CSCI 5521: Introduction to Machine Learning (Spring 2016)[1]

Homework 2

Questions

1. Implement a program to fit two multivariate Gaussian distributions to the 2-class data in "training_data.txt" and classify the test data in "test_data.txt" by computing the log odds $log\frac{P(C_1|x)}{P(C_2|x)}$ with $P(C_1) = 0.4$ and $P(C_2) = 0.6$. Your program should learn μ_1, μ_2, $\mathbf{S_1}$ and $\mathbf{S_2}$, the mean and covariance estimate for class 1 and class 2, in each of the following three different models:

   (a) (**8 points**) Assume independent $\mathbf{S_1}$ and $\mathbf{S_2}$ (learned from the data from each class).

   (b) (**7 points**) Assume $\mathbf{S_1} = \mathbf{S_2}$ (learned from the data from both classes).

   (c) (**10 points**) Assume $\mathbf{S_1}$ and $\mathbf{S_2}$ are diagonal: $\mathbf{S_1} = \alpha_1\mathbf{I}$, $\mathbf{S_2} = \alpha_2\mathbf{I}$ (α_1 and α_2 are learned from the data from each class).

   Return and print out the learned parameters μ_1, μ_2 to the MATLAB command window. For (a) and (b), return and print out the learned parameters $\mathbf{S_1}, \mathbf{S_2}$ and print out the error rates on the test set to the MATLAB command window. For (c), return and print out the learned parameters α_1 and α_2 and print out the error rate on the test set to the MATLAB command window. Also explain the results in the report.

2. In this problem, we will observe the impact of dimensionality reduction on classification on the Optdigits dataset.

   (a) (**15 points**) Implement k-Nearest Neighbor (KNN) on the Optdigits dataset for $k = \{1, 3, 5, 7\}$. Print out the error rate on the test set for each value of k to the MATLAB command window.

   (b) (**20 points**) Using PCA, compute a projection only using the Optdigits training data. Then project both the training and test data to $\mathbb{R}^2$ using the first two principal components. Run KNN on the projected data for

[1] Instructor: Rui Kuang (kuang@cs.umn.edu). TAs: Raphael Petegrosso (peteg001@cs.umn.edu); Zhuliu Li (lixx3617@umn.edu).

$k = \{1, 3, 5, 7\}$. Print out the error rate on the test set for each value of k to the MATLAB command window. Also, for $k = \{1, 3, 5\}$, plot all samples in the projected space and label each data point with the corresponding digit in 10 different colors for the 10 types of digits. Include in the plots the decision boundaries found by KNN.

(c) (**25 points**) Using LDA, compute a projection only using the Optdigits training data. Do this by first projecting the data to $D - 1$ dimensions using PCA then check whether the inverse of the projected data S_w exists. If not, then project the data again using PCA to $D - 2$ dimensions and check again. Do this until the inverse exists and then compute a projection using LDA. Use the computed projection to project both the training and test data to $\mathbb{R}^2$ using the first two dimensions. Run KNN on the projected data for $k = \{1, 3, 5, 7\}$. Print out the error rate on the test set for each value of k to the MATLAB command window. Also, for $k = \{1, 3, 5\}$, plot all samples in the projected space and label each data point with the corresponding digit in 10 different colors for the 10 types of digits. Include in the plot the decision boundaries found by KNN.

3. Apply PCA to visualize the modified Faces dataset from the UCI repository[2]. This dataset has been modified from the UCI version by only containing half-resolution (60 x 64) images of only straight head position. There are 4 facial expressions (neutral, happy, sad, angry) and 2 eye states (open, sunglasses) of each of the 20 people. Some images were corrupted in the original dataset so there are a total of 156 images in this dataset. The Faces dataset is given in the file “faces.txt”. Each row represents an image and each column represents a pixel. Use the following MATLAB command to visualize image i:

```
imagesc(reshape(faces_data(i,:),60,64)).
```

- (**10 points**) Implement PCA and apply it to find the principal components in the dataset. In two separate figures, plot the first two eigenvectors to visualize the eigen faces of the dataset using a similar command as above.
- (**15 points**) Use the first d principle components to reconstruct the first image (first row of the data matrix)by projecting the centered data using first d principal components then back project to the original space and add the mean, where $d = \{10, 50, 100\}$. For each d, plot the reconstructed image. Explain your observations.

[2] https://archive.ics.uci.edu/ml/datasets/CMU+Face+Images

Instructions

- Solutions to all questions must be included in a report including result explanations, learned parameter values, e.g., $\mu_1, \mu_2, \mathbf{S_1}, \mathbf{S_2}$, and all error rates and plots.

- All programming questions must be written in Matlab, no other programming languages will be accepted. The code must be able to be executed from the Matlab command window on the cselabs machines. Each function must take the inputs in the order specified and print/display the required output to the Matlab command window. For each part, you can submit additional files/functions (as needed) which will be used by the main functions specified below. Put comments in your code so that one can follow the key parts and steps. **Please follow the rules strictly. If we cannot run your code, you will receive no credit.**

- **Question 1:**

  - Problem1(*training_data.txt*: a file containing training data, *test_data.txt*: a file containing test data). The function must print the test set error rate and the learned parameters $\mu_1, \mu_2, \mathbf{S_1}, \mathbf{S_2}$ to the Matlab command window. It must also return the learned parameters in variables to the Matlab workspace.

- **Question 2:**

  - myKNN(*optdigits_train.txt*: a file containing Optdigits training data, *optdigits_test.txt*: a file containing Optdigits test data, k: the number of nearest neighbors). The function must print to the Matlab command window the test set error rates.
  - Problem2b(*optdigits_train.txt*: a file containing Optdigits training data, *optdigits_test.txt*: a file containing Optdigits test data, k: the number of nearest neighbors). The function must print to the Matlab command window the test set error rates and display all plots.
  - Problem2c(*optdigits_train.txt*: a Optdigits training data, *optdigits_test.txt*: a file containing Optdigits test data, k: the number of nearest neighbors). The function must print to the Matlab command window the test set error rates and display all plots.

- **Question 3:**

  - myPCA(*faces.txt*: a file containing the Faces dataset). The function must display all plots.

- For each dataset, rows are the samples and columns are the features with the last column containing the label (except the Faces dataset which has no labels).

- You can use the *eig* function to calculate eigenvalues and eigenvectors. To visualize the projected data, you can use the *text* function. To specify the color, use the *Color* parameter in the *text* function. If the figure doesn't show all the data, you can use the *axis* function to scale the axis.

Submission

- **Things to submit:**

  1. hw2_sol.pdf: A document which contains the report with solutions to all questions.
  2. Problem1.m: Code for Question 1.
  3. myKNN.m, Problem2b.m, Problem2c.m Code for Question 2.
  4. myPCA.m: Code for Question 3.
  5. Any other files, except the data, which are necessary for your code.

- **Submit**: All material must be submitted electronically via Moodle.