CSCI 5521: Machine Learning Fundamentals (Fall 2022)[1]

Homework 2 (Tuesday, Oct. 11)

Due on Gradescope at 11:59 PM, Friday, Oct. 21

Instructions:

- This homework has 4 questions, 100 points, on 2 pages.
- For each questions, especially programming, please provide the detailed explanation/calculation to avoid point reduction.

1. (**30 points**) Let $\mathcal{X} = \{x_1, \ldots, x_n\}$ be a set of n samples drawn i.i.d. from an uni-variate distribution with density function $p(x \mid \theta)$, where θ is an unknown parameter. In general, θ will belong to a specified subset of $\mathbb{R}$, the set of reals. For the following choices of $p(x \mid \theta)$, derive the maximum likelihood estimate of θ based on the samples $\mathcal{X}$:

   (a) $p(x \mid \theta) = \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x^2}{2\theta^2}\right), \theta > 0.$

   (b) $p(x \mid \theta) = e^{-\theta}\frac{\theta^x}{x!}, \theta > 0, x > 0.$

   Programming assignments: The next three problems involve programming. We gave you skeleton files and you don't need to modify Q2.py, Q3.py, and Q4.py.

2. (**25 points**) In this programming exercise you will implement K-means and Principal Component Analysis (PCA):

   (a) You will implement the K-means algorithm ($K = 3$) to the provided dataset Digits089.csv. The dataset contains 3000 samples, where each sample has 784 features. y_train includes three class labels (i.e., 0, 8, 9) and X_train store features. Your program should iteratively update the center of each cluster based on the input samples and record the reconstruction error after each iteration. **Write a summary of K-mean algorithm and report the number of iterations for convergence. Plot the reconstruction error with respect to the iteration. Is the plot shape following what you expect?** The code for plotting is included in Q2.py, and you do not need to modify the file.

   (b) Repeat the above, but use low-dimensional data obtained from PCA. Implement the PCA algorithm, which should reduce the original samples to dimensions needed to capture $> 90\%$ of the variance. **How many dimensions are necessary in this case? Write a summary of PCA algorithm and plot the the reconstruction error with respect to the iteration. Does PCA help to cluster? Explain.** (Hint: Consider the number of iterations for convergence and running time)

   (c) For both questions above, display the cluster centroid and the cluster assignment using only the first two principal components. Distinguish between the three classes $(0, 8, 9)$ with a different plot symbol and color.

[1]Instructor: Kshitij Tayal (tayal@umn.edu). TA: Xianyu Chen, and Yoshitaka Inoue (csci5521@umn.edu).

3. (**25 points**) In this programming exercise, you will implement two multivariate Gaussian classifiers and train it on the provided dataset training_data.txt and test it on test_data.txt with two different assumption which are as follows:

   - Assume S_1 and S_2 are learned from the data from each class.
   - Assume $S_1 = S_2$ (learned from the data from both classes).

   Each of the data files contains a matrix $\mathcal{M} \in \mathbb{R}^{N\times 9}$ with N samples, the first 8 columns include the features (i.e. $x \in \mathbb{R}^8$) used for classifying the samples while the last column stores the corresponding class labels (i.e. $r \in \{1, 2\}$). You can assume prior probability for C_1 (Class 1) and C_2 (Class 2) as 0.3 and 0.7 respectively. Here S_1 refers to covariance matrix for class 1 and S_2 refers to covariance matrix for class 2. Answer the following questions:

   (a) Write an overview of each discriminant function for each case. Briefly explain.
   (b) Plot the confusion matrix for class-dependent covariance, given by Q3.py.
   (c) Plot the confusion matrix for class-independent covariance, given by Q3.py.
   (d) Briefly explain the difference of the results.

4. (**20 points**) In this exercise, you will implement Bernoulli Naive Bayes to solve binary text classification problems. Specifically, you will use Naive Bayes to classify whether a text email is spam or not. The dataset provided has 2000 samples. We utilize 1600 samples to estimate parameters and the rest 400 samples for testing. **Write a summary of your algorithm and show the associated confusion matrix.**

Submission

- Additional Instructions: You need to submit the code written in Python. Other programming languages are not accepted and would cause credit loss. We provide the template, including the function's corresponding inputs and outputs; you need to fill in the function and generate the expected results.
- Things to submit:
  - `hw2_sol.pdf`: the document including all of your answers to the problem **including the summary or figure of the programming problems**. The PDF format is accepted. If you would like to use the hand-written document, you can scan it and make sure that the scanned copy is clearly readable. If it is difficult to read, credit would be lost.
  - `Mykmeans.py`: a text file containing the python function for Problem 2. Use the skeleton file `Mykmeans.py` and then fill in the missing parts.
  - `MyPCA.py`: a text file containing the python function for Problem 2. Use the skeleton file `MyPCA.py` and then fill in the missing parts.
  - `MyDiscriminant.py`: a text file containing the python function for Problem 3. Use the skeleton file `MyDiscriminant.py` and then fill in the missing parts.
  - `MyNaiveBayes.py`: a text file containing the python function for Problem 4. Use the skeleton file `MyNaiveBayes.py` and then fill in the missing parts.
- Submit: You need to submit the materials electronically to the Gradescope. **For this assignment, you need to submit the materials [Hw2-Written (for `hw2_sol.pdf`) and the HW2 programming (the compressed file such zip file including the Python codes)].** We will grade your code in vanilla Python.