CARNEGIE MELLON UNIVERSITY 10-315
HOMEWORK 10

DUE: Thursday, April 17, 2025

https://www.cs.cmu.edu/~10315

INSTRUCTIONS

- **Format:** Use the provided LaTeX template to write your answers in the appropriate locations within the *.tex files and then compile a pdf for submission. We try to mark these areas with STUDENT SOLUTION HERE comments. Make sure that you don't change the size or location of any of the answer boxes and that your answers are within the dedicated regions for each question/part. If you do not follow this format, we may deduct points.

  You may also type you answer or write by hand on the digital or printed pdf. Illegible handwriting will lead to lost points. However, we suggest that try to do at least some of your work directly in LaTeX.

- **How to submit written component:** Submit to Gradescope a pdf with your answers. Again, make sure your answer boxes are aligned with the original pdf template.

- **How to submit programming component:** See section Programming Submission for details on how to submit to the Gradescope autograder.

- **Policy:** See Piazza for updated policy for this assignment.

Name	
Andrew ID	
Hours to complete all components (nearest hour)	

1 [5 pts] K-Means

1.1 Non-increasing with K

Given a set of N observations $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$, and cluster centers $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K$, assign each observation to its closest center. Each point has corresponding cluster assignment, $z_i \in \{1, \ldots, K\}$ for i-th point.

Consider the K-means optimization as a function of the hyperparameter for the number of clusters, K:

$$J(K) = \min_{\boldsymbol{\mu}_1,...,\boldsymbol{\mu}_K, z_1,...,z_N} \sum_{i=1}^{N} ||\mathbf{x}^{(i)} - \boldsymbol{\mu}_{z_i}||_2^2$$

Prove that $J(K)$ is a non-increasing function of $K \in \mathbb{Z}^+$. You write your proof in a series of sentences. To receive full credit, each statement in your proof should have sound reasoning and the proof must be complete, e.g., cover all cases for K.

Hint: Try to consider two cases: 1) $K \geq N$, and 2) $K < N$.

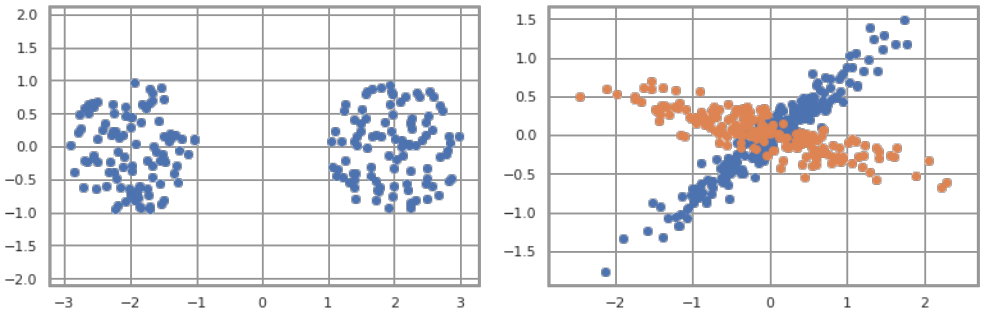
Your Answer

2 [18 pts] PCA

2.1 [6 pts] PCA Warm-Up

1. **[3 pts]** Principal component analysis is a dimensionality reduction method that projects a dataset into its most variable components. You are given the following 2D datasets, draw the first and second principal components on each plot. Be sure to label which vector is the first principal component and which is the second.

Update plots



2. **[2 pts]** Assume we are given a dataset for which the eigenvalues of the covariance matrix are: (2.1, 1.8, 1.3, 0.9, 0.4, 0.2, 0.15, 0.02, 0.001). What is the smallest value of K (dimension after reduction) we can use if we want to retain 75% of the variance (sum of all the variances in value) using the first K principal components? Justify your answer.

K

Justification:

Justification

3. **[1 pts] Select one:** Assume we apply PCA to a matrix $X \in \mathbb{R}^{N \times M}$ and obtain a set of PCA features, $X \in \mathbb{R}^{N \times M}$. We divide this set into two, Z_1 and Z_2. The first set, Z_1, corresponds to the top principal components. The second set, Z_2, corresponds to the remaining principal components.

   Which is more common in the training data?

   - ◯ A point with large feature values in Z_1 and small feature values in Z_2
   - ◯ A point with small feature values in Z_1 and small feature values in Z_2
   - ◯ A point with large feature values in Z_1 and large feature values in Z_2
   - ◯ A point with small feature values in Z_1 and large feature values in Z_2

2.2 [12 pts] PCA Computation

Given 6 data points in $\mathbb{R}^5$, represented as rows in a 6 x 5 matrix X below:

$$X = \begin{bmatrix} -2 & -2 & -2 & 0 & 0 \\ -2 & -2 & -2 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 0 & 0 & 0 & -2 & -2 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

1. **[1 pts]** Is X centered?

   ◯ Yes ◯ No

2. **[3 pts]** Write X in its SVD form (a product of three matrices) where the dimensions of the three matrices are as small as possible such that X may be recovered.

   SVD decompositions are not unique. To help with grading (and with the rest of PCA below), write the left matrix with columns that have L2 norm equal to one and the right matrix with rows that have L2 norm equal to one.

   You may calculate the values however you like.

   Your Answer

3. **[2 pts]** What's first principal component of the original data set?

   Your Answer

4. **[3 pts]** If we project the original data set onto 1-D space using the first principal component, what's the variance of the projected data?

   Your Answer

   Optional space to show your work:

   Your Answer

5. **[3 pts]** For the projected data in the previous part, what is the reconstruction error? Report your answer up to three decimal places.

   Your Answer

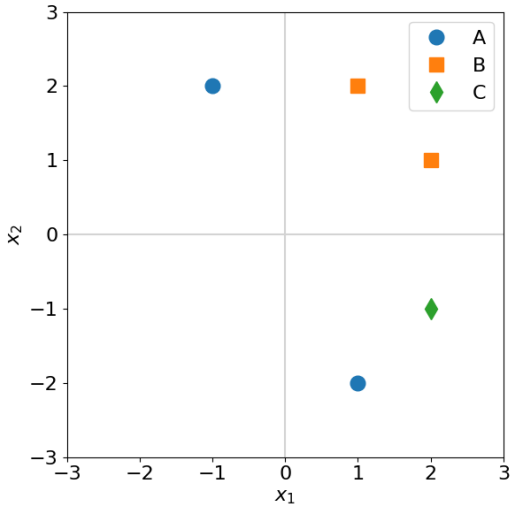
   Optional space to show your work:

   Your Answer

3 [12 pts] Generative: Categorical Gaussian

Consider the following dataset, $\mathcal{D}$, with five points and three classes, A, B, and C:

i	$x_1^{(i)}$	$x_2^{(i)}$	$y^{(i)}$
1	-1.0	2.0	A
2	1.0	2.0	B
3	2.0	1.0	B
4	1.0	-2.0	A
5	2.0	-1.0	C



We are going to use a generative model for this dataset using a categorical distribution for $p(y)$ and 2-D Gaussian distributions for $p(\mathbf{x} \mid y)$.

- We'll assume that data points are i.i.d. as usual.
- We won't make any other independence assumptions, such as naive Bayes.
- Reminder, covariance matrices are symmetric, so they actually have fewer parameters than it may appear. E.g., a 3x3 covariance matrix would have 6 parameters (3 variance values on the diagonal and the three covariance values).

Note: You may use a calculator or Python as long as you are doing the calculations yourself.

For this problem, we'll assume that our parameters are fixed at:

$$p(y \mid \boldsymbol{\pi}): \quad \pi_A = 0.4 \quad \pi_B = 0.4 \quad \pi_C = 0.2$$

$$p(\mathbf{x} \mid Y = A, \boldsymbol{\mu}_A, \Sigma_A): \quad \boldsymbol{\mu}_A = \begin{bmatrix} -1.0 \\ -1.0 \end{bmatrix} \quad \Sigma_A = \begin{bmatrix} 2.0 & -2.0 \\ -2.0 & 4.0 \end{bmatrix}$$

$$p(\mathbf{x} \mid Y = B, \boldsymbol{\mu}_B, \Sigma_B): \quad \boldsymbol{\mu}_B = \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix} \quad \Sigma_B = \begin{bmatrix} 2.0 & -0.5 \\ -0.5 & 1.0 \end{bmatrix}$$

$$p(\mathbf{x} \mid Y = C, \boldsymbol{\mu}_C, \Sigma_C): \quad \boldsymbol{\mu}_C = \begin{bmatrix} 2.0 \\ -1.0 \end{bmatrix} \quad \Sigma_C = \begin{bmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{bmatrix}$$

3.1 [2 pts] Number of Parameters

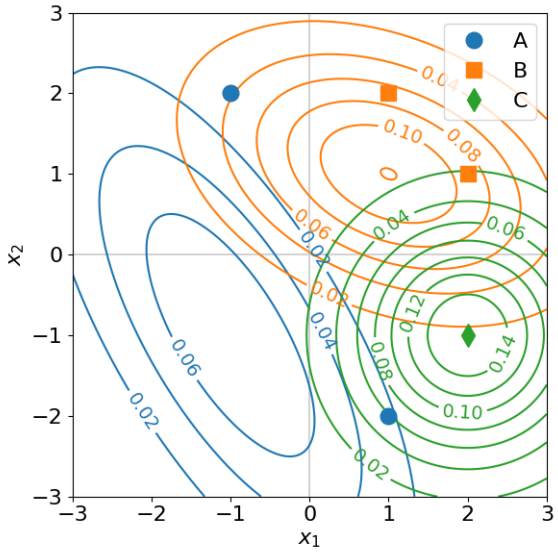
How many total parameters are in this model?

3.2 [4 pts] Distribution Plots

For the following two plots, select which distributions are being plotted for the specific parameters listed above. Blue contour lines represent $Y = A$, orange lines represent $Y = B$, and green lines represent $Y = C$. Please let us know if you need any help differentiating between these colors.

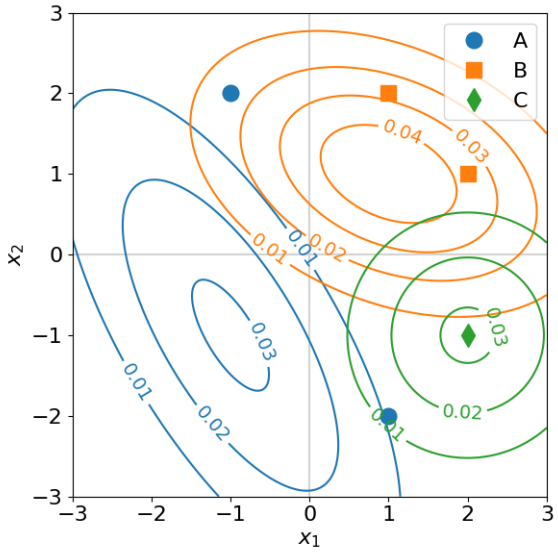
Hint: It will help to consider both plots together and/or calculate specific densities at a point or two.

1. Which distribution does the below graph plot?



- ◯ $p(\mathbf{x})$
- ◯ $p(y)$
- ◯ $p(\mathbf{x})p(y)$
- ◯ $p(\mathbf{x} \mid y)$
- ◯ $p(\mathbf{x} \mid y)p(y)$
- ◯ $p(y \mid \mathbf{x})$
- ◯ None of the above

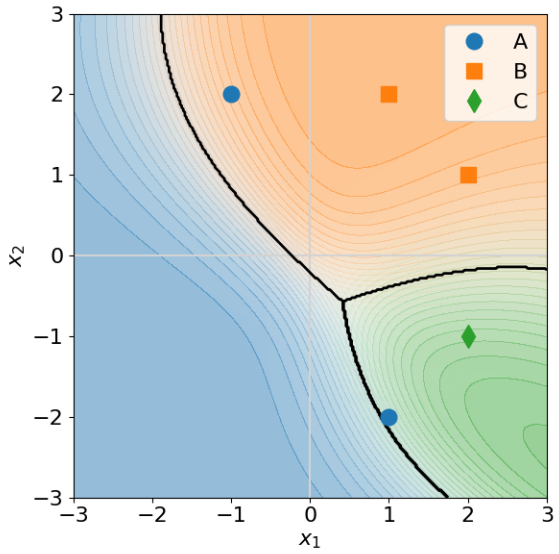
2. Which distribution does the below graph plot?



- ◯ $p(\mathbf{x})$
- ◯ $p(y)$
- ◯ $p(\mathbf{x})p(y)$
- ◯ $p(\mathbf{x} \mid y)$
- ◯ $p(\mathbf{x} \mid y)p(y)$
- ◯ $p(y \mid \mathbf{x})$
- ◯ None of the above

3.3 [6 pts] Numerical Values

Below is a plot of the data and $p(y \mid \mathbf{x})$ for our categorical Gaussian generative model with the parameters given at the beginning of this question:



Warning: The following questions in this section will be all or nothing. Make sure to check your work carefully.

Give your answers accurate to at least three significant figures.

Note: You do not need to show your work. As a sanity check, $p(\mathbf{x}^{(1)}, y^{(1)}) = 0.003355$.

1. What are the following probabilities/densities related to the fourth input, $\mathbf{x}^{(4)}$?

$p(\mathbf{x}^{(4)} \mid Y = A)$	$p(\mathbf{x}^{(4)}, Y = A)$	$p(Y = A \mid \mathbf{x}^{(4)})$

2. What is the log-likelihood (not the log *conditional* likelihood) for our generative model, $\log \prod_{i=1}^{N} p(\mathbf{x}, y)$, using the parameters given at the beginning of this question?

4 [9 pts] Decision Boundary

1. **[1 pts]** Consider a generative classifier to model a continuous random variable X with two classes $Y \in \{0, 1\}$. If we care equally about predicting both classes correctly, what does $P(Y = 1 \mid X)$ equal on the decision boundary?

2. **[6 pts]** Find the decision boundary of the following classifiers given the probabilities, assuming we care equally about predicting both classes correctly. The decision boundary can be written in the form $ax^2 + bx + c = 0$, where $a, b, c \in \mathbb{R}$ and first non-zero coefficient of a, b, c is 1 (in this order). Find the values of these coefficients and round your answers to the nearest 3 decimal places.

   Let $p(Y = 1) = \frac{1}{4}$.
   Let $p(X \mid Y = 0) = \mathcal{N}(\mu = 4, \sigma^2 = 9)$.
   Let $p(X \mid Y = 1) = \mathcal{N}(\mu = -2, \sigma^2 = 9)$.

   a b c

   Let $p(Y = 1) = \frac{2}{3}$.
   Let $p(X \mid Y = 0) = \mathcal{N}(\mu = 0, \sigma^2 = 1)$.
   Let $p(X \mid Y = 1) = \mathcal{N}(\mu = 1, \sigma^2 = 4)$.

   a b c

3. **[2 pts]** Which of the following pairs of Gaussian class conditional distributions are guaranteed to produce a linear decision boundary for Gaussian discriminative analysis? **Select all that apply.**

   ☐ $\mu_{Y=0} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \Sigma_{Y=0} = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}, \mu_{Y=1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma_{Y=1} = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$

   ☐ $\mu_{Y=0} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \Sigma_{Y=0} = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}, \mu_{Y=1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma_{Y=1} = \begin{bmatrix} b & 0 \\ 0 & b \end{bmatrix}$

   ☐ $\mu_{Y=0} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \Sigma_{Y=0} = \begin{bmatrix} c & d \\ e & f \end{bmatrix}, \mu_{Y=1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma_{Y=1} = \begin{bmatrix} c & d \\ e & f \end{bmatrix}$

   ☐ $\mu_{Y=0} = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \Sigma_{Y=0} = \begin{bmatrix} c & d \\ e & f \end{bmatrix}, \mu_{Y=1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \Sigma_{Y=1} = \begin{bmatrix} a & 0 \\ 0 & a \end{bmatrix}$

5 [4 pts] Generative vs Discriminative

Select all that apply. Naive Bayes is a generative model as opposed to discriminative because:

- ☐ It accounts for features that are never observed in the training data.
- ☐ It uses Bayes rule to build a model for $p(Y|X)$.
- ☐ It assumes that all input features are independent given the class label.
- ☐ It defines a distribution for $p(X)$.
- ☐ None of the above.

6 [20 pts] Variational Autoencoders

6.1 [10 pts] Close to standard normal

Derive the KL divergence between a univariate Gaussian distribution with mean μ and variance σ^2 and a standard normal Gaussian (mean 0 and variance 1). That is, show the following:

$$D_{KL}(\mathcal{N}(\mu, \sigma^2) \parallel \mathcal{N}(0, 1)) = \frac{1}{2}\left(\sigma^2 + \mu^2 - 1 - ln(\sigma^2)\right)$$

Hints: The following will be useful: the definition of variance; and the fact that $\mathbb{E}[X^2] = \mu^2 + \sigma^2$

Your Answer

6.2 [5 pts] KL divergence is non-negative

Prove that KL divergence is always non-negative, i.e. show that:

$$D_{KL}(p(x) \parallel q(x)) \geq 0$$

In order to prove this, use the fact that the negative of the log function is a convex function and apply Jensen's inequality for convex functions.

Your Answer

6.3 [3 pts] VAE Objective

Which of the following would we like to *minimize* when training a variational autoencoder, where $q_\phi(\mathbf{z} \mid \mathbf{x})$ is the encoder, $p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})$ is the decoder, $\mathbf{z}^{(i)} \sim q_\phi(\mathbf{z} \mid \mathbf{x}^{(i)})$, and $\hat{\mathbf{x}}^{(i)} \sim p_\theta(\mathbf{x} \mid \mathbf{z}^{(i)})$.

- ☐ $\frac{1}{N}\sum_{i=1}^{N} \|\mathbf{x}^{(i)} - \hat{\mathbf{x}}^{(i)}\|_2^2 + KL\left(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \,||\, \mathcal{N}(\mathbf{0}, \mathbf{I})\right)$
- ☐ $\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[-\log p_{\boldsymbol{\theta}}(\mathbf{x} \mid \mathbf{z})\right] + KL\left(q_{\boldsymbol{\phi}}(\mathbf{z} \mid \mathbf{x}) \,||\, p_{\boldsymbol{\theta}}(\mathbf{z})\right)$
- ☐ $-\text{ELBO}(q_\phi)$
- ☐ None of the above

6.4 [2 pts] VAE Reparameterization

True or False: The reparameterization trick is used to avoid having a random function on the computation path between the generator network weights and the objective.

◯ True ◯ False

7 Programming

The programming portion of this assignment will consist of working on 5 small problems that all pertain to the topics that you will have learned throughout the later half of the semester. The topics are as follows

1. Recommender Systems
2. K-Means Algorithm
3. Latent Variable Models

Complete the necessary programming portions of the assignment in the handout notebook, `hw10.ipynb`, in either Google Colab (preferred) or Jupyter Notebook. There will additionally be some related written questions that you will have to provide analyses or plots in the questions below.

7.1 [3 pts] Recommender Systems

Complete the recommender systems programming. Afterwards, answer the following question: after running `matrix_factorization_alt_min` with `K=2`, `alpha=0.01`, and `num_epoch=200`, which of the books (listed in the programming) will user 0 enjoy the most (report the index of the book)? Which will user 0 enjoy the least (report the index of the book)? What are the respective predicted ratings by user 0 for the two books?

Book index with **highest** rating:

Index

Rating

Book index with **lowest** rating:

Index

Rating

Visualization

7.2 [4 pts] K-Means

Complete the K-Means programming problem. Afterwards, provide the plots of the centers that are produced by the K-Means algorithm below for the various K.

K=2

K=5

K=10

7.3 Latent Variable Models

1. **[3 pts]** Train PCA, Autoencoder, and VAE using a latent dimension of 2 and the provided training code and hyperparameters. Visualize and include samples from each model after training.

PCA Samples with Latent Dimension 2

Autoencoder Samples with Latent Dimension 2

VAE Samples with Latent Dimension 2

2. **[3 pts]** Train PCA, Autoencoder, and VAE using a latent dimension of 16 with the other hyperparameters provided in the code. Visualize and include samples from each model after training.

PCA Samples with Latent Dimension 16

Autoencoder Samples with Latent Dimension 16

VAE Samples with Latent Dimension 16

3. **[2 pts]** Compare the quality of the samples generated from the 2D and 16D latent space. Which latent dimension seems to give better results? Why do you think that is?

Sample Comparison: 2D vs 16D

4. **[3 pts]** For each of PCA, Autoencoder, and VAE, include your generated images of linearly interpolated latent sample. See programming notebook for detailed instructions.

PCA Latent Interpolation Visualization

Autoencoder Latent Interpolation Visualization

VAE Latent Interpolation Visualization

5. **[2 pts]** Compare the interpolation results above for the different techniques. Why do you think different techniques worked well/poorly?

Interpolation Comparison

6. **[3 pts]** Using the latent dimension 2 models for PCA, Autoencoder, and VAE, encode all training data and plot the 2D latent encodings for each method (the code for this is provided for you in the programming). Include your three plots below.

PCA Latent Embeddings (2D)

Autoencoder Latent Embeddings (2D)

VAE Latent Embeddings (2D)

7. **[2 pts]** What do you observe about the 2D latent embeddings for each model? How do they differ, and what might explain the differences in structure?

Latent Embedding Observations

8 Collaboration Policy

After you have completed all other components of this assignment, report your answers to the following collaboration questions.

1. Did you receive any help whatsoever from anyone in solving this assignment? If so, include full details including names of people who helped you and the exact nature of help you received.

2. Did you give any help whatsoever to anyone in solving this assignment? If so, include full details including names of people you helped and the exact nature of help you offered.

3. Did you find or come across code that implements any part of this assignment? If so, include full details including the source of the code and how you used it in the assignment.