Homework 2

DSE 220: Machine Learning

Due Date: 30 April 2017

1 Instructions

The answers to the questions should be submitted on Gradescope and the code should be submitted on github by 30 April 2017. You may submit the ipython notebook downloaded as PDF to Gradescope, but please make sure that the questions are clearly segmented and labelled. The code will only be evaluated from github so make sure you have correct code on github. To secure full marks for a question both the answer and the code should be correct. Completely wrong (or missing) code with correct answer will result in zero marks. Please complete this homework individually.

2 Data

Download the MNIST train and test data from github along with their corresponding label files. The train and test data consist of 6000 and 1000 binarized MNIST images respectively.

3 Generative Learning

Please don't use the direct function from scikit-learn library for questions 1, 2, 3 and write your own implementation for them.

Question 1: Compute and report the prior probabilities π_j for all labels. (10 marks)

Question 2: For each pixel X_i and label j, compute $P_{ji} = P(X_i = 1|y = j)$ (Use the maximum likelihood estimate shown in class). Use Laplacian Smoothing for computing P_{ji}. Report the highest P_{ji} for each label j. (15 marks)

Question 3: Use naive bayes (as shown in lecture slides) to classify the test data. Report the accuracy. (5 marks)

Note: You can use the scikit-learn function from Question 4 onwards

Question 4: Compute the confusion matrix (as shown in the lectures) and report the top 3 pairs with most (absolute number) incorrect classifications. (10 marks)

Question 5: Visualizing mistakes: Print two MNIST images from the test data that your classifier misclassified. Write both the true and predicted labels for both of these misclassified digits. (10 marks)

Now, we will implement Gaussian Mixture Model and Linear Discriminant Analysis on the *breast cancer* data (sklearn.datasets.load_breast_cancer) available in *sklean.datasets*. Load the data and split it into train-validation-test (40-20-40 split). Don't shuffle the data, otherwise your results will be different.

Question 6: Implement Gaussian Mixture model on the data as shown in class. Tune the covariance_type parameter on the validation data. Use the selected value to compute the test accuracy. As always, train the model on train+validation data to compute the test accuracy. (10 mark)

Question 7: Apply Linear Discriminant Analysis model on the train+validation data and report the accuracy obtained on test data. Report the transformation matrix (w) along with the intercept. (5 mark)

4 Evaluating Classifiers

Question 8: Load the digits dataset (scikit-learn's toy dataset) and take the last 1300 samples as your test set. Train a K-Nearest Neighbor (k=5, l_{inf} distance) model and then without using any scikit-learn method, report the final values for Specificity, Sensitivity, TPR, TNR, FNR, FPR, Precision and Recall for Digit 3 (this digit is a positive, everything else is a negative). (15 marks)

5 Regression

An ablation experiment consists of removing one feature from an experiment, in order to assess the amount of additional information that feature provides above and beyond the others. For this section, we will use the diabetes dataset from scikit-learn's toy datasets. Split the data into training and testing data as a 90-10 split with random state of 10.

Question 9: Perform least squares regression on this dataset. Report the mean squared error and the mean absolute error on the test data. (5 marks)

Question 10: Repeat the experiment from Question 10 for all possible values of ablation (i.e., removing the feature 1 only, then removing the feature 2 only, and so on). Report all MSEs. (10 marks)

Question 11: Based on the MSE values obtained from Question 11, which features do you deem the most/least significant and why? (5 marks)