### University of Waterloo

**ECE 657A: Data and Knowledge Modeling and Analysis**

**Winter 2024**

**Assignment 1:** Data Cleaning and Dimensionality Reduction

**Due:** Mar 25th, 2024 11:59pm

**Overview**

Assignment Type: Done in groups of up to three students.

Hand in: One report (PDF) or python notebook per group, via the LEARN dropbox. Also submit the code / scripts needed to reproduce your work. (If you are submitting by PDF, if you don’t know LATEX should try to use it, it’s good practice and it will make the project report easier)

**Objective**: To gain experience on the use of classification.

**Datasets**

Available on LEARN

Dataset A: This data is the splice junctions on DNA sequences. The given data set includes 2200 samples with 57 features, in the matrix ’fea’. It is a binary class problem. The class labels are either +1 or -1, given in the vector ’gnd’. Parameter selection and classification tasks are conducted on this data set. (File :DataA.csv)

Dataset B: This data consists of 3 different types of irises’ (Setosa, Versicolour, and Virginica) petal and sepal length. The rows being the samples and the columns being: Sepal Length, Sepal Width, Petal Length and Petal Width. (File: DataB.csv)

Dataset C : Handwritten digits of 0, 1, 2, 3, and 4 (5 classes). This dataset contains 2066 samples with 784 features corresponding to a 28 x 28 gray-scale (0-255) image of the digit, arranged in column-wise. This data is used to illustrate the difference between feature extraction methods. (File: DataC.csv)

**Guidelines**

* No late submissions will be accepted.
* The answer sheets are checked for plagiarism.
* The code will check for plagiarism with the online websites. • For all the random state use seed=42

**Nonlinear Dimensionality Reduction**

Refer to DataC.csv

Apply the nonlinear dimensionality reduction methods Locally Linear Embedding (LLE) and ISOMAP to the dataset C, set the number of nearest neighbors to be 5, the projected low dimension to be

4.

* 1. Apply LLE to the images of digit ’3’ only. Visualize the original images by plotting the images corresponding to those instances on 2-D representations of the data based on the first and second components of LLE. Describe qualitatively what kind of variations is captured.
  2. Repeat step 1 using the ISOMAP method. Comment on the result. Does ISOMAP do better in some way? Are the patterns being found globally based or locally based?
  3. Use the Naive Bayes classifier to classify the dataset based on the projected 4-dimension representations of the LLE and ISOMAP. Train your classifier by randomly selected 70% of data, and test with remained 30%. Retrain for multiple iterations (using different random partitions of the data) and use the average accuracy of multiple runs for your analysis. Justify why your number of iterations was sufficient. Based on the average accuracies compare their performance with PCA and LDA. Discuss the result.

**Binary Classification**

Refer to DataA.csv

Classify data set A using four classifiers: k-NN, Support Vector Machine (with rbf kernel), Naïve Bayes Classifier, and Decision Tree. The objective is to experiment with parameter selection in training classifiers and to compare the performance of these well-known classification methods.

1. Preprocess the data using the Z-score normalization, and randomly split the data into a training set and a test set using the hold-out scheme: 70% of samples for training and the other 30% for testing.
2. Use 5-fold cross validation on the training set to select the parameters k for k-NN from the set [1, 3, 5, 7, ..., 31]. Plot a figure that shows the relationship between the accuracy and the parameter k. Report the best k in terms of classification accuracy.
3. For the RBF kernel SVM, there are two parameters to be decided: the soft margin penalty term c and the kernel width parameter gamma. Again use 5-fold cross validation on the training set to select the parameter c from the set [0.1, 0.5, 1, 2, 5, 10, 20, 50] and select the parameter gamma from the set [0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10]. Report the best parameters in terms of classification accuracy.
4. Using the chosen parameters from the above parameter selection process for k-NN and SVM, and the default setups for Naïve Bayes classifier and Decision Tree, classify the test set. Repeat each classification method 20 times by varying the split of training-test set as in Step (1). Report the average and standard deviation of classification performance on the test set regarding accuracy, precision, recall, and F1-score.
5. Comment on the obtained results.

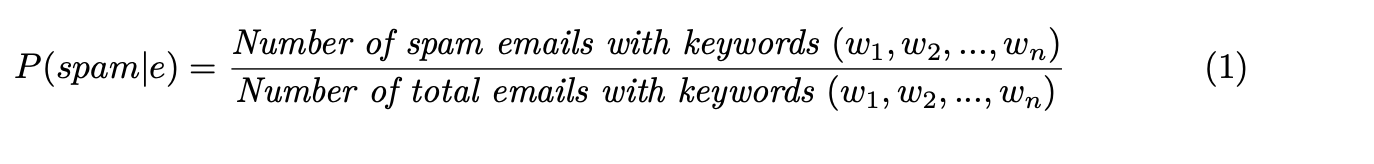
**Multi-Class Classification**

Refer to DataB.csv

1. Describe and develop the training and classification procedures by using the “one-versus-all” and “one-versus-one” strategy for SVM.
2. Classify data set B by using binary Support Vector Machine classifiers with linear kernel and default parameters. Randomly split the data into 70% training and 30% test set. Report the classification overall accuracy, precision, recall, F1-score, and the confusion matrix of the classification results on the test set.
3. How does the decision tree classifier deal with the multi-class problem? Classify data set B using decision tree with default parameters, report the classification results. Comment and compare the methods of SVM and decision tree.

**Theoretical Questions**

1. Suppose we have a data set with k features and N samples and N ≫ k then:  
   1. In the event of discrete features (2 category only), what is the highest number of leaves in decision tree?
   2. In the event of continuous features, what is the highest number of leaves in decision tree?
2. You are classifying emails as legitimate or spam and in doing so you would like to estimate the probability that a new email e containing the keywords (w1 , w2 , ..., wn ) is spam by taking all the emails in the training set with those keywords. In order words the probability is calculated as:



* 1. Explain why the plan may not work.
  2. Describe the data sets for which this plan might work.
  3. Using Naive Bayes assumption, explain how to get the probability of an email being spam? Show it mathematically.
  4. How does Naive Bayes assumption caters the problem with your plan?

1. For the following question, consider that we are using SVM with quadratic kernel and Fig.1

A graph with red and green dots

Description automatically generated

* When slack penalty (C) is large, draw the decision boundary. (Consider C → ∞)
* When slack penalty (C) is small, draw the decision boundary. (Consider C ≈ 0)
* Which one of the above will give better results in terms of classification.
* When C is large, draw a data point which will not alter the learnt decision boundary. Give your reasons.
* When C is large, draw a data point which will alter the decision boundary learnt drastically. Give your reasons.

Note: You can print the Fig. 1 on a paper and draw on it, then scan it and submit. Alter- natively, you can use software that allows you to digitally draw the boundaries. Submit plots for each part separately.