CS-ECE4720/7720 Spring 2025
Introduction to Machine Learning and Pattern Recognition

Homework Assignment 3

due Thursday 3/20/2025

Problem 1 (30 pts)
The Iris dataset was created by R. A. Fisher and has been well studied in machine learning: **iris-class.txt** and **iris-data.txt**. There are 3 classes in the Iris dataset: Iris Setosa (class 1), Iris Versicolour (class 2) and Iris Virginica (class 3), each class has 50 samples in 4-dimesional space. For each class, take the first 20 samples as the testing data and the last 30 samples as the training data.

You are to use one multivariate Gaussian density with a full covariance matrix to model the 4-dimensional data distribution of each class.

(a) (15 pts) Implement the MLE equations for the 4D Gaussian mean and covariance matrix in MATLAB, and report the mean vectors and covariance matrices for the three classes.

(b) (9 pts) Use maximum likelihood classification to classify the 60 test samples, and tally your classification errors in a confusion table as shown below:

Classified class True class	Setosa	Versicolour	Virginica
Setosa			
Versicolour			
Virginica			

(c). (6 pts) Include your MATLAB code.

Problem 2 (28 pts) Perform 3-class logistic regression on the dataset **iris-data.txt** that you used in Problem 1.

- Use extended data vector $\tilde{\underline{x}}_n = \begin{bmatrix}1\\ \underline{x}_n\end{bmatrix}$ for all data samples;
- Partition the dataset into training and testing sets in the same way as in Problem 1;

Use parameters in the file **Theta_init_S25.txt** to initialize $\Theta^{(0)} = \left[\underline{\theta}_1^{(0)}\ \underline{\theta}_2^{(0)}\ \underline{\theta}_3^{(0)}\right]$

- Set the number of iterations to 2000;
- Set the step size to $\mu = 0.0001$.

(a). (10 pts) Use the training set to estimate the parameter matrix $\Theta = [\underline{\theta}_1\ \underline{\theta}_2\ \underline{\theta}_3]$ of logistic regression, and report $\boldsymbol{\Theta}^{(2000)}$.

(b). (5 pts) Plot the L2-norm of $\boldsymbol{\Theta}^{(l)} - \boldsymbol{\Theta}^{(l-1)}$ for $l = 1,2,\cdots,2000$. Note that you can first convert the parameter matrix to vector by $\underline{\boldsymbol{\theta}} = \boldsymbol{\Theta}(:)$ (MATLAB notation) and then compute the norm.

(c). (5 pts) Compute the softmax function values on the test set samples, $s(\underline{\theta}_i{}^T \tilde{\underline{x}}_n)$, by using the parameters obtained at the last training iteration. Plot $s(\underline{\theta}_i{}^T \tilde{\underline{x}}_n)$ vs. n for the 60 test data samples with $i = 1,2,3$ in one figure, where different colors or symbols should be use for the three curves.

(d). (2 pts) Summarize your classification results in a confusion table as in Problem 1 (max posterior probability-based classification).

(e). (6 pts) Include your MATLAB code.

Problem 3 (28 pts)

(a) (4 pts) Carry out Principal component analysis (PCA) on the Iris training dataset (as defined Problem 1).

(b) (4 pts) Use the PCA result of (a) to reduce the training and test data feature dimension from 4 to 2 (note: there should be only one transform matrix W derived from the training data, and the test data should not be used in deriving W).

(c) (4 pts) Estimate the 2-D mean vectors and full covariance matrices for each class by using the estimation code you wrote for Problem 1.

(d) (4 pts) Perform maximum likelihood classification on the 2-D test data by using the models you obtained in (c).

(e) (2 pts) Summarize your results of (d) in a confusion table as in Problem 1.

(f) (4 pts) Make a scatter plot for the PCA projected Iris test data, where the data samples of different classes should be marked by different symbols or colors.

(g) (6 pts) Include your MATLAB code.