ENG EK381 Probability, Statistics and Data Science for Engineers — Fall 2022

 — December 20, 2022

Exam 3

Last Name	First Name	Student ID #
Lee	Jake	U62785048

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$(0,4)$.

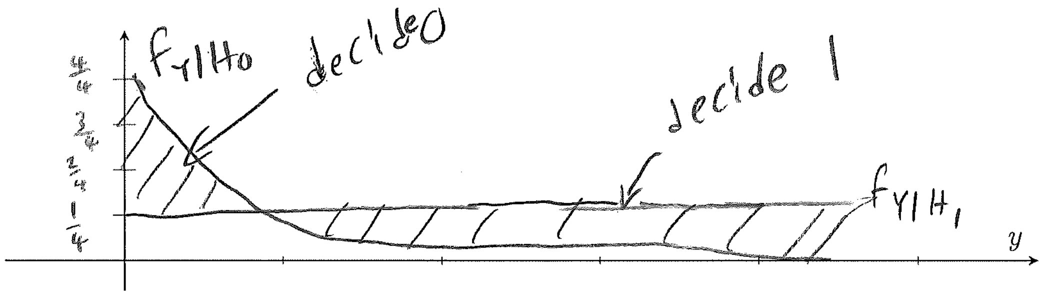
(a) Determine the ML rule. Simplify your expression as much as you can.

$P_{Y|H_0} = e^{-y} \quad P_{Y|H_1} = \frac{1}{4}$

$$D^{ML} = \begin{cases} 1, & \frac{1}{4} \geq e^{-y} \\ 0, & \frac{1}{4} < e^{-y} \end{cases} = \begin{cases} 1, & y = 2,3,4 \\ 0, & y = 0,1 \end{cases} \qquad D^{ML} = \begin{cases} 1, & \ln(\frac{1}{4}) + y \geq 0 \\ 0, & \ln(\frac{1}{4}) + y < 0 \end{cases}$$

$$\ln(L(y)) = \ln\left(\frac{1/4}{e^{-y}}\right) = \ln\left(\frac{1}{4} e^{y}\right) = \ln\left(\frac{1}{4}\right) + y \qquad D^{ML} = \begin{cases} 1, & y \geq -\ln(\frac{1}{4}) \\ 0, & y < -\ln(\frac{1}{4}) \end{cases}$$

(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.

$$P_e = P_{FA}\, P[H_0] + P_{MD}\, P[H_1]$$

$$P_{FA} = P_{Y|H_0}(0) + P_{Y|H_0}(1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{16}$$

$$P_{MD} = P_{Y|H_1}(2) + P_{Y|H_1}(3) + P_{Y|H_1}(4) = e^{-2} + e^{-3} + e^{-4} = e^{-9}$$

$$P_e = \frac{1}{\cancel{16}\ 4}\left(\frac{4}{5}\right) + e^{-9}\left(\frac{1}{5}\right) = \boxed{\frac{1}{20} + \frac{1}{5e^{9}}}$$

(CONTINUED ON NEXT PAGE)

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

$$D^{MAP}(y) = \begin{cases} 1, & \ln(\frac{1}{4}) + y \geq \ln(4) \\ 0, & \ln(\frac{1}{4}) + y < \ln(4) \end{cases}$$

$$\begin{cases} 1, & y \geq \ln(4) - \ln(\frac{1}{4}) \\ 0, & y < \ln(4) - \ln(\frac{1}{4}) \end{cases} = \begin{cases} 1, & y \geq \ln(8) \\ 0, & y \leq \ln(8) \end{cases}$$

$$P_{Y|H_0} P[H_0] = \frac{4}{5e^{y}} \qquad P_{Y|H_1} P[H_1] = \frac{1}{20}$$

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

$$P_e = P_{FA}\, P[H_0] + P_{MD}\, P[H_1]$$

$$P_{FA} = P_{Y|H_0}(0) + P_{Y|H_0}(1) = \frac{1}{4} + \frac{1}{4} = \frac{1}{16}$$

$$P_{MD} = P_{Y|H_1}(2) + P_{Y|H_1}(3) + P_{Y|H_1}(4) = e^{-9}$$

$$P_e = \frac{1}{16}\left(\frac{4}{5}\right) + e^{-9}\left(\frac{1}{5}\right) = \boxed{\frac{1}{20} + \frac{1}{5e^{-9}}}$$

Problem 2 (Estimation) *20 points*

(*This is a two-page problem with 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 $\mathsf{Cov}[X, Y]$.

$E[Y] = E[X+Z] = E[X] + E[Z] = 3 + 0 = \boxed{3}$

$Var[Y] = Var[X+Z] = 4 + 2$

$E[XY] = E[X^2 + XZ] = 9 + 0 = 9$

$Cov[X, Y] = Cov[X, X+Z] = Var[X] + Var[Y]$

$Cov[X, Y] = 9 - (3)(3) = 0$

(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.

$\hat{X}_{MMSE}(Y) = E[(X - \hat{X}(Y))^2]$

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

(CONTINUED ON NEXT PAGE)

Scenario 2: $Y_1 = X_1 + Z_1$ where X_1, X_2, Z_1, Z_2 are independent random variables with
$Y_2 = 2X_2 + Z_2$

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

(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.)

$\mathrm{Var}[Y_1] = \mathrm{Var}[X_1 + Z_1] = 4 + 2 = 6$

$\mathrm{Var}[Y_2] = \mathrm{Var}[2X_2 + Z_2] = 2^2(4) + 2 = 18$

$\Sigma_Y = \begin{bmatrix} 6 & \\ & 18 \end{bmatrix}$

(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 $a_{11}, a_{12}, a_{21}, a_{22}, b_1, b_2$.

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 $F_{T_m}^{-1}(\beta)$ is its inverse.

m	1	2	3	4	5	6	7	8	9	10
$F_{T_m}^{-1}(0.025)$	-12.71	-4.30	-3.18	-2.78	-2.57	-2.45	-2.36	-2.31	-2.62	-2.23
$F_{T_m}^{-1}(0.05)$	-6.31	-2.92	-2.35	-2.13	~~-2.02~~	-1.94	-1.89	-1.86	-1.83	-1.81
$F_{T_m}^{-1}(0.1)$	-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.

$\hat{\mu}_A = \frac{1}{2}(2+4) = \boxed{3}$ $\quad V_{nA} = \frac{1}{2-1}(2-3)^2+(4-3)^2 = 1+1 = \boxed{2}$

$\hat{\mu}_B = \frac{1}{4}(4+4+5+5) = \frac{18}{4} = \boxed{\frac{9}{2}}$ $\quad V_{nB} = \frac{1}{4-1}(4-\frac{9}{2})^2+(4-\frac{9}{2})^2+(5-\frac{9}{2})^2+(5-\frac{9}{2})^2$

$\frac{1}{3}\left(\frac{1}{4}+\frac{1}{4}+\frac{1}{4}+\frac{1}{4}\right) = \boxed{\frac{1}{3}}$

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

$T = \frac{\sqrt{2}}{\sqrt{2}}(3-$

$6.31 \times 2 = 12.62$

$P_{val} = 2F_{T_1}(0.05) = 2(-6.31) = -12.62$

$[3 \pm 12.62]$

$\mu_n = \frac{24}{6} = 4$

(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. 2 sample T test

$\sigma^2 = ((2-1)(2)+(4-1)(\frac{1}{3}))/(2+4-2) = \frac{2+1}{4} = \boxed{\frac{3}{4}}$

(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?

2 sample T test

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

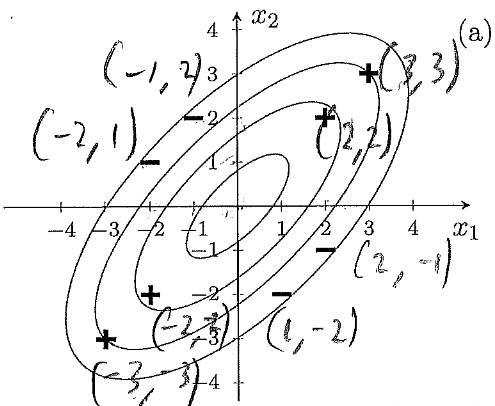
$$T = \frac{(\mu_{n_1}-\mu_{n_2})}{\sqrt{\sigma^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{\left(\frac{6}{2}-\frac{9}{2}\right)}{\sqrt{\frac{3}{4}\left(\frac{1}{2}+\frac{1}{4}\right)}} = \frac{-\frac{3}{2}}{\sqrt{\frac{6}{4}}} = \frac{-3(\cancel{2})}{\sqrt{2}(\cancel{2})} = \frac{-3}{\sqrt{2}}$$

$-\frac{3\sqrt{2}}{2} \approx -\frac{3}{2} = -1.5 < F_{T_5}^{-1}(0.1) = -1.48$

reject because the p val will be < 0.1

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.

$a_1 = 1$

$a_2 = \frac{1}{2}$

$b = 3$

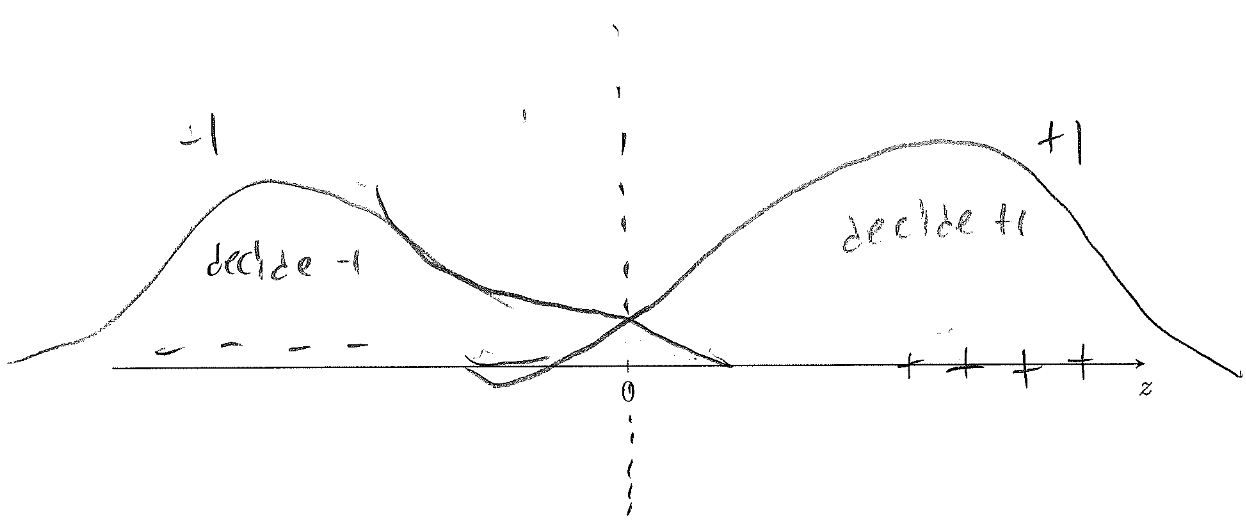
(b) Sketch the reduced one-dimensional dataset on the plot at the bottom of the page. (You do not need to exactly evaluate the one-dimensional coordinates or label your axes, but the relative spacing of the 8 points should be correct.)

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

$\frac{1}{2}$ error rate any linear line through the origin would give half +1 and half -1 on either side of the line

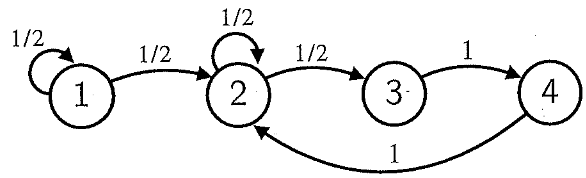
(d) Using **dashed lines**, sketch decision boundaries below that will result in 0 training errors.

(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.

$C_1 = \{1\}$ period $= 1$ transient

$C_2 = \{2,3,4\}$ period $= 1$ recurrent

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

$$P[X_2=1 | X_0 = 4] = \frac{P[X_0=4 | X_2=1]\, P[X_2=1]^2}{P_0[X_0=4]} = 0$$

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

$$P[X_2 = 1] = P_0(1)\, P_{12} = \frac{1}{4}\left(\frac{1}{2}\right) = \frac{1}{8}$$

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

A unique limiting state probability vector π exists

$$\begin{bmatrix} \frac12 & \frac12 & 0 & 0 \\ 0 & \frac12 & \frac12 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \rightarrow P^T \pi = \pi = \begin{bmatrix} \frac12 & 0 & 0 & 0 \\ \frac12 & \frac12 & 0 & 1 \\ 0 & \frac12 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{bmatrix} = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{bmatrix}$$

$$\frac12 \pi_1 = \pi_1$$

$$\frac12 \pi_1 + \frac12 \pi_2 = \pi_2$$

$$\frac12 \pi_2 = \pi_3$$

$$\pi_3 = \pi_4$$

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

3