frac{\mathsf{Cov}[X, Y]}{\mathsf{Var}[Y}(y - \mathbb{E}[Y]) \\ &= 3 + \frac{4}{6}(y - 3) = \frac{2}{3}y + 1 \implies a = \frac{2}{3},\ b = 1\end{aligned}$$

(c) Determine the mean-squared error of your estimator from part (b).

**Solution:**

$$\mathsf{MSE}_{\text{LLSE}} = \mathsf{Var}[X] - \frac{\big(\mathsf{Cov}[X, Y]\big)^2}{\mathsf{Var}[Y]} = 4 - \frac{4^2}{6} = \frac{24 - 16}{6} = \frac{8}{6} = \frac{4}{3}$$

**(CONTINUED ON NEXT PAGE)**

**Scenario 2:** $\begin{array}{l} Y_1 = X_1 + Z_1 \\ Y_2 = 2X_2 + Z_2 \end{array}$ where $X_1, X_2, Z_1, Z_2$ are independent random variables with

$$\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0 \qquad \mathbb{E}[Z_1] = \mathbb{E}[Z_2] = 0 \qquad \mathsf{Var}[X_1] = \mathsf{Var}[X_2] = 4 \qquad \mathsf{Var}[Z_1] = \mathsf{Var}[Z_2] = 2$$

(d) Determine $\mathbf{\Sigma}_{\underline{Y}} = \begin{bmatrix} \mathsf{Var}[Y_1] & \mathsf{Cov}[Y_1, Y_2] \\ \mathsf{Cov}[Y_2, Y_1] & \mathsf{Var}[Y_2] \end{bmatrix}$ and $\mathbf{\Sigma}_{\underline{X},\underline{Y}} = \begin{bmatrix} \mathsf{Cov}[X_1, Y_1] & \mathsf{Cov}[X_1, Y_2] \\ \mathsf{Cov}[X_2, Y_1] & \mathsf{Cov}[X_2, Y_2] \end{bmatrix}$.

(*Hint: Note that* $g(X_1, Z_1)$ *and* $h(X_2, Z_2)$ *are independent for any functions* $g$ *and* $h$*.*)

**Solution:**

Using independence, we have that $\mathsf{Cov}[Y_1, Y_2] = 0$, $\mathsf{Cov}[X_1, Y_2] = 0$, and $\mathsf{Cov}[X_2, Y_1] = 0$. From part (a), we know that $\mathsf{Var}[Y_1] = 6$ and $\mathsf{Cov}[X_1, Y_1] = 4$. Similarly,

$$\begin{aligned} \mathsf{Var}[Y_2] &= 4\mathsf{Var}[X_2] + \mathsf{Var}[Z_2] + 2 \cdot 4\mathsf{Cov}[X_2, Z_2] = 16 + 2 + 0 = 18 \\ \mathsf{Cov}[X_2, Y_2] &= \mathsf{Cov}[X_2, 2X_2 + Z_2] = 2\mathsf{Var}[X_2] + \mathsf{Cov}[X_2, Z_2] = 8 + 0 = 8 \ . \end{aligned}$$

Therefore, $\mathbf{\Sigma}_{\underline{Y}} = \begin{bmatrix} \mathsf{Var}[Y_1] & \mathsf{Cov}[Y_1, Y_2] \\ \mathsf{Cov}[Y_2, Y_1] & \mathsf{Var}[Y_2] \end{bmatrix} = \begin{bmatrix} 6 & 0 \\ 0 & 18 \end{bmatrix}$ and
$\mathbf{\Sigma}_{\underline{X},\underline{Y}} = \begin{bmatrix} \mathsf{Cov}[X_1, Y_1] & \mathsf{Cov}[X_1, Y_2] \\ \mathsf{Cov}[X_2, Y_1] & \mathsf{Cov}[X_2, Y_2] \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 8 \end{bmatrix}$.

(e) We would like to find a linear estimator of $\begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ of the form $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$ that minimizes the mean-squared error. Determine the optimal values of $a_{11}, a_{12}, a_{21}, a_{22}, b_1, b_2$.

**Solution:**

The optimal estimator of this form is the vector LLSE estimator

$$\underline{\hat{x}}_{\text{LLSE}}(\underline{Y}) = \mathbb{E}[\underline{X}] + \mathbf{\Sigma}_{\underline{X},\underline{Y}} \mathbf{\Sigma}_{\underline{Y}}^{-1} (\underline{Y} - \mathbb{E}[\underline{Y}]) \ .$$

Note that both $\underline{X}$ and $\underline{Y}$ are both mean zero. Thus, $b_1 = 0$, $b_2 = 0$. We also have

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} 4 & 0 \\ 0 & 8 \end{bmatrix} \begin{bmatrix} 6 & 0 \\ 0 & 18 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{2}{3} & 0 \\ 0 & \frac{4}{9} \end{bmatrix}$$

**Problem 3** (Statistics) *20 points*

You have been asked to evaluate the performance of two new stores. The table below summarizes how many online reviews each store received for a given star count (from 1 to 5 stars).

| | 1 Star | 2 Stars | 3 Stars | 4 Stars | 5 Stars |
|---|---|---|---|---|---|
| Store A Review Count | 0 | 1 | 0 | 1 | 0 |
| Store B Review Count | 0 | 0 | 0 | 2 | 2 |

You may find the following table useful. Recall that $F_{T_m}(t)$ is the CDF for a t-distribution with $m$ degrees-of-freedom and $F^{-1}_{T_m}(\beta)$ is its inverse.

| $m$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $F^{-1}_{T_m}(0.025)$ | -12.71 | -4.30 | -3.18 | -2.78 | -2.57 | -2.45 | -2.36 | -2.31 | -2.62 | -2.23 |
| $F^{-1}_{T_m}(0.05)$ | -6.31 | -2.92 | -2.35 | -2.13 | -2.02 | -1.94 | -1.89 | -1.86 | -1.83 | -1.81 |
| $F^{-1}_{T_m}(0.1)$ | -3.08 | -1.89 | -1.64 | -1.53 | -1.48 | -1.44 | -1.41 | -1.40 | -1.38 | -1.37 |

(a) Determine the sample mean and sample variance for Store A as well as for Store B.

> **Solution:**
> $$M^{(A)}_{n_1} = \frac{1}{2}(2+4) = 3 \qquad V^{(A)}_{n_1} = \frac{1}{2-1}\big((4-3)^2 + (2-3)^2\big) = 2$$
> $$M^{(B)}_{n_2} = \frac{1}{4}(4+4+5+5) = 4.5 \qquad V^{(B)}_{n_2} = \frac{1}{4-1}\big(2\cdot(5-4.5)^2 + 2\cdot(4-4.5)^2\big) = \frac{1}{3}$$

(b) Construct a confidence interval for the Store A average review with confidence level $0.9$.

> **Solution:**
> The variance is unknown and we have $n_1 = 2 < 30$ samples so we use the T-distribution. Here, $1-\alpha = 0.9$ so $\alpha/2 = 0.05$. Our confidence interval is $[M^{(A)}_{n_1} \pm \epsilon]$ with $\epsilon = -\frac{\sqrt{V_n}}{\sqrt{n_1}}F^{-1}_{T_{n_1-1}}(\alpha/2) = -\frac{\sqrt{2}}{\sqrt{2}}F^{-1}_{T_1}(0.05) = 6.31$. Equivalently, $[3 \pm 6.31]$ or $[-3.31, 9.31]$.

(c) You have good reason to believe that the review variance is **equal** across stores. Use this new information to calculate the pooled sample variance.

> **Solution:** $$\hat{\sigma}^2 = \frac{(n_1-1)V^{(A)}_{n_1} + (n_2-1)V^{(B)}_{n_2}}{n_1+n_2-2} = \frac{(2-1)\cdot 2 + (4-1)\cdot\frac{1}{3}}{2+4-2} = \frac{3}{4}$$

(d) You would like to evaluate whether gap between the average review for Store A and Store B is statistically significant. Assuming the review variance is **equal** across stores, what kind of significance test should you use?

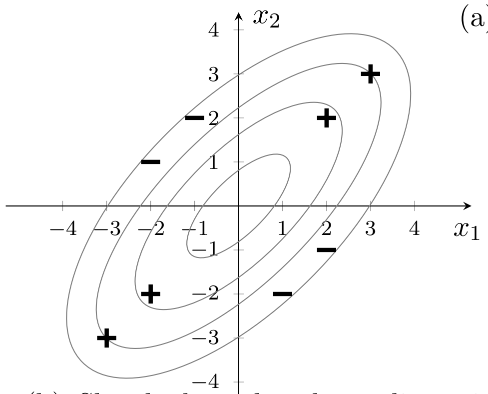
> **Solution:**
> Since we are comparing the means of two datasets of equal variance (with less than 30 samples), a two-sample T-test is appropriate.

(e) Should we reject the null hypothesis at a significance level of $0.1$? Justify your answer.

> **Solution:**
> $$T = \frac{(M^{(A)}_{n_1} - M^{(B)}_{n_2})}{\sqrt{\hat{\sigma}^2\big(\frac{1}{n_1}+\frac{1}{n_2}\big)}} = \frac{3-4.5}{\sqrt{\frac{3}{4}\big(\frac{1}{2}+\frac{1}{4}\big)}} = \frac{-3/2}{3/4} = -2$$
> We also have p-value $= 2F_{n_1+n_2-2}(-|T|) = 2F_4(-2) > 2F_4(-2.13) = 2\cdot 0.05 = 0.1$ so we fail to reject the null hypothesis.

**Problem 4** (Machine Learning) *20 points*

You are given the 8 training data points on the figure, denoted by $+$ and $-$ symbols. The ellipses represent a contour plot for a vector Gaussian distribution fit to the entire training dataset. You will use PCA dimensionality reduction to create a one-dimensional version of this training dataset. **Each part can be solved mainly with plots and illustrations.**




(a) The PCA transform is of the form: $z = a_1x_1 + a_2x_2 + b$ Determine the values of $a_1, a_1, b$.

**Solution:**

The mean vector is clearly at the origin, so no recentering is necessary. Thus, $b = 0$. From the contour plot, we see that the largest eigenvector is in the direction $\begin{bmatrix}1\\1\end{bmatrix}$. Normalizing this to a unit vector, we get $a_1 = a_2 = \frac{1}{\sqrt{2}}$.

(b) Sketch the reduced one-dimensional dataset on the plot at the bottom of the page. (You do not need to exactly evaluate the one-dimensional coordinates or label your axes, but the relative spacing of the 8 points should be correct.)

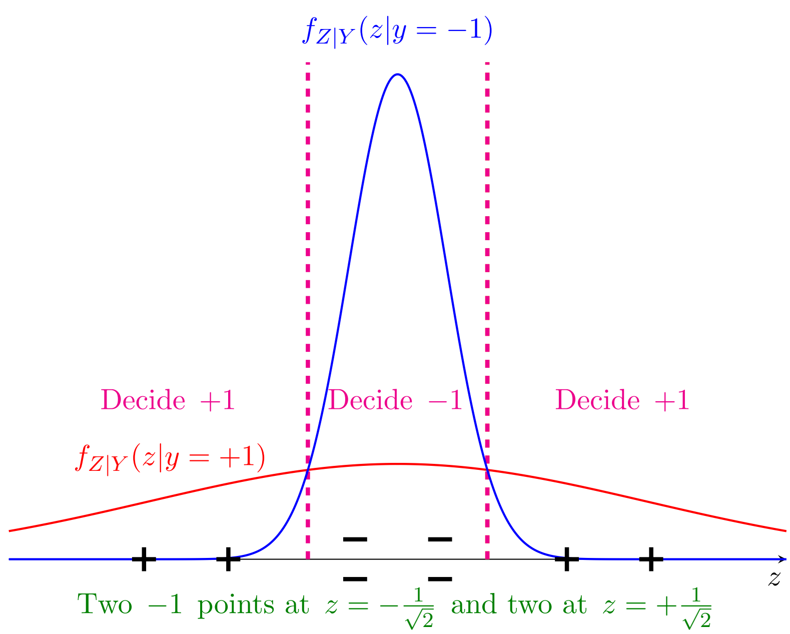
(c) For the reduced one-dimensional dataset, determine the training error rate for the closest average classifier. Justify your answer.

**Solution:**

Both labels have mean 0 in the reduced space. Thus, the closest average classifier will assign every training point to $+1$ to break the tie. The training error rate is $0.5 = 50\%$.

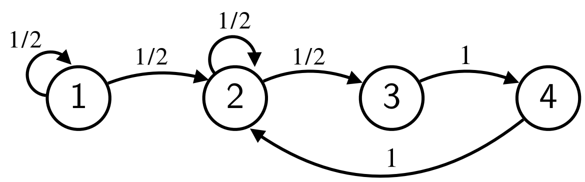
(d) Using **dashed lines**, sketch decision boundaries that will result in 0 training errors.

(e) For this reduced dataset, it turns out the QDA classifier has 0 training errors. Below, sketch the likelihoods of the two Gaussian distributions used to determine these decision boundaries. No calculations are necessary, just an approximate sketch.



**Problem 5** (Markov Chains) *20 points*

Consider the following discrete-time Markov chain. $X_0$ is equally likely to be $1$, $2$, $3$, or $4$.




(a) List the communicating classes. For each communicating class, determine the period and whether it is transient or recurrent.

> **Solution:**
>
> $C_1 = \{1\}$ which has period $1$ and is transient. $C_2 = \{2,3,4\}$ which has period $1$ and is recurrent.

(b) Determine $\mathbb{P}[X_2 = 1 | X_0 = 4]$.

> **Solution:**
>
> Since there is no path from state $4$ to state $1$, this probability is $0$.

(c) Determine $\mathbb{P}[X_2 = 1]$.

> **Solution:**
>
> Once we leave state $1$, it is impossible to return. Thus, the only valid path is $X_0 = 1, X_1 = 1, X_2 = 1$. The probability of $X_0 = 1$ is $1/4$ and the remaining transitions occur with probability $1/2$ each. Thus, $\mathbb{P}[X_4 = 1] = \frac{1}{4} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{16}$.

(d) Does a unique limiting state probability vector $\underline{\pi}$ exist? If so, argue why and solve for it. If not, argue why.

> **Solution:**
>
> Since there is a single recurrent communicating class with period $1$, there is a unique limiting state probability vector $\underline{\pi}$. Since state $1$ is transient, we first set $\pi_1 = 0$. From the steady-state equation $\mathbf{P}^T \underline{\pi} = \underline{\pi}$,
>
> $$\pi_4 = \pi_3; \qquad \pi_3 = \frac{1}{2}\pi_2 \implies \pi_4 = \frac{1}{2}\pi_2$$
>
> From normalization, $\pi_1 + \pi_2 + \pi_3 + \pi_4 = 0 + \pi_2 + \dfrac{\pi_2}{2} + \dfrac{\pi_2}{2} = 2\pi_2 = 1$
>
> $$\implies \pi_1 = 0,\ \pi_2 = \frac{1}{2},\ \pi_3 = \frac{1}{4},\ \pi_4 = \frac{1}{4}$$

(e) Given that the Markov chain starts in state $3$, find the expected number of steps until it returns to state $3$.

> **Solution:**
>
> We can use what we know from Geometric random variables to solve this problem. First, note that after state $3$, the chain always jumps to $4$ and then $2$, for a total of $2$ steps. Now, let $Y$ be a Geometric$(1/2)$ random variable that counts the number of steps until the Markov chain jumps from state $2$ to state $3$. We know that $\mathbb{E}[Y] = 2$. Thus, the expected number of steps is $4$.