CS 559: Homework Set 3
Due: November 2, 6:00pm

Philippos Mordohai
Department of Computer Science
Stevens Institute of Technology
Philippos.Mordohai@stevens.edu

Collaboration Policy. Homeworks will be done individually: each student must hand in their own answers. Use of partial or entire solutions obtained from others or online is strictly prohibited.

Late Policy. No late submissions will be allowed without consent from the instructor. If urgent or unusual circumstances prohibit you from submitting a homework assignment in time, please e-mail me explaining the situation.

Submission Format. Electronic submission of a **zip** file on Canvas is mandatory. Include code in your pdf file as needed to make your answers clear. Submit all code separately.

Problem 1. (10 points) In Matlab, or the programming language of your choice, do the following:

- Generate 100 observations from a $3D$ normal distribution: $data = randn(3, 100);$
- Offset the data by the vector $m = [1 \quad 2 \quad 3]^T$: $data1 = data + repmat([1; 2; 3], 1, 100);$
- Scale the data anisotropically by multiplying by a diagonal matrix: $data2 = diag([10 \quad 3 \quad 1]) * data1;$
- Define a rotation matrix R:

$$R = \begin{bmatrix} 0.6651 & 0.7427 & 0.0775 \\ 0.7395 & -0.6696 & 0.0697 \\ 0.1037 & 0.0109 & -0.9946 \end{bmatrix}$$

- Rotate the data: $data3 = R * data2;$

Repeat the experiment for $10,000$ observations.

Questions (for both 100 and 10,000 observations):

1. What are the means and covariance matrices of $data$, $data1$, $data2$ and $data3$?

2. Compare the directions of maximum variance for $data2$ and $data3$? Explain your answer. (You can use the pca() function in Matlab or similar functions in other languages to answer this part.)

3. Compute the first principal component of $data3$ and compare with R. Is the answer what you expected?

4. Compare your answers to the answers you expect based on the way the data were constructed. What is the effect of using more samples in the second experiment?

Do not submit code for this problem.

Hints:

- Remember to remove the mean before computing covariance matrices. For example, use $data1mean0 = data1 - repmat(mean(data1, 2), 1, 100);$

- Make sure that the rows and columns represent the information as needed by the PCA function you use. See documentation as needed. Remember that in this case each data point consists of three measurements. Therefore, the principal components should have the correct dimensionality. (Transpose the matrices if necessary.)

Problem 2. (35 points) Download the "Pima Indians Diabetes Database" from `http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes` (if you have not already done so). Use a 50-50 split of the data for training and testing. Apply Principal Component Analysis to reduce the dimensionality of the data from 8 (do not forget to exclude the class label before doing PCA) to 3. Explain how you selected the appropriate principal components.

Train a classifier using MLE after the data have been projected.

Submit code (as a separate file), average classification accuracy over at least 10 runs and the three principal components you selected for one of the runs.

Hints:

- See hints for Problem 1.

- Do not forget that there are two classes in this problem, but the test data are unlabeled.

Problem 3. (10 points) Consider the following data drawn from two distributions in 2D.

Class 1: $D_1 = \{[-2\ \ 1], [-5\ \ -4], [-3\ \ 1], [0\ \ -3], [-8\ \ -1]\}$
Class 2: $D_2 = \{[2\ \ 5], [1\ \ 0], [5\ \ -1], [-1\ \ -3], [6\ \ 1]\}$

Classify the data using the Fisher Linear Discriminant method. Show all steps, including the computed class means, within-class scatter matrices and the optimal line direction. Also show which points are classified correctly and which points are not classified correctly. You can assign all points with positive projections to one class and all points with negative projections to the other class for this problem. Submit code as a separate file.

Hint:

- As above, make sure that the scatter matrices and projection direction are of the right dimensions. (Transpose matrices, if necessary.)

Problem 4. (30 points) Now apply the Fisher Linear Discriminant method to the Pima Indians Diabetes database. Use all 8 features, excluding the class label. **Train a classifier using MLE after the data have been projected.** Use a 50-50 split of the data for training and testing.

Submit code (as a separate .m file), average classification accuracy over at least 10 runs and the optimal projection direction for one of the runs.

Problem 5. (15 points) Consider the following 5D dataset in which the first column is the class label:

ω_2	1	1	-1	0	2
ω_1	0	0	1	2	0
ω_2	-1	-1	1	1	0
ω_1	4	0	1	2	1
ω_1	-1	1	1	1	0
ω_1	-1	-1	-1	1	0
ω_2	-1	1	1	2	1

(a) State the desired condition for correct classification for each sample using a linear discriminant function, before and after normalization.

(b) Train a perceptron using the *single sample rule* with the learning rate kept at 1 for all iterations. Use $[3\ 1\ 1\ -1\ 2\ -7]$ as the initial weight vector. Make sure that the first element of the weight vector corresponds to class label. Show all steps.