HW 3

Daniel Kim
3035781607

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

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

$P(Y=C_1) = P(Y=C_2) = \frac{1}{2}$; $\mu_2 > \mu_1$

1) Bayes optimal decision boundary

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

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

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

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

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

* Decision Boundary is the point equidistant to μ_1 and μ_2

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

Loss function:
- Class 1 ; $x < \frac{\mu_1 + \mu_2}{2}$
- Class 2 ; otherwise

2) Prove: Bayes Error

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

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

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

∟, $P(C_1) = P(C_2) = \frac{1}{2}$

↓

PDF :

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

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

Bayes Error:

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

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

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

* b^* must be between μ_1 and μ_2 *

Min $P_e(b)$:

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

Fundamental theorem of calculus:

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

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

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

$\downarrow$

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

$$\log\left(f_1(b^*)\right) = \log\left[f_2(b^*)\right]$$

$$\frac{(b^* - \mu_2)^2}{2\sigma^2} = \frac{(b^* - \mu_1)^2}{2\sigma^2}$$

$$(b^* - \mu_2)^2 = (b^* - \mu_1)^2$$

$$\therefore \quad b^* = \frac{\mu_1 + \mu_2}{2}$$

Question 3: Classification and Risks

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

Loss function:

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

$\lambda_c \geq 0$ loss from misclassification

$\lambda_d \geq 0$ loss from doubting

- i) correct = no penalty
- ii) Incorrect = λ_c penalty
- iii) Doubt: λ_d penalty

Risk Classifying new Data point:

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

1) Simplify Risk function

a) Prove $R(r(x) = i \mid x) = \lambda_c (1 - P(Y = i \mid x))$ if doubt ie $i \neq c+1$

Risk Classifying new Data point:

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

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

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

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

Both cases:

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

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

All prob of will pdf = 1 $\therefore \sum_{j=1}^{c} P(Y = j \mid x) = 1$

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

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

b) Prove $R(r(x) = c+1 \mid x) = \lambda_d$

Risk classifying new Data point:

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

If classifier picks doubt $(i = c+1)$ loss must be λ_d regardless of actual class j

↳ $R(r(x) = c+1 \mid x) = \sum_{j=1}^{c} \lambda_d P(Y = j \mid x)$

All prob of will pdf = 1 $\therefore \sum_{j=1}^{c} P(Y = j \mid x) = 1$

↳ $R(r(x) = c+1 \mid x) = \lambda_d \cdot 1 = \lambda_d$

* Must be λ_d

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

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

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

R3) doubt otherwise

Based of part 1:

- if i not in doubt classes:

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

- if classify as doubt

$$R(r(x)=c+1|x) = \lambda_d$$

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

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

dividing both sides by λ_c:

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

Rearranging:

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

matches R2

Conclusion:

$r_{opt}(x)$ derived from min expected risk by comparing classification risk to doubt penalty. Finds best threshold for trade off value.

* Minimizes error rate:

Scenario 1: $P(Y=i|x) >> 1 - \frac{\lambda_d}{\lambda_c}$

↳ choosing class i incurs smaller risk than doubt

Scenario 2: $P(Y=i|x) << 1 - \frac{\lambda_d}{\lambda_c}$

↳ risk from misclassification is high so doubt minimizes expected loss

In total:

By always picking the class w highest prob, we ensure choosing best class i before comparing doubt

3) if $\lambda_3 = 0$:

if $\lambda_3 = 0$, then we will either classify x in class i if we are almost certain of it or then choosing another, this matches what we predicted for our current $\hat{y}_{opt}(x)$ as it works extremely well at 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 $\hat{y}_{opt}(x)$ since the threshold we established classifies the x in the class w/ highest prob.

Question 4: Maximum Likelihood Estimation and Bias

1) Max Likelihood Est:

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

$\downarrow$ log

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

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

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

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

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

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

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

Prove

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

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

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

$$= \frac{\sum_{i=1}^{n} \nu i}{\sum_{i=1}^{n} i} - \nu$$

$$= \nu \frac{\sum_{i=1}^{n} \nu i}{\sum_{i=1}^{n} i} - \nu$$

$$= \nu - \nu$$

$$= 0$$

$\hookrightarrow$ Bias $= 0$

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

Disprove

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

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

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

$$\therefore E[\hat{\mu}^2] = \frac{\sigma^2}{2} + \mu^2$$

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

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

↓ linearity of Expectation

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

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

↓ $\hat{\mu} = \frac{1}{2}\sum x_i \, i$

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

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

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

$$= \frac{1}{n}\left[n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right] - \sigma^2$$

$$= \frac{n-1}{n}\sigma^2 - \sigma^2$$

$$= \frac{-1}{n}\sigma^2$$

Thus, the MLE of σ^2 is biased

4) Variance Fair γ = only est μ. What kind of MLE of μ?

We know:

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

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

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

↳ Bias = 0

$$R(\hat{\mu}) = \text{Var}(\hat{\mu})$$

$$R(\hat{\mu}) = \frac{2\sigma^2}{n(n+1)}$$

Question 5: Covariance Matrix 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 f(x) if the eigenvector of Σ is the largest eigenvalue.

Min: x will min PDF f(x) if the eigenvector of Σ is the smallest eigenvalue.

* Σ^{-1} penalizes distance from the mean $\mu = 0$ largest in the direction of eigenvector w/ smallest eigenvalue

4) $E[p] = E[y^T x] \overset{L.O.E}{=} y^T E[x]$

$$\text{Var}[p] = E[(p - E[p])^2] = E[(p - E[p])(p - E[p])^T]$$

$$= E[y^T (x - E[x])(x - E[x])^T y]$$

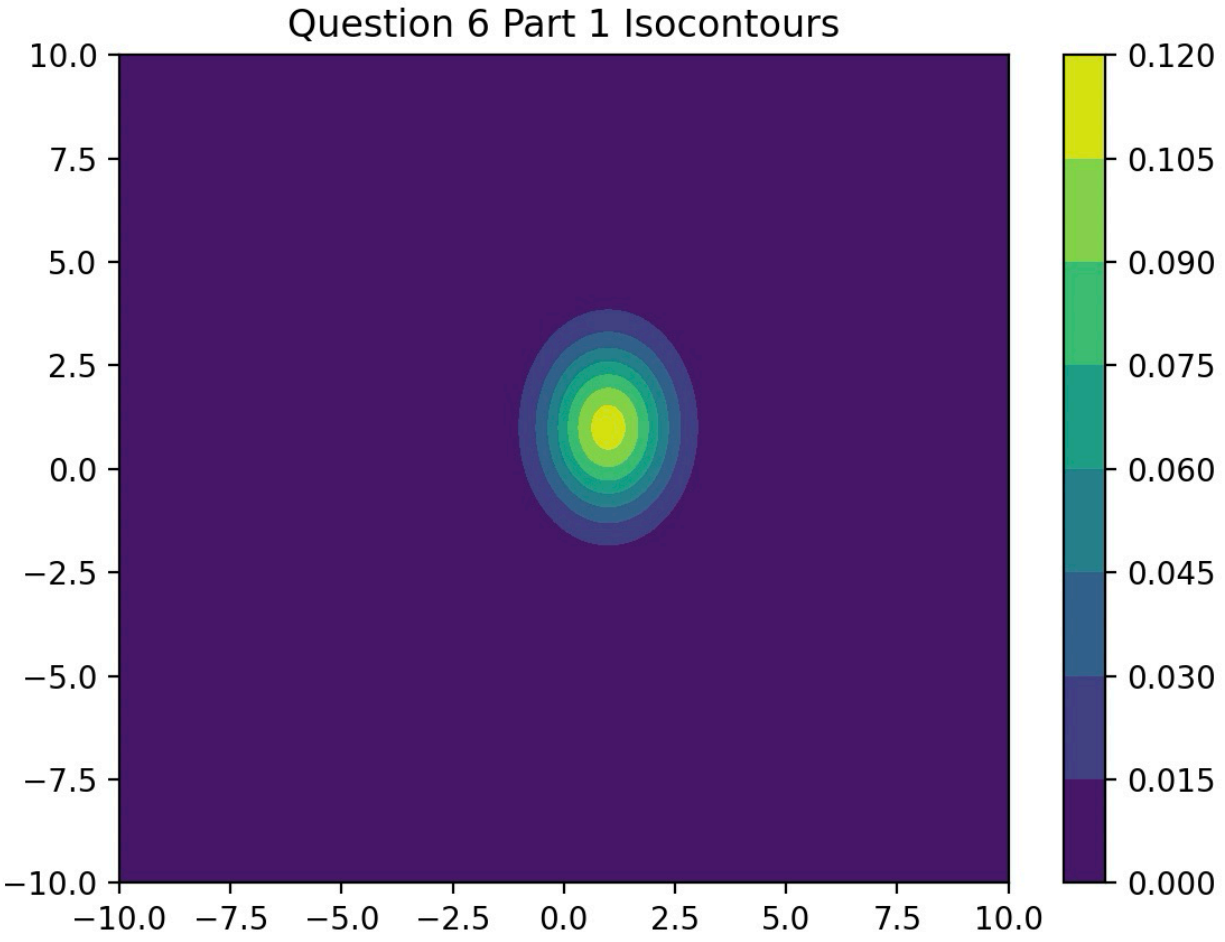
$$\overset{LOE}{=} y^T E[(x - E[x])(x - E[x])^T] y$$

$$= y^T \Sigma y$$

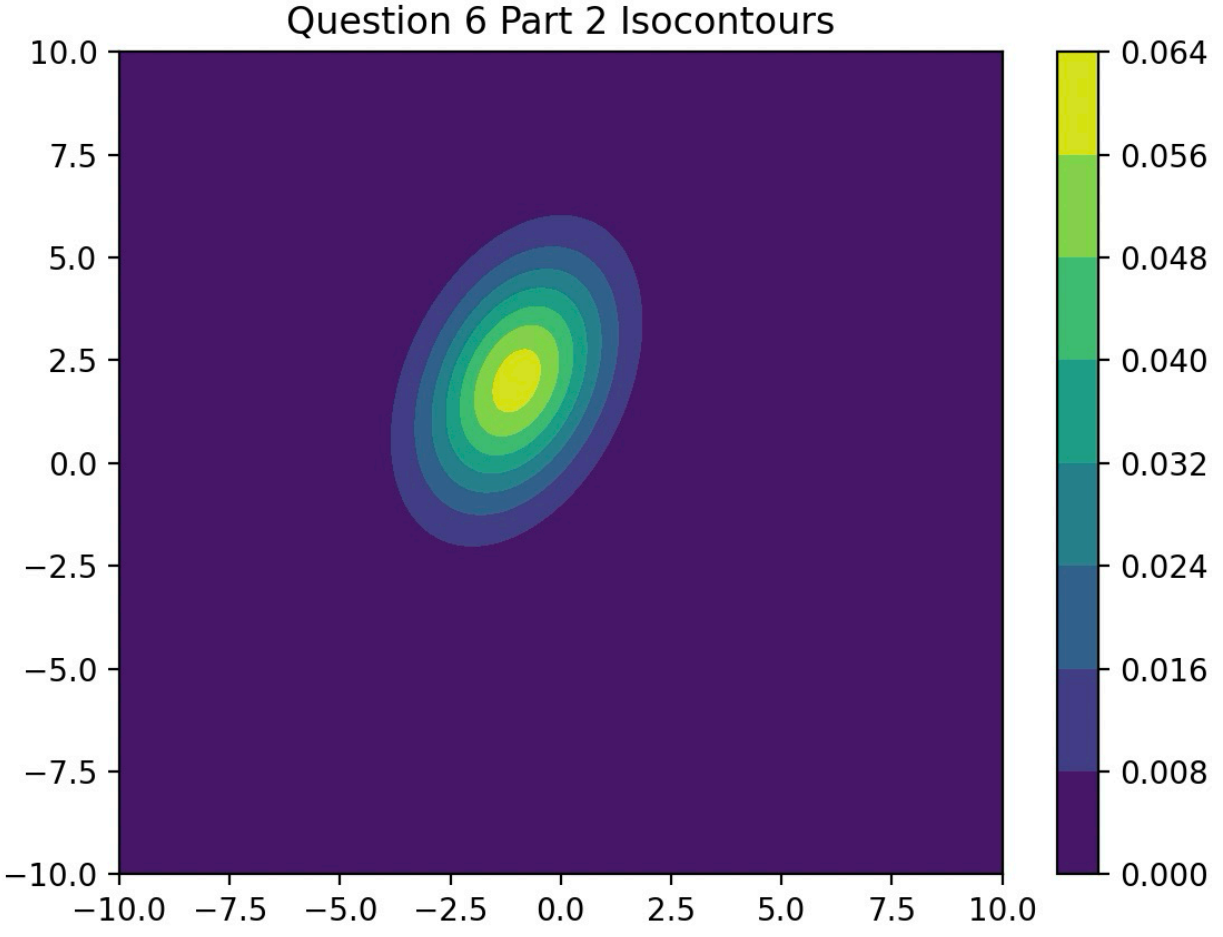
↳ * Largest eigenvalue λ_{max} tells me the direction of max variance

Question 6 : Isocontours of Normal Distributions

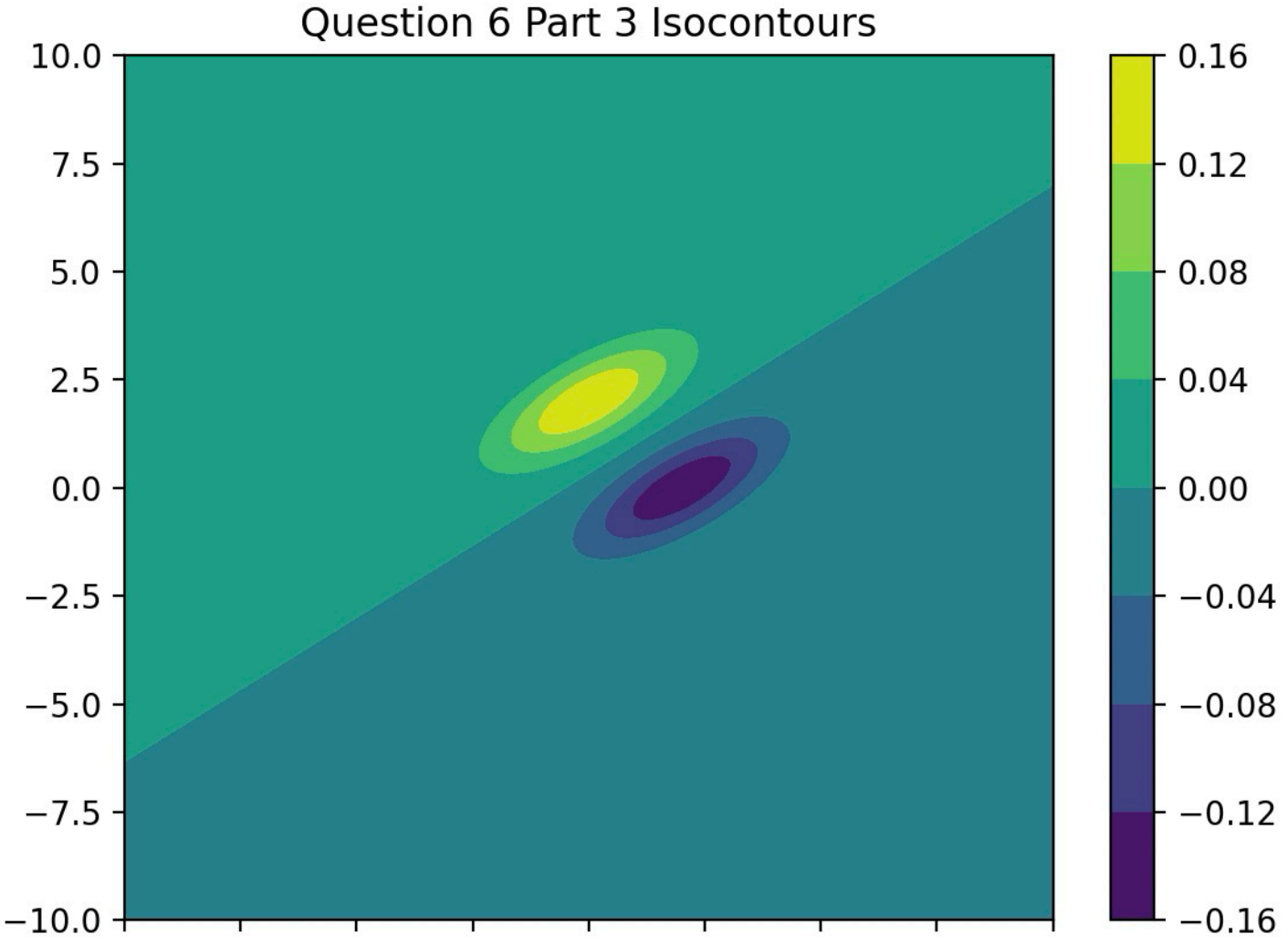
1)



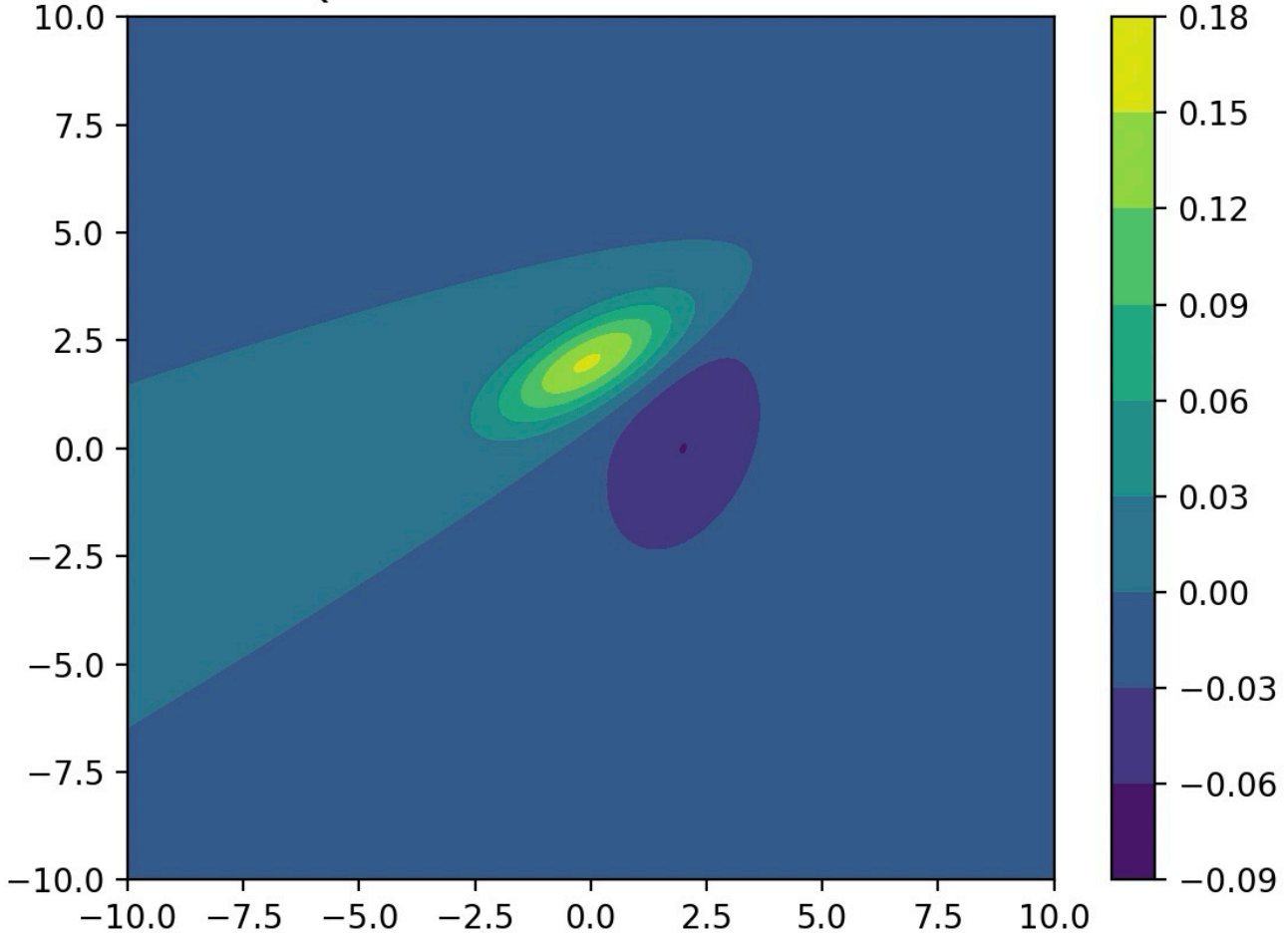
2)



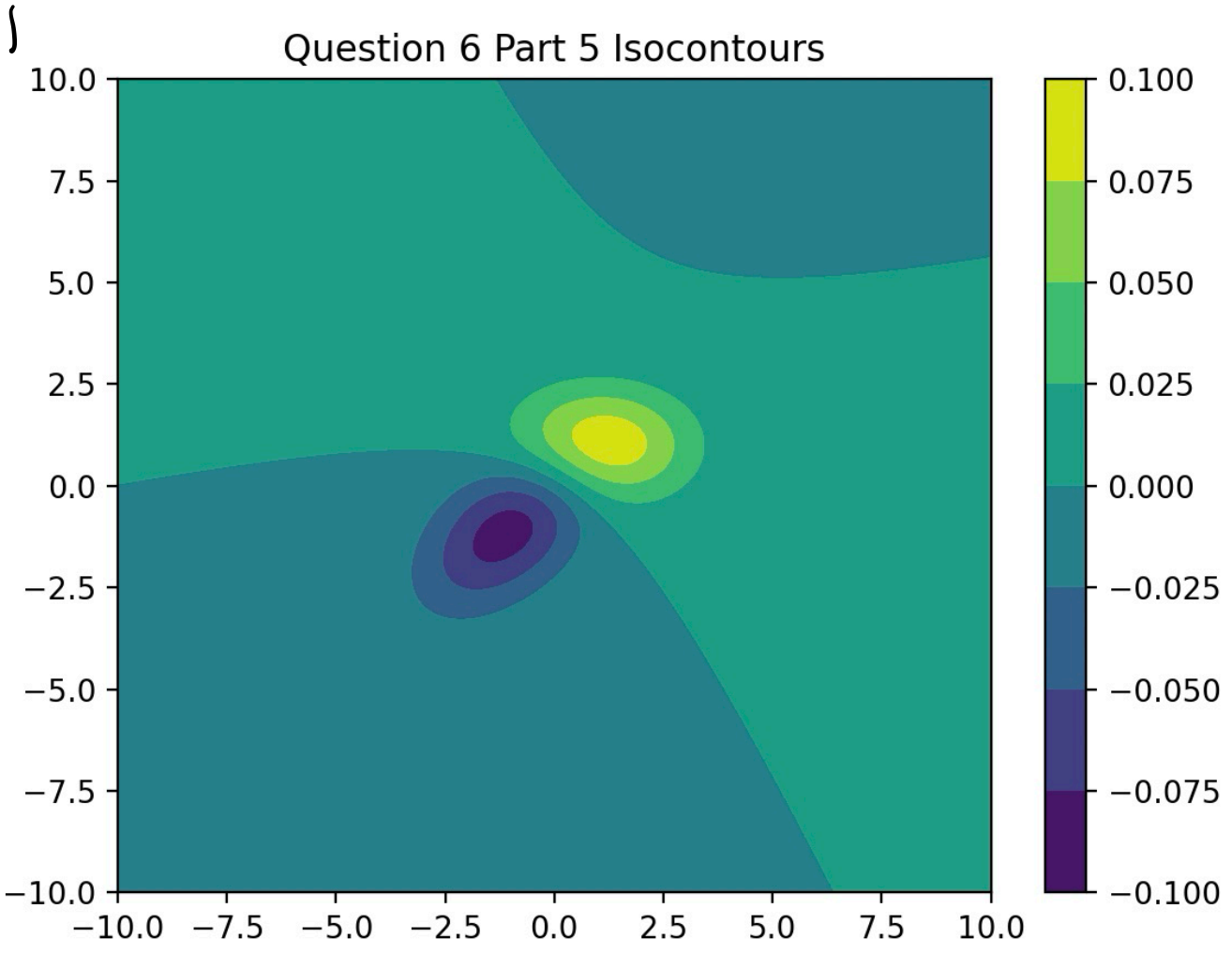
3)


Question 6 Part 3 Isocontours
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
0.16
0.12
0.08
0.04
0.00
−0.04
−0.08
−0.12
−0.16

4)

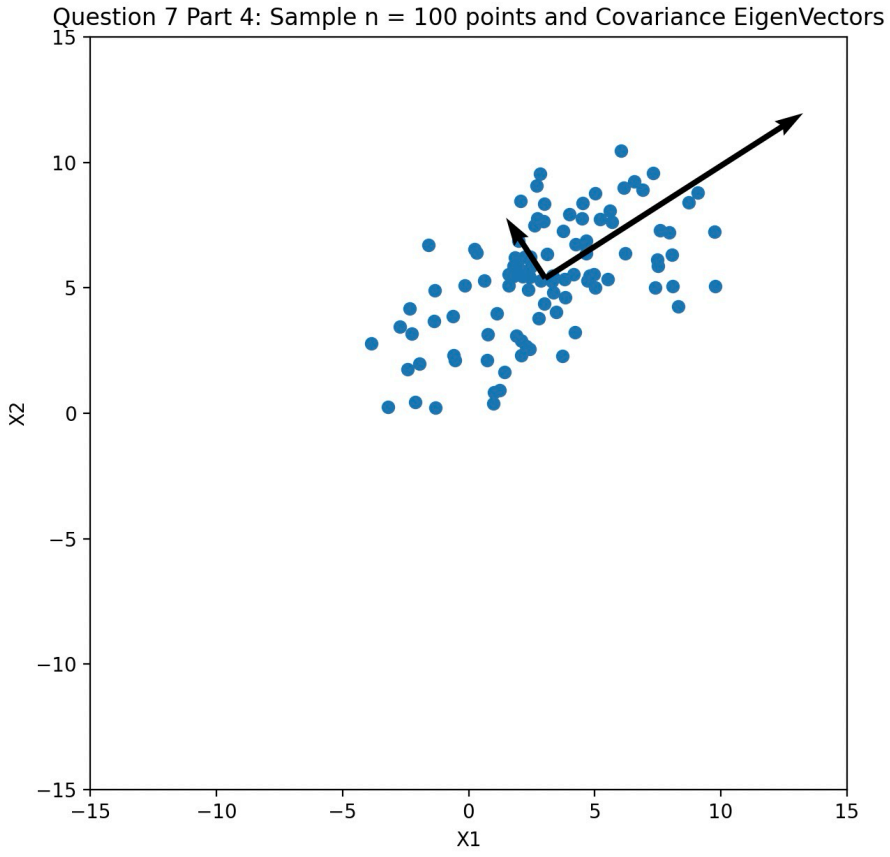

Question 6 Part 4 Isocontours
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
0.18
0.15
0.12
0.09
0.06
0.03
0.00
−0.03
−0.06
−0.09

5)

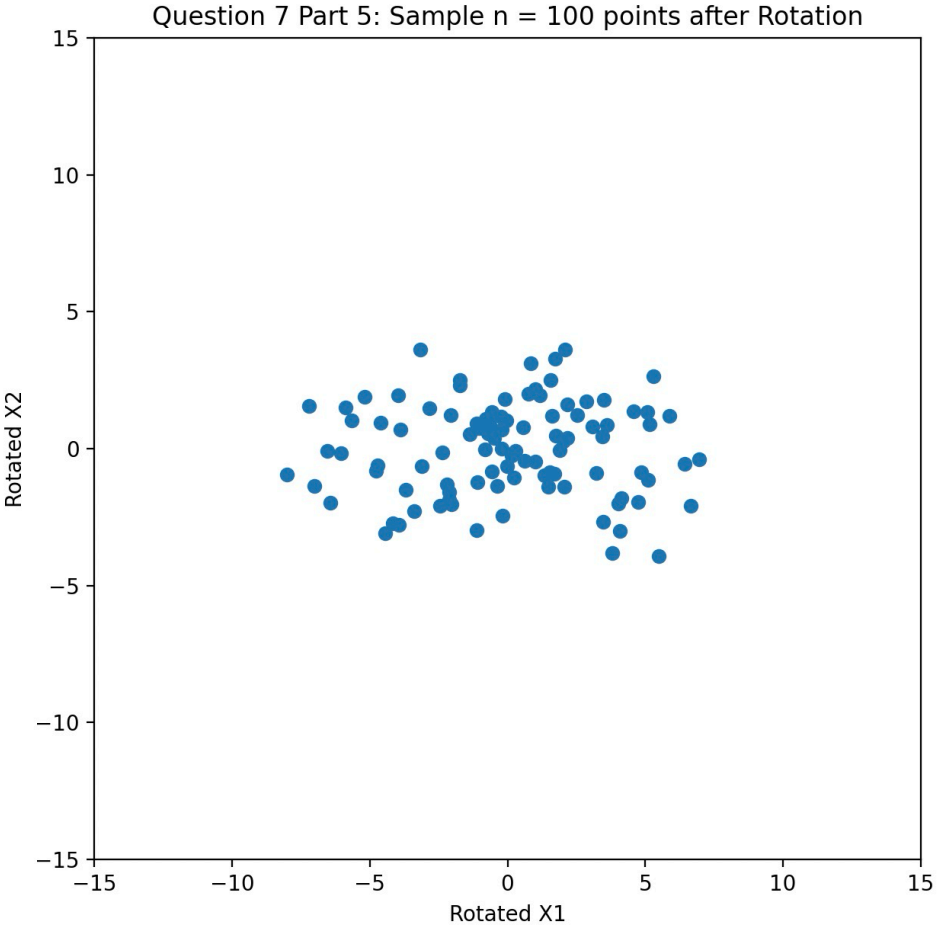

Question 6 Part 5 Isocontours
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
0.100
0.075
0.050
0.025
0.000
−0.025
−0.050
−0.075
−0.100

Question 7: Eigen Vectors of Gaussian Cov Matrix

4)

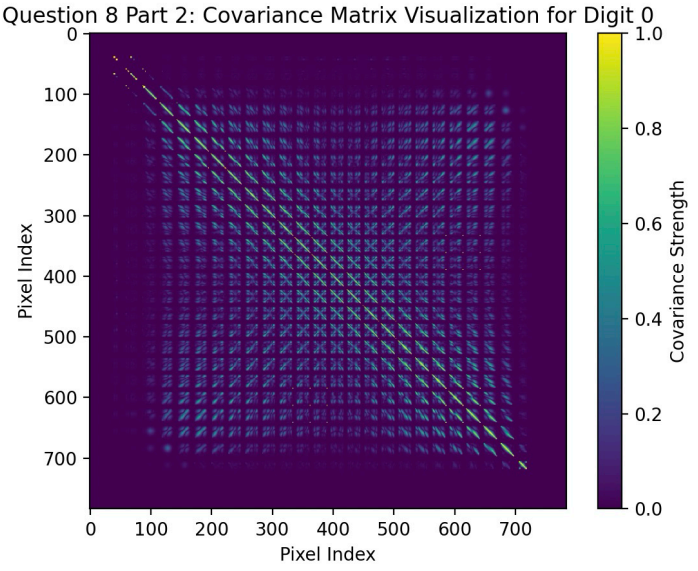


5)



Question 8: Gaussian Classifier for Digits and Spam

2)

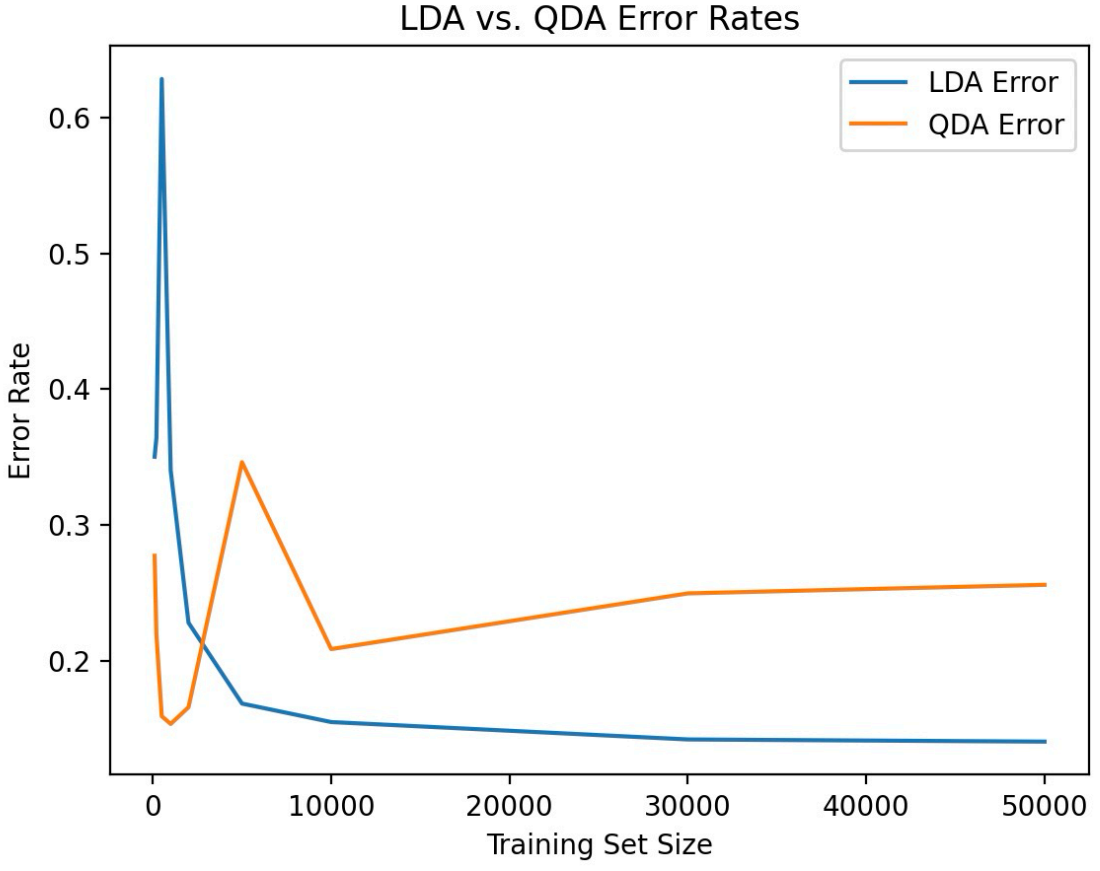


* In my code, I saved digit 0 shown by my idx. looking at my graph, I visually see that diagnol terms are generally increasing in cov strength than off-diagnol terms. This tells me that cov for neighboring pixels are greater than those that are far.

3)

a)

b)

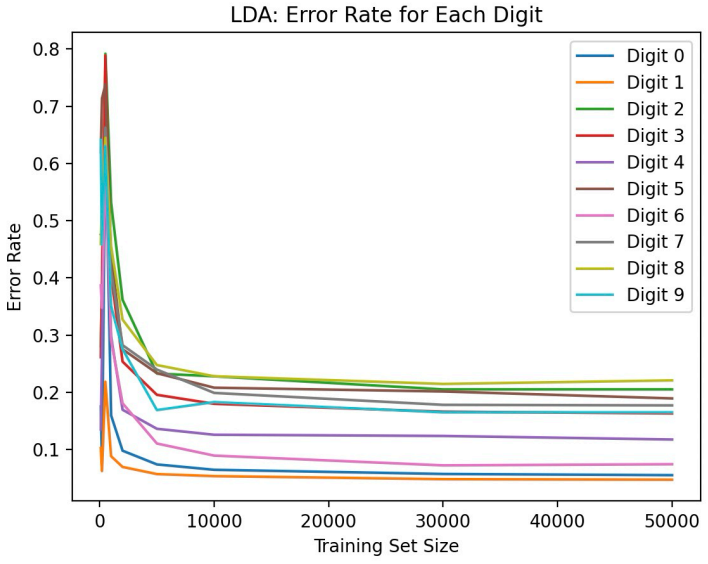


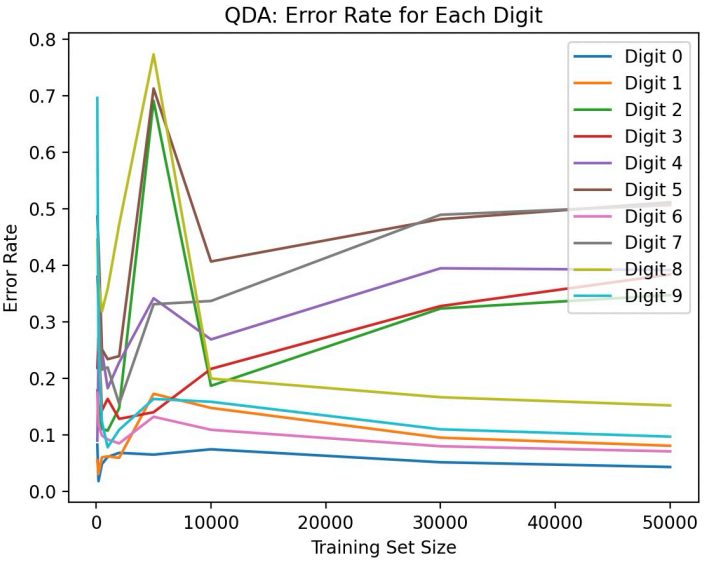
c) Looking at both graph's for LDA/QDA error rate, while initially with smaller training size, QDA and LDA are relatively similar, it seems to me that with higher training size and more certainty, that LDA seems to be the better LDA for this scenario.

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 Boundry which favors LDA.

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

d)





Based on our LDA on MNIST, It would appear that digit 1 is easiest for LDA and Digit 0 easiest for QDA. This informs me that digit 1 has more distinct features allowing an easier separation by a linear boundary for LDA. Digit 0 for QDA informs me that it has more of a flexible quad 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

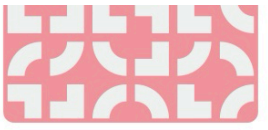
All Successful Selected Errors

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	☑