Homework 2

DSE 220: Machine Learning

Due Date: 30 April 2020, 11:59 PM

1 Instructions

The answers should be submitted on Gradescope. You should submit the PDF of the jupyter notebook and also submit a zip of the notebook with any additional files needed to run the notebook. Please make sure that the questions are clearly segmented and labeled. 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 Dataset for Section 3

Download the 'wine' (train, validation and test) data from Canvas (hw1.zip). Use this data for the the k-Nearest Neighbours problems.

Download from: Canvas/Files/Lecture1/hw1.zip

3 k-Nearest Neighbours

Normalize Data: Normalize features such that for each feature the mean is 0 and the standard deviation is 1 in the train+validation data. Use the normalizing factors calculated on train+validation data to modify the values in train, validation and test data.

Question 1: Train k-nn model on train + validation data and report accuracy on test data. Use Euclidean distance and k=3. (1 mark)

Question 2: Train the model on train data for distance metrics defined by l_1, l_{inf}, l_2. Report the accuracies on the validation data. Select the best metric and report the accuracy on the test data for the selected metric. Use k=3. (1 mark)

Question 3: Train the k-nn model on train data for k=1,3,5,7,9. Report and plot the accuracies on the validation data. Select the best 'k' value and report the accuracy on the test data for the selected 'k'. Use Euclidean distance. (2 marks)

4 Data

Download the MNIST train and test data from Canvas (hw2_SP20.zip) along with their corresponding label files. The train and test data consist of 6000 and 1000 binarized MNIST images respectively.
Download from: Canvas/Files/Lecture2/hw2_SP20.zip

5 Generative Learning

Please don't use the direct function from scikit-learn library for questions 4, 5, 6 and write your own implementation for them.

Question 4: Compute and report the prior probabilities π_j for all labels. (10 marks)

Question 5: 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 6: 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 functions from Question 7 onwards

Question 7: Compute the confusion matrix (as shown in the lectures) and report the top 3 pairs with most (absolute number) incorrect classifications. (10 marks)

Question 8: 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 9: 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 10: 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)

6 Evaluating Classifiers

Question 11: Load sklearn's digits dataset (sklearn.datasets.load_digits) 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)

7 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 12: Perform least squares regression on this dataset. Report the mean squared error and the mean absolute error on the test data. (5 marks)

Question 13: Repeat the experiment from Question 9 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 14: Based on the MSE values obtained from Question 10, which features do you deem the most/least significant and why? (5 marks)

8 Logistic Regression

For the following question use the wine dataset (wine original.csv). Download the file from Canvas (hw2_SP20.zip)
Download from: Canvas/Files/Lecture2/hw2_SP20.zip

Question 15: Perform a 80-20 split using *train_test_split* on the data to obtain the train and the test data. Set *random_state* = 3 while performing the train test split. Use Logistic Regression to classify the wines according to their cultivators. Tune the classifier using *Lasso* and *Ridge* regularization techniques under different values of *'C'* using *GridSearchCV.* Clearly report the parameters of the best classifier and the accuracy on the test data. (10 marks).