

1 Honor Code

1. Please list the names and SIDs of all students you have collaborated with below.

2. Declare and sign the following statement (Mac Preview, PDF Expert, and FoxIt PDF Reader, among others, have tools to let you sign a PDF file):

"I certify that all solutions are entirely my own and that I have not looked at anyone else's solution. I have given credit to all external sources I consulted."



Signature: _____

Question 2: Gaussian Classification

$$f_{X|Y=C_i}(x) \sim N(\mu_i, \sigma^2) \quad \Delta = 1$$

$$P(Y = C_1) = P(Y = C_2) = \frac{1}{2} \quad ; \quad N_2 > N_1$$

1) Bayes optimal decision bounds

$$P(C_1|x) = P(C_2|x)$$

$$f(x|C_1) \frac{P(C_1)}{f(x)} = f(x|C_2) \frac{P(C_2)}{f(x)}$$

$$f(x|C_1) = f(x|C_2)$$

$$N(\mu_1, \sigma^2) = N(\mu_2, \sigma^2)$$

$$(x - \mu_1)^2 = (x - \mu_2)^2$$

* Decision Bound is the point equidistant to μ_1 and μ_2

$$b = \frac{\mu_1 + \mu_2}{2}$$

Loss function:

$$\begin{cases} \text{Class 1} ; & x < \frac{\mu_1 + \mu_2}{2} \\ \text{Class 2} ; & \text{otherwise} \end{cases}$$

2) Prove: Bayes Error

$$P_e = P((C_1 \text{ misclassified as } C_2) \cup (C_2 \text{ misclassified as } C_1))$$

$$= P(C_1 \text{ misclassified as } C_2) + P(C_2 \text{ misclassified as } C_1)$$

$$= P(\text{misclassified as } C_2 | C_1) P(C_1) + P(\text{misclassified as } C_1 | C_2) P(C_2)$$

$$\therefore P(C_1) = P(C_2) = \frac{1}{2}$$



PPF:

$$P(\text{misclassified as } C_1 | C_2) = \int_{-\infty}^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_2)^2}{2\sigma^2}} dx$$

$$P(\text{misclassified as } C_2 | C_1) = \int_b^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx$$

Bayes Error:

$$P_e(b) = \frac{1}{2} \left(P(\text{misclassified as } C_2 | C_1) + P(\text{misclassified as } C_1 | C_2) \right)$$

$$P_e(b) = \frac{1}{2} \left(\frac{1}{\sqrt{2\pi}\sigma} \right) \left(\underbrace{\int_{-\infty}^b e^{-\frac{(x-\mu_2)^2}{2\sigma^2}} dx}_{+} + \underbrace{\int_b^{\infty} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx}_{+} \right)$$

3) Optimal Decision Boundary $b^* = \min P_e(b)$

$\Rightarrow b^*$ must be between μ_1 and μ_2

Min $P_e(b)$:

$$\frac{d P_e(b)}{d b} = \frac{1}{2} \left(\frac{1}{\sqrt{2\pi}\sigma} \right) \left(\underbrace{\int_{-\infty}^b e^{-\frac{(x-\mu_2)^2}{2\sigma^2}} dx}_{+} + \underbrace{\int_b^{\infty} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} dx}_{+} \right)$$

Fundamental theorem of calculus:

$$\frac{d}{dx} \int_{-\infty}^x f(x) dx = f(x)$$

$$\Rightarrow \frac{1}{2} \left(\frac{1}{\sqrt{2\pi}\sigma} \right) \left(f_1(b^*) - f_2(b^*) \right)$$

$$\begin{aligned} 0 &= \frac{1}{2} \left(\frac{1}{\sqrt{2\pi}\sigma} \right) \left(\underbrace{f_1(b^*) - f_2(b^*)}_{=} \right) \\ &= 0 \end{aligned}$$



$$f_1(b^*) = f_2(b^*)$$

$$\log(f_1(b^*)) = \log(f_2(b^*))$$

$$\frac{(b^* - \nu_2)^2}{2\sigma^2} = \frac{(b^* - \nu_1)^2}{2\sigma^2}$$

$$(b^* - \nu_2)^2 = (b^* - \nu_1)^2$$

∴ $b^* = \frac{\nu_1 + \nu_2}{2}$

Question 3: Classification and Risks

Classification w/ $r: \mathbb{R}^d \rightarrow \{1, \dots, c+1\}$ decision rule

Loss function:

$$L(r(x)=i, y=s) = \begin{cases} 0 & ; i = s \quad (s \in 1, \dots, c) \\ \lambda_c & ; i \neq s \quad (s \in \{1, \dots, c\}) \\ \lambda_2 & ; i = c+1 \end{cases}$$

$\lambda_c \geq 0$ loss from misclassification

$\lambda_2 \geq 0$ loss from doubt

↳ i) correct = no penalty Risks Classifying over Data points:
 ii) Incorrect = λ_c penalty $R(r(x)_i = i | x) = \sum_{j=1}^c L(r(x)_i = i, y=j) P(Y=j | x)$
 iii) Doubt = λ_2 penalty

1) Simplify Risks function

a) prove $R(r(x)_i = i | x) = \lambda_c (1 - P(Y=i | x))$ if $i \neq c+1$

Risks Classifying over Data points:

$$R(r(x)_i = i | x) = \sum_{j=1}^c L(r(x)_i = i, y=j) P(Y=j | x)$$

In addition to $i \neq c+1$, the loss functions can only be 2 values:

1) correct classification: $L(r(x)_i = i, y=i) = 0$

2) misclassification: $L(r(x)_i = i, y=j) = \lambda_c$

Both cases:

↳ $R(r(x)_i = i | x) = \sum_{\substack{j=1 \\ j \neq i}}^c \lambda_c P(Y=j | x) + 0 \cdot P(Y=i | x)$



$$R(r_{C2j} = c \mid x) = \lambda_c \sum_{\substack{j=1 \\ j \neq i}}^c P(Y=j \mid x)$$

$$\text{All prob of valid pdf} = 1 ; \quad \sum_{j=1}^c P(Y=j \mid x) = 1$$

$$\hookrightarrow \sum_{\substack{j=1 \\ j \neq i}}^c P(Y=j \mid x) = 1 - P(Y=i \mid x)$$

$$\hookrightarrow R(r_{C2j} = i \mid x) = \lambda_c (1 - P(Y=i \mid x))$$

b) Prove $R(r_{C2j} = c+1 \mid x) = \lambda_d$

Ridge classifying error Data points:

$$R(r_{C2j} = c+1 \mid x) = \sum_{j=1}^c L(r_{C2j} = i, r=j) P(Y=j \mid x)$$

If classifier picks wrong learnt ($i = c+1$) loss must be \geq regular loss
of actual class j

$$\hookrightarrow R(r_{C2j} = c+1 \mid x) = \sum_{j=1}^c \lambda_d P(Y=j \mid x)$$

$$\text{All prob of valid pdf} = 1 ; \quad \sum_{j=1}^c P(Y=j \mid x) = 1$$

$$\hookrightarrow R(r_{C2j} = c+1 \mid x) = \lambda_d \cdot 1 = \lambda_d$$

Must be λ_d

2) Show $r_{\text{opt}}(x)$ obtains min Risks:

R1) non-doubt class i: $P(Y=i|x) \geq P(Y=j|x)$

R2) Class i: $P(Y=i|x) \geq 1 - \lambda_2/\lambda_c$

R3) doubt otherwise

Based on part 1:

- if i not in doubt class:

$$R(r(x)=i|x) = \lambda_c (1 - P(Y=r|x))$$

- if class i is doubt

$$R(r(x)=c-1|x) = \lambda_2$$

In order to minimize risks, need to compare 2 values and pick decision with lowest risk:

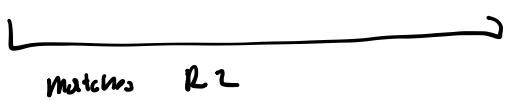
$$\lambda_c (1 - P(Y=i|x)) \leq \lambda_2$$

Dividing both sides by λ_c :

$$1 - P(Y=i|x) \leq \frac{\lambda_2}{\lambda_c}$$

Rearranging:

$$P(Y=i|x) \geq 1 - \frac{\lambda_2}{\lambda_c}$$

 matches R2

* Minimizes overall risk:

$$\text{Scenario 1: } P(Y=i|x) \gg 1 - \frac{\lambda_2}{\lambda_c}$$

\hookrightarrow choosing class i always smaller risk than others

$$\text{Scenario 2: } P(Y=i|x) \ll 1 - \frac{\lambda_2}{\lambda_c}$$

\hookrightarrow risk from misclassification is high so doubt minimizes expected loss

In total:

\Rightarrow always picking the class w/ higher P(i), we can choose best class i before comparing doubts

Conclusion:
 $r_{\text{opt}}(x)$ derived from min overall
 risks by comparing classification risks
 to doubt probability. Find best threshold
 for trade off value.

3) if $\lambda_2 = 0$:

If $\lambda_2 = 0$, then we will either classify x in class i if we are almost certain it is or then choosing j such that this matches what we predicted for our current $v_{opt}(t)$ as it works extremes well as dealing with extremities.

If $\lambda_2 > \lambda_c$:

If $\lambda_2 > \lambda_c$, then we will classify x in class $\{1, \dots, c\}$ that gives the highest probability of correct classification. This matches our $v_{opt}(t)$ since the thresholds we established classifies the x in the class w/ highest prob.

Question 4: Maximum Likelihood Estimation and Bias

1) Max Likelihood Est:

$$L(\nu, \sigma; x) = \prod_{i=1}^n \frac{\sqrt{i}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \nu)^2}{2\sigma^2}\right)$$

$\downarrow \log$

$$\ell(\nu, \sigma; x) = \frac{1}{2} \sum_{i=1}^n \ln\left(\frac{1}{2\pi}\right) - n \ln(\sigma) - \sum_{i=1}^n \frac{(x_i - \nu)^2}{2\sigma^2}$$

$$\frac{d\ell}{d\nu} = \sum_{i=1}^n \frac{(x_i - \nu)}{\sigma^2} = 0$$

$$\hookrightarrow \hat{\nu} = \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n i}$$

$$\frac{d\ell}{d\sigma} = \frac{-n}{\sigma} + \sum_{i=1}^n \frac{(x_i - \nu)^2}{\sigma^3} = 0$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\nu})^2 :$$

$$\hookrightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\nu})^2 :$$

2) Prove/Disprove MLE sample est $\hat{\nu}$ unbiased

Prove

$$\text{bias}(\hat{\nu}) = E[\hat{\nu}] - \nu$$

$$= E\left[\frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n i}\right] - \nu \quad \Rightarrow \quad = \frac{\sum_{i=1}^n \nu_i}{\sum_{i=1}^n i} - \nu \quad \Rightarrow \quad = \nu - \nu$$

$$= E\left[\frac{\sum_{i=1}^n E[x_i]}{\sum_{i=1}^n i}\right] - \nu \quad \Rightarrow \quad = N \frac{\sum_{i=1}^n \nu_i}{\sum_{i=1}^n i} - \nu \quad \Rightarrow \quad \text{Bias} = 0$$

3) Prove / Disprove MLE sample est $\hat{\sigma}^2$ unbiased

Disprove

$$\text{Var}[x] = E[x^2] + E[x]^2$$

$$\hookrightarrow z = \sum_{i=1}^n r_i \quad ; \quad r_i = \frac{n(n+1)}{2}$$

$$\text{Since } \hat{v} = \frac{1}{2} \sum x_i, \text{ Var}[\hat{v}] = \frac{\sigma^2}{2}.$$

$$\hookrightarrow E[\hat{v}^2] = \frac{\sigma^2}{2} + v^2$$

$$\text{Bias}(\hat{\sigma}^2) = E[\hat{\sigma}^2] - \hat{\sigma}^2$$

$$= E\left[\frac{1}{n} \sum_{i=1}^n (x_i - \hat{v})^2\right] - \sigma^2$$

\downarrow Linearity of Expectation

$$= E\left[\frac{1}{n} \sum_{i=1}^n (x_i^2 - 2x_i \hat{v} + \hat{v}^2)\right] - \sigma^2$$

$$= \frac{1}{n} E\left[\sum_{i=1}^n x_i^2 + \sum_{i=1}^n \hat{v}^2 - 2 \hat{v} \sum_{i=1}^n x_i\right] - \sigma^2$$

$\downarrow \hat{v} = \frac{1}{2} \sum x_i$

$$\Rightarrow \frac{1}{n} E\left[x_i^2 + 2\hat{v}^2 - 2\hat{v}\hat{v}\right] - \sigma^2$$

$$= \frac{1}{n} \left[\sum_{i=1}^n E[x_i^2] - 2E[\hat{v}^2] \right] - \sigma^2$$

$$= \frac{1}{n} \left[\sum_{i=1}^n \left(\frac{\sigma^2}{i} + v^2 \right) - 2 \left(\frac{\sigma^2}{2} + v^2 \right) \right] - \sigma^2$$

$$\begin{aligned}
 &= \frac{1}{n} \left[n\sigma^2 + 2\mu^2 - \sigma^2 - 2\mu^2 \right] - \sigma^2 \\
 &= \frac{n-1}{n} \sigma^2 - \sigma^2 \\
 &= \underbrace{-\frac{1}{n}}_{\text{bias}} \sigma^2
 \end{aligned}$$

Thus, the MLE of σ^2 is biased

4) Variance for $\nu = \text{only est } \nu$. What will be MLE of ν ?

We know:

$$\hat{\nu} = \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n i}; \quad E[\hat{\nu}] = \nu \quad \text{w/ bias} = 0$$

$$\begin{aligned}
 \text{Var}[\hat{\nu}] &= \left(\sum_{i=1}^n \frac{1}{\sigma^2 i} \right)^{-1} = \left(\frac{n(n+1)}{2\sigma^2} \right)^{-1} \\
 &= \frac{2\sigma^2}{n(n+1)}
 \end{aligned}$$

$\hookrightarrow \text{Bias} = 0$

$$E(\hat{\nu}) = \text{Var}(\hat{\nu})$$

$$E(\hat{\nu}) = \frac{2\sigma^2}{n(n+1)}$$

Question 5: Covariance Matrices and Decomposition

1) $\hat{\Sigma}$ is not invertible only if all points lie on a common hyperplane in feature space

2) Fix singular covariance Matrix:

↳ Projecting the covariance matrix onto a lower-D subspace by removing features that are linear combinations of remaining features. Continue the process until all features are independent

3) Max/Min PDF $f(x)$

Max: x will max PDF for it the eigenvector of Σ is the largest eigenvalue.

Min: x will min PDF for it the eigenvector of Σ is the smallest eigenvalue.

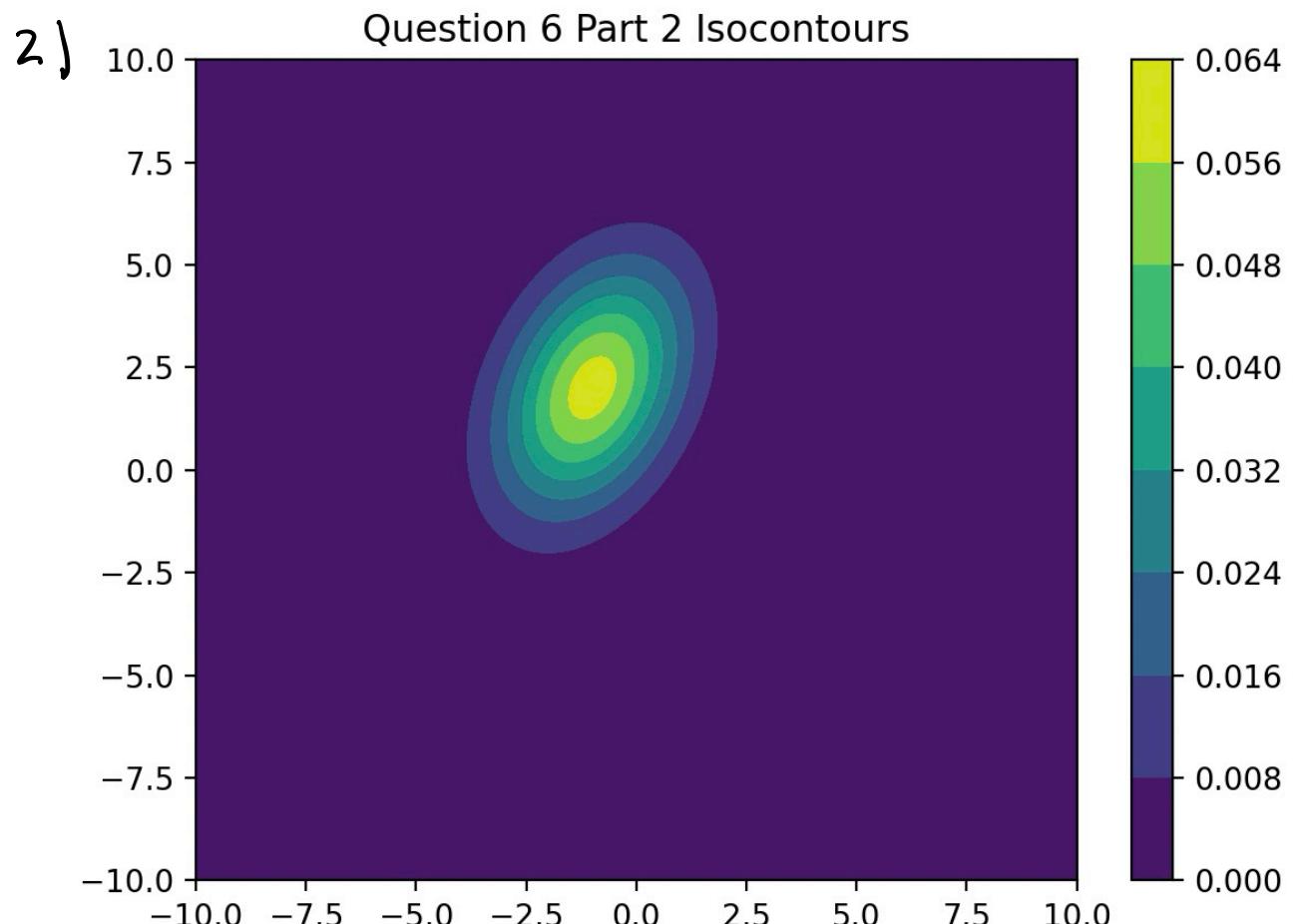
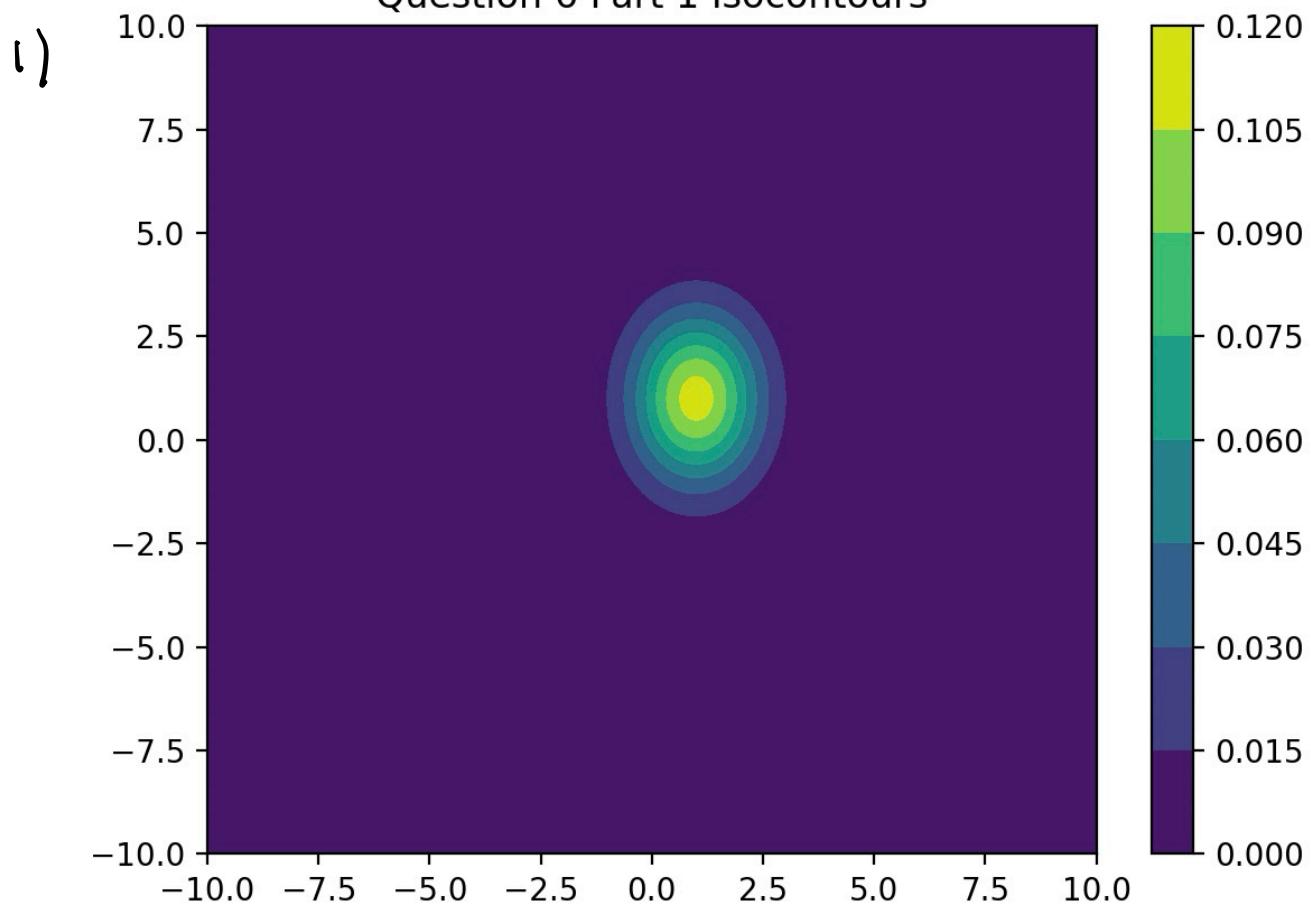
* Σ^{-1} penalizes distance from the mean $\mu = 0$ least in the direction of eigenvectors with smaller eigenvalues

$$4) E[\rho] = E[y^T x] \stackrel{L.E}{=} y^T E[x]$$

$$\begin{aligned} \text{Var}[\rho] &= E[(\rho - E[\rho])^2] = E[(\rho - E[\rho])(\rho - E[\rho])^T] \\ &= E[y^T(x - E[x])(x - E[x])^T y] \\ &\stackrel{L.E}{=} y^T E[(x - E[x])(x - E[x])^T] y \\ &= y^T \Sigma y \end{aligned}$$

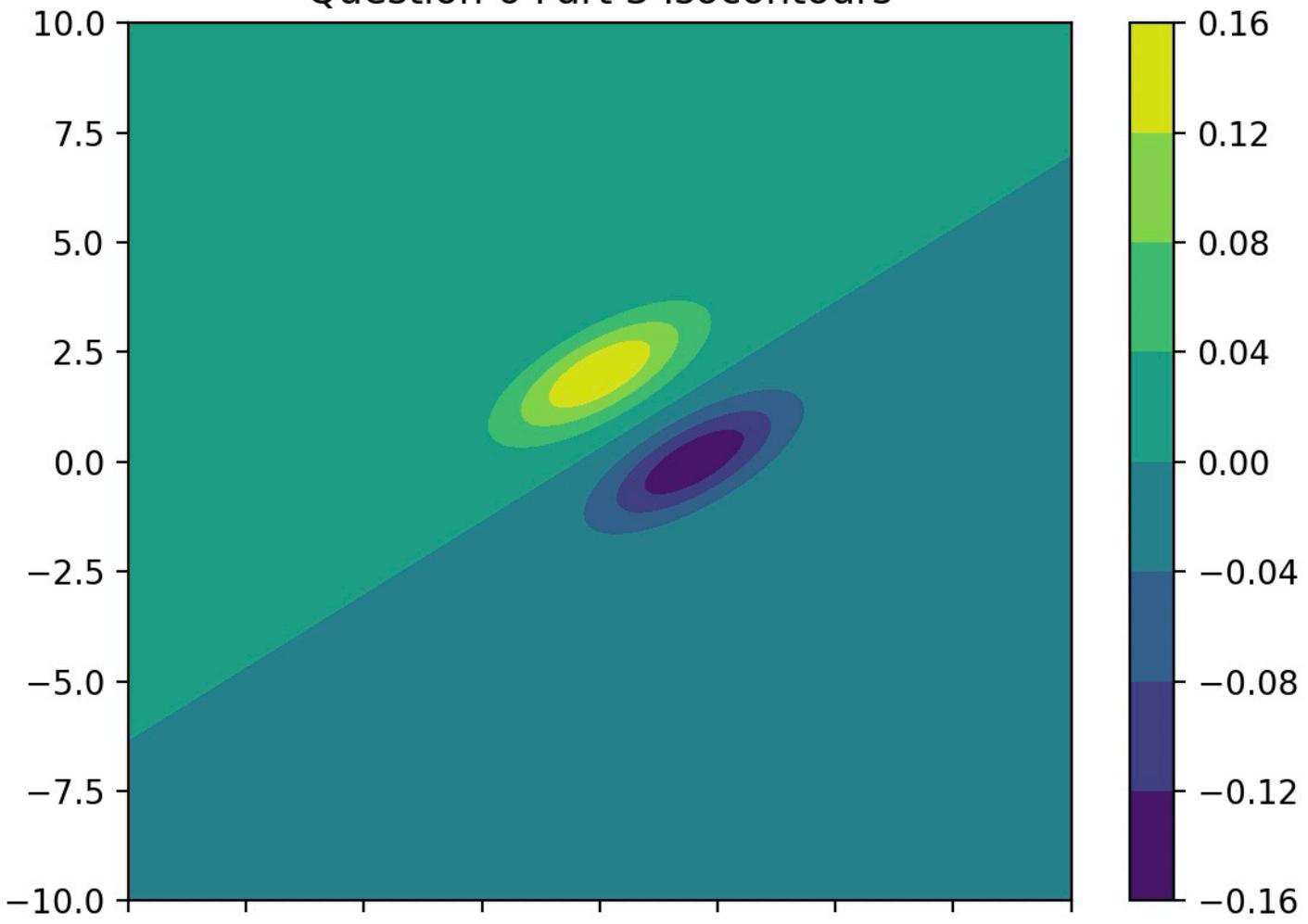
↳ * Largest eigenvalue λ_{\max} tells me the direction of max variance

Question 6 : Iso contours of Normal Distributions



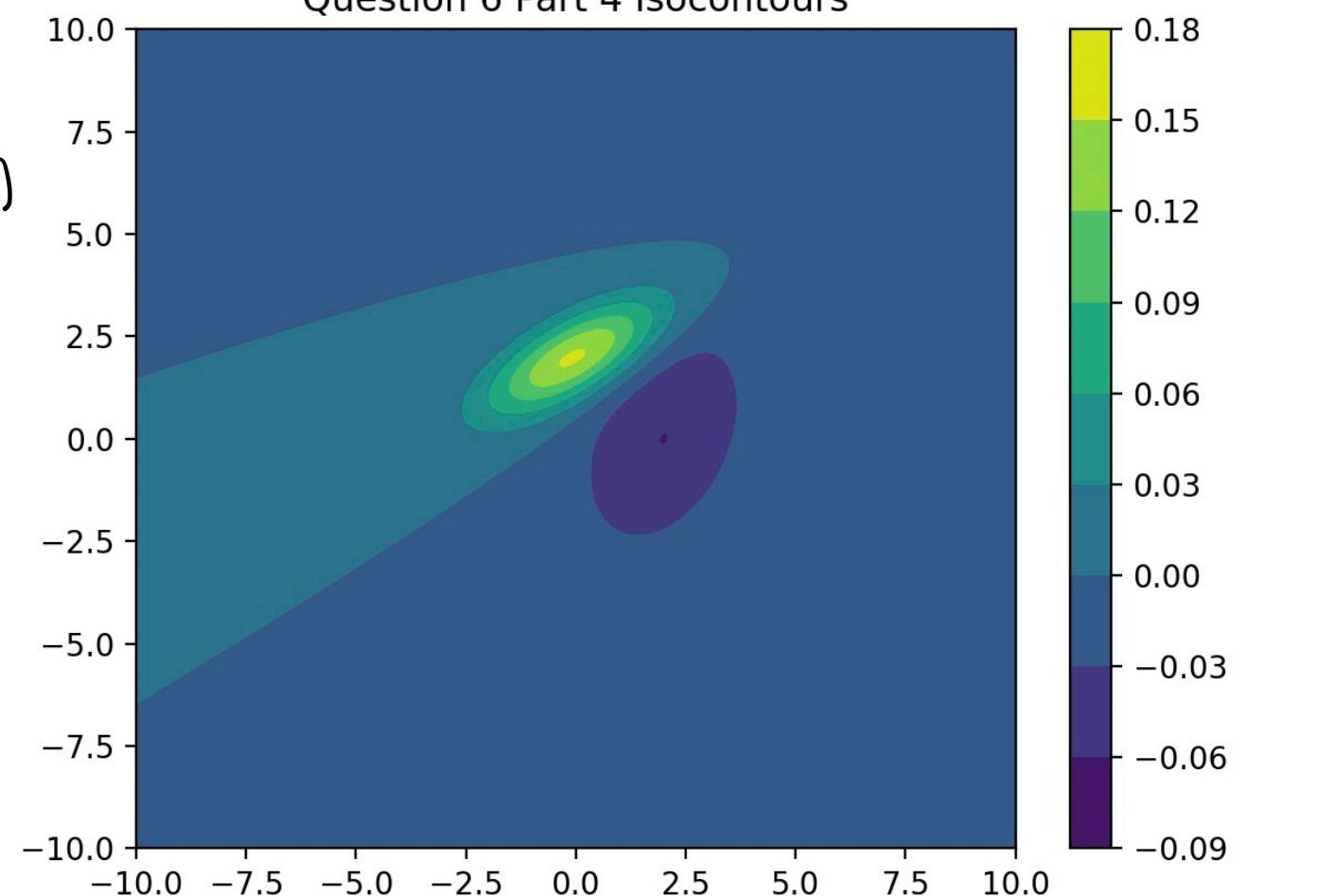
3)

Question 6 Part 3 Isocontours

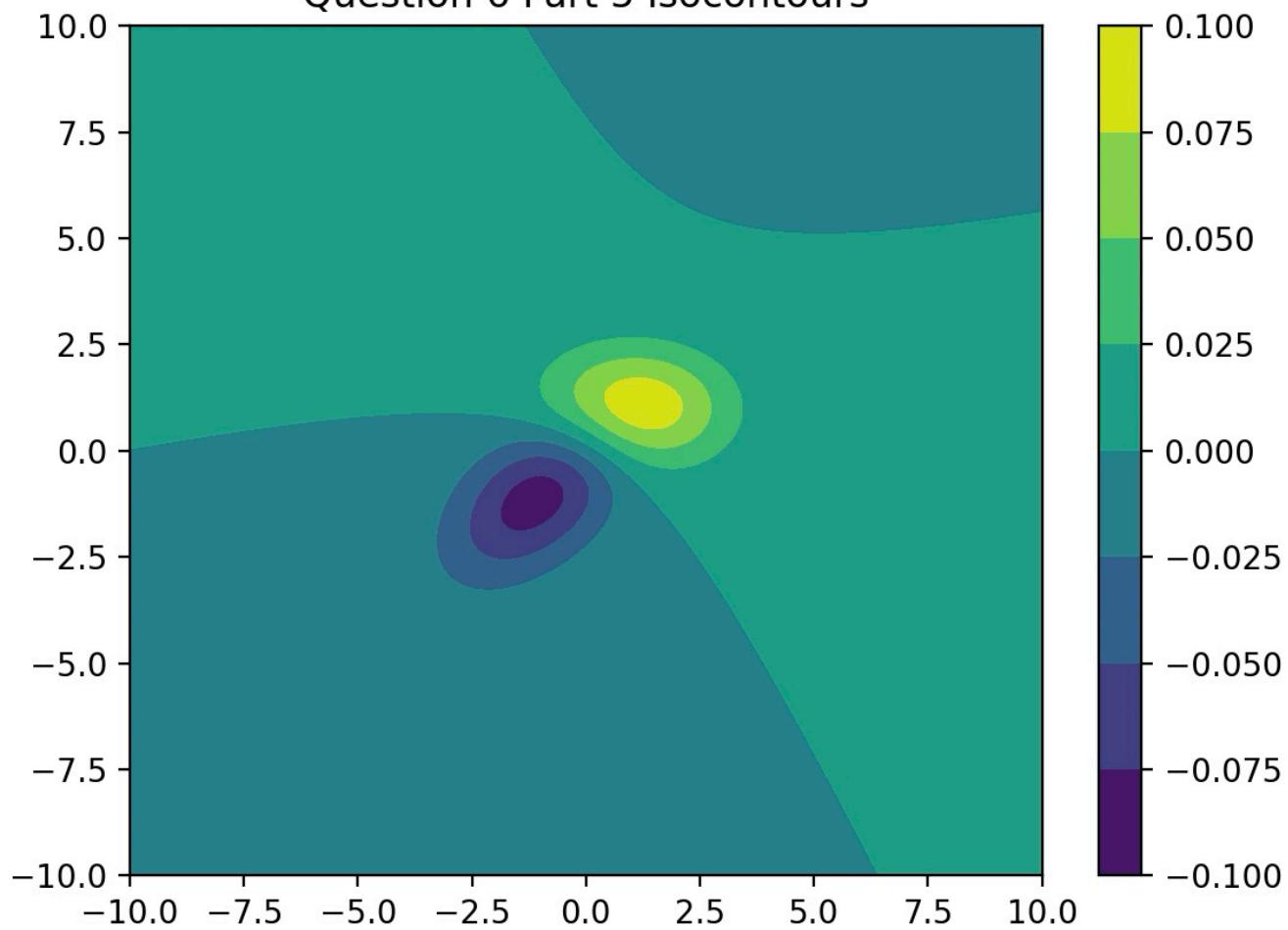


4)

Question 6 Part 4 Isocontours

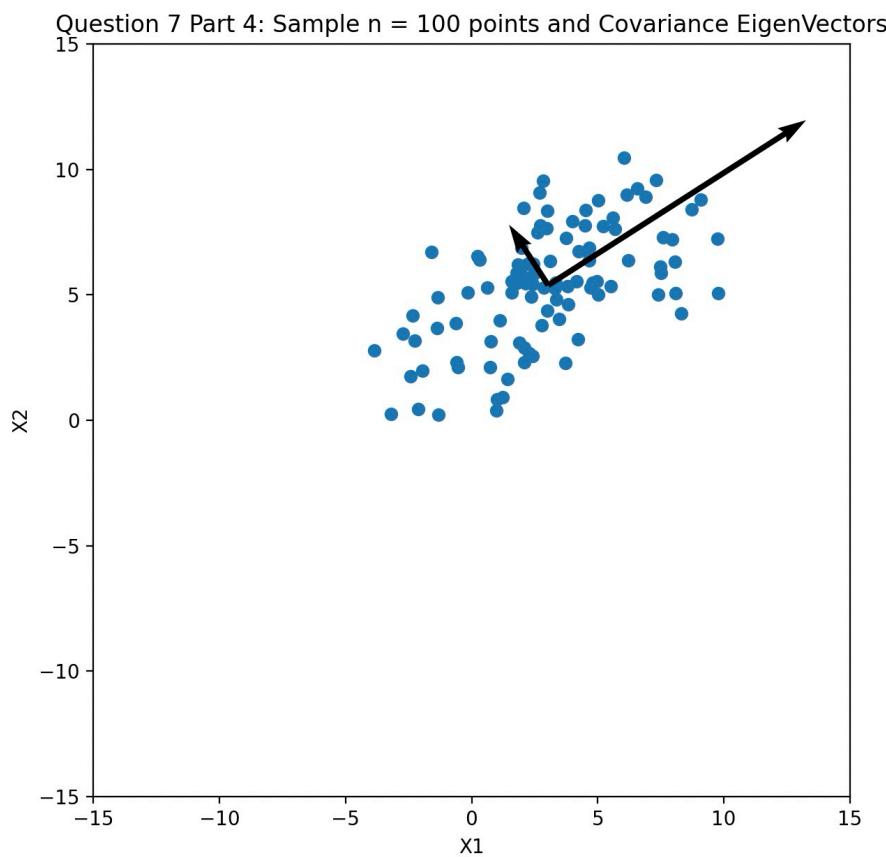


ζ Question 6 Part 5 Isocontours

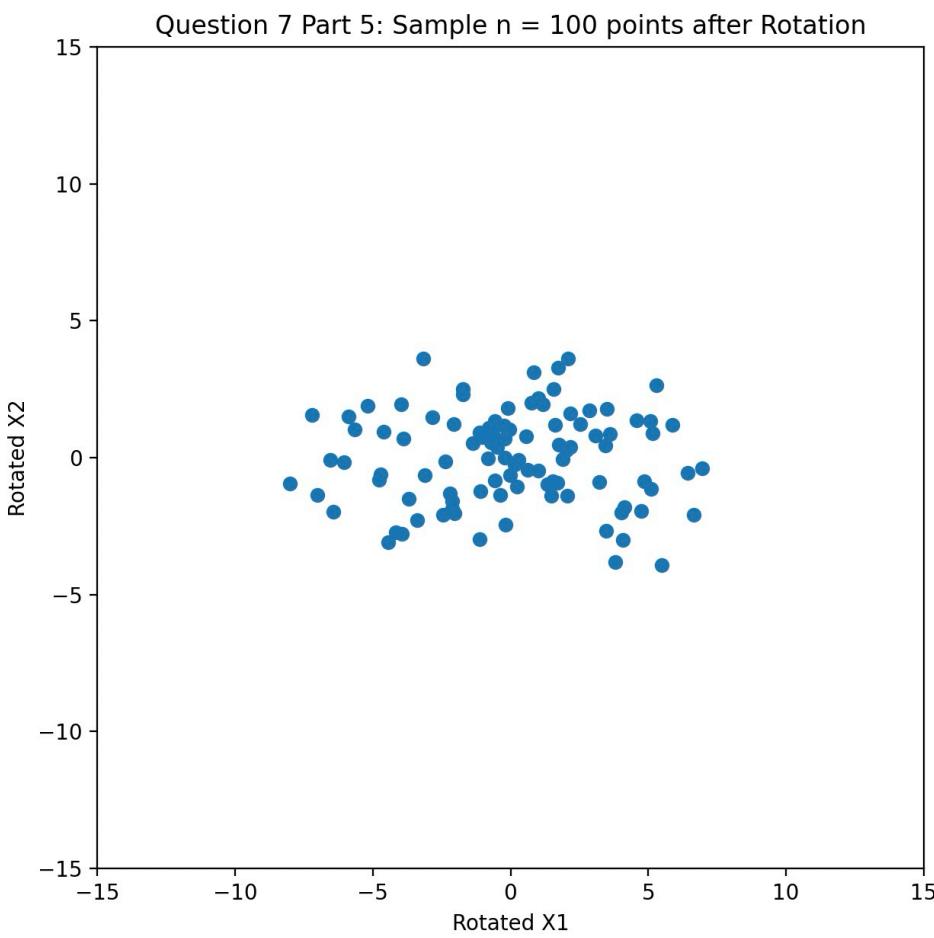


Question 7: Eigen Vectors of Gaussian Cov Matrix

4)

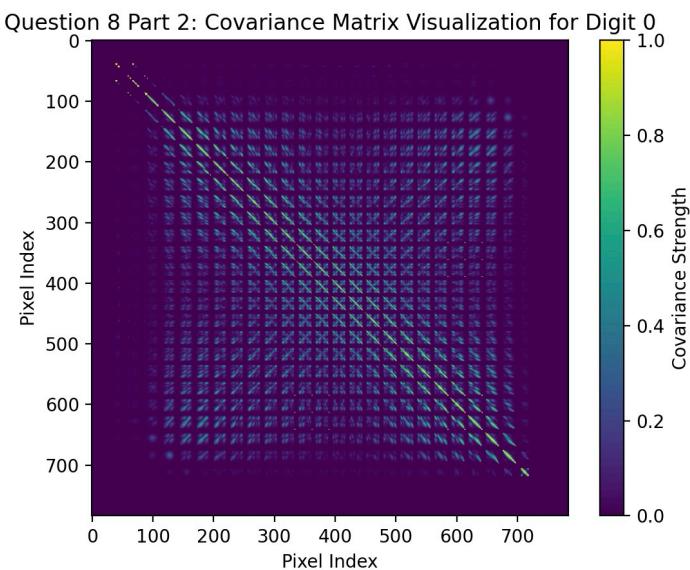


5)



Question 8: Gaussian Classifier for Digits and Spam

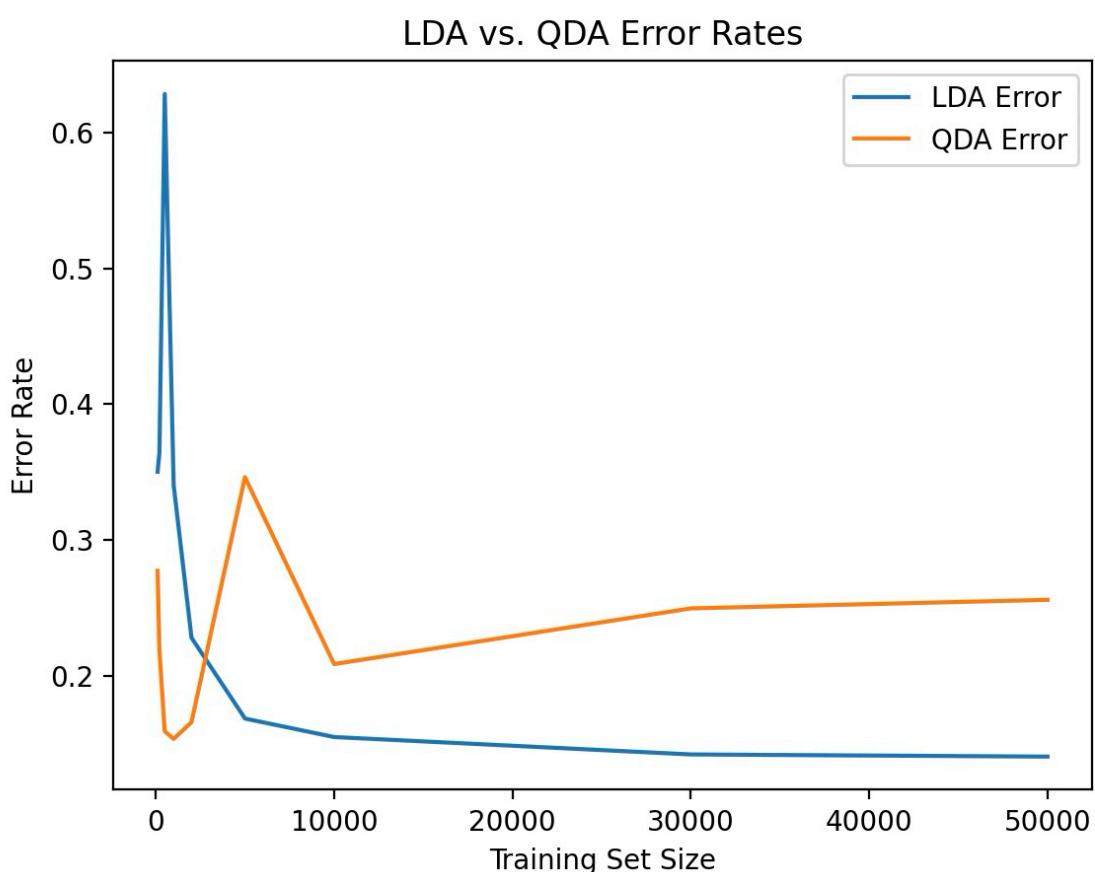
2)



* In my code, I selected digit 0 shown by me idx. looking at my graph, I visually see that diagonal terms are generally increasing in cov strength than off-diagonal terms. This tells me that cov for neighboring pixels are greater than those that are far.

3)

a)
b)

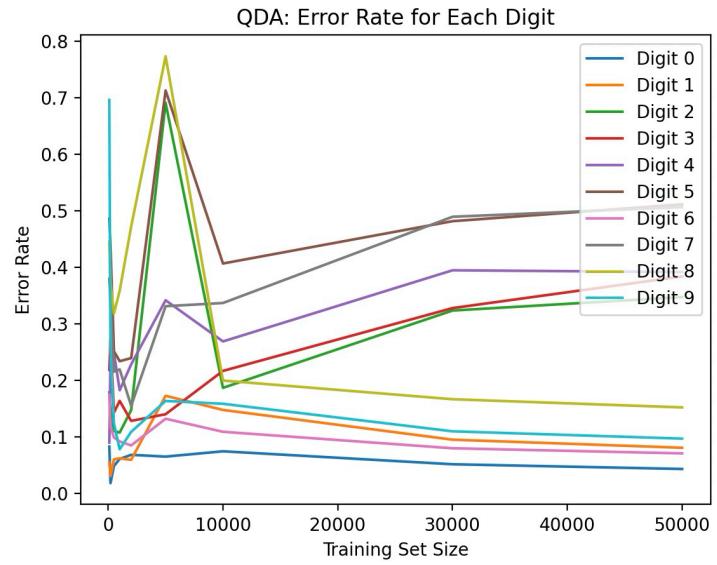
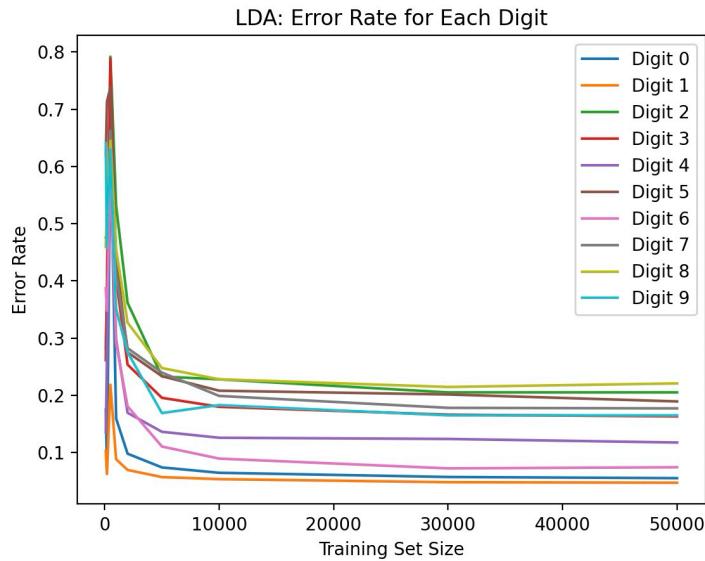


c) Looking at both graph's for LDA/QDA error rate, while initially with smaller training size, QDA and LDA are relatively similar, it seems to me that with higher training size and more contours, that LDA seems to be the better bDA for this scenario.

As mentioned in lecture, QDA tends to overfit while LDA tends to underfit. Since LDA seems to perform better, I suspect the data has a more linear Bayesian Decision Bound which favors LDA.

This is also shown with LDA performing significantly better than QDA.

d)



Based on our GDA on MNIST, it would appear that **Digit 1** is easiest for LDA and **Digit 0** easiest for QDA. This informs us that digit 1 has more distinct features allowing an easier separation by a linear boundary for LDA. Digit 0 for QDA informs us that it has more or a flaccidive wavy boundary helping the model.

Kaggle Submissions

UC Berkeley CS 189 HW3 (MNIST) Spring 2025

MNIST submissions for HW 3 for Spring 2025 CS 189



Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Submissions

Select up to 14 submissions that will count towards your final leaderboard score. If less than 14 are selected, Kaggle will automatically select from your best scoring submissions. [Learn More](#)

1/14

Auto-selection candidates [?](#)

All Successful Selected Errors

Recent ▾

Submission and Description

Public Score [?](#)

Select

MNIST_Test_Labels.csv

Complete · 25s ago · Daniel Kim MNIST Kaggle HW3 Submission #2

0.862

UC Berkeley CS189 HW3 (SPAM) Spring 2025

Kaggle competition for SPAM hw3 for CS 189 2025



Overview Data Code Models Discussion Leaderboard Rules Team Submissions

Submissions

Select up to 14 submissions that will count towards your final leaderboard score. If less than 14 are selected, Kaggle will automatically select from your best scoring submissions. [Learn More](#)

1/14

Auto-selection candidates [?](#)

All Successful Selected Errors

Recent ▾

Submission and Description

Public Score [?](#)

Select

spam_Test_Labels.csv

Complete · 24s ago · Daniel Kim SPAM Kaggle HW3 Submission #2

0.786