C S 488/508 Introduction to Data Mining
Homework 4: Classification

Objective

In this homework, you will do exercises to understand several different classification algorithms.

Q1. (20 points) Naive Bayesian approach.

This is a non-programming question. Consider the data set shown in Table 1.

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	+
8	1	0	1	+
9	1	1	1	+
10	1	0	1	+

Table 1: Data set for Q3

(a) (5 points) Estimate the conditional probabilities for $P(A|+)$, $P(B|+)$, $P(C|+)$, $P(A|-)$, $P(B|-)$, $P(C|-)$.

(b) (5 points) Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample (A=0, B=1, C=0) using the naive Bayes approach. Work out the solution manually and show the detailed manual calculation steps.

(c) (5 points) Estimate the conditional probabilities using the m-estimate approach, with $p = \frac{1}{2}$ and $m = 4$.

(d) (**CS 508 only**) (5 points) Repeat part (b) using the conditional probabilities given part (c).

Q2. (10 points) Linear model.

This is a non-programming question. You can consider utilizing the idea behind linear SVM or Perceptron.

(a) Demonstrate how the perceptron model can be used to represent the following boolean functions

`A OR B`

(Hint: draw a table with all the possible values of A and B and the result of A OR B. Then, derive a linear equation that can separate the two different classes.)

(b) (**CS 508 only**) Is the problem `NOT A AND C` linearly separable? Explain why.
(Hint: draw a table with all the possible values of A and C and the result of `NOT A AND C`. Then, check and see whether a linear equation that can separate the two different classes exists or not.)

Your answer to this question should not be more than half of an A4 page.

Q4. (70 points) Programming question

This question uses a data set called Glass Identification Data Set from the UCI machine learning repository. Its csv file can be downloaded from Canvas (glass.csv). For this dataset, we want to train classifiers to identify types of glass. There are 6 types of glass defined in terms of their oxide content (i.e., Na, Fe, K, etc.). Refer to the data set description at `https://archive.ics.uci.edu/ml/datasets/glass+identification` for more details about its attribute information, instances, and classes.

(a) (10 points) (Holdout) Create a training set that contains 80% of the labeled data and export it to a .csv file called `training.csv`. Create a test set that contains the remaining 20% and export it to a .csv file called `testing.csv`. Submit the two .csv files.
(Hint: One way is using `train_test_split` from `sklearn.model_selection` and `savetxt` from `numpy`)

(b) (20 points) (k-Nearest Neighbor Classifier) Use the training and testing sets in the above question, train a k-nearest neighbor classifier and measure performance on both the training set and the testing set. Vary the settings in the following ways:

- Try it for k = [1, 3, 5, 7, 9, 11]
- Try it with Euclidean distance and Manhattan distance.

Submit source code and plots showing the trend as k varies from 1 to 25 for each of the three distances, focusing on both training set and test set accuracies. What do you find?
(Hint: Use `KNeighborsClassifier` from `sklearn.neighbors`)

(c) (40 points) (Compare Logistic regression classifier and support vector machine classifier; 5-fold cross validation) In this question, you need to do 5-fold cross validation.

- First, create training datasets and testing datasets for 5-fold cross validation. These training and testing sets should be the same for training the logistic regression and the SVM models. This is to guarantee that their classification performance can be compared.
- Train logistic regression classifiers and measure performance (using accuracy) using 5-fold cross validation.
- Train non-linear SVM models using RBF kernel and measure performance (using accuracy) using 5-fold cross validation.
- Compare the performance of logistic regression and SVM models. Does one model outperform the other? What is the possible reason?

Submit source code, the performance measurement, and your analysis.
(Hints: Use class `LogisticRegression` from `sklearn.linear_model`.
Use class `SVC` from `sklearn.SVM`.
Use `KFold` from `sklearn.model_selection` to split the data into 5 folds. See `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html` for an example on how to use `KFold`)

General requirements

- For questions that are not required to be done manually, you can write code or conduct manual calculation to answer the questions.
- Put the code for all these questions to one file. Please properly organize the code to make grading easy.

Submission instructions

A zipped file `hw-lastname.zip` consisting of all the code and the PDF file.

Grading criteria

(1) CS 508 students need to answer all the questions.

(2) CS 488 students do not need to answer questions marked with **(CS 508 only)** although you have the freedom to work on them. Your scores will be scaled to 100. If CS 488 students answer the questions marked with **(CS 508 only)**, you will not have any points deducted if your answers are wrong; you will not get any extra points either if your answers are correct.

(3) The score allocation has been put beside the questions.

(4) Please make sure that you test your code **thoroughly**.

(5) FIVE points will be deducted if files are not submitted in the required format.