COMPSCI 361: Machine Learning
Assignment 3: Group Project (Worth 10% in Total)
Due date: 23:59 27 May 2022

General Instruction:

Recall that we have covered a series of supervised learning approaches, in particular, classification in our lectures, including **Logistic Regression** (week 5), **Naive Bayes** (week 5), **SVM** (week 8-9), and **Neural Networks** (week 10). The goal of this assignment is to investigate supervised learning algorithms for article classification on BBC news datasets using LR, NB, SVM, NNs. This assignment can be divided into four parts. Please write a python program to complete the following tasks using the Scikit-learn library:

- Task 1: Exploratory data analytics on text data
- Task 2: Perform classification models (LR, NB, SVM, NNs) to build article classifiers using the given dataset.
- Task 3: Investigate the impact of multiple hyperparameters. Compare the classification quality across four classification models in terms of F1 measure. Learn to manage overfitting or underfitting situations.
- Task 4: Report your answers for each question and summarize your insights

Datasets:

Let's consider two classes of BBC news articles: tech news and entertainment news. After loading the data (.csv file) using pandas library, you will see that each news article is a comma-separated line with three columns: `news ID, processed news body, news class`. The processed news bodies are tokenized and lower-cased with removal of stop words and special characters. You can find the two data files, train.csv and test.csv, in `A3.zip` on Canvas.

Submission:

Each group leader submits a single **report** ("`Your_Tutorial_Group_Name.pdf`" or in `.HTML`) and the **source code with detailed comments** ("`Your_Tutorial_Group_Name.py`" or in `.ipynb`) on Canvas by **23:59, Friday 27 May 2022**. Your report should be no more than five pages. Each group member shall submit a **peer review form** (`Your_UPI.pdf`), including contribution weights of your own and the rest of the group members. The total contribution weights of your own and the rest of the group members should be added up to 1. Individual marks will be adjusted according to final weighting. You may submit to Canvas many times.

Penalty Dates:

The assignment will not be accepted after the last penalty date unless there are special circumstances (e.g., sickness with medical certificate, family/personal emergencies). Penalties will be calculated as follows as a percentage of the mark for the assignment.

- By 23:59, Saturday 28 May 2022 (10% penalty)
- By 23:59, Sunday 29 May 2022 (30% penalty)

Task 1: Exploratory Data Analytics [1 pts]

(a) Load the dataset and construct a feature vector for each article in the. You need to report the number of articles, and the number of extracted features. Show 5 example articles with their extracted features using a dataframe. **[0.5 pts]**

(b) Conduct term frequency analysis and report three plots: (i) top-50 term frequency distribution across the entire dataset, (ii) term frequency distribution for respective class of articles, and (iii) class distribution. **[0.5 pts]**

Task 2: Classification Models Learning [4 pts]

(a) **LR.** Train your logistic regression classifier with L2-regularization. Consider different values of the regularization term λ. Describe the effect of the regularization parameter λ on the outcome in terms of bias and variance. Report the plot generated for specific λ values with training loss on the y-axis versus λ on the x-axis to support your claim. **[1 pts]**

(b) **NB**. Train a Naive Bayes classifier using all articles features. Report the (i) top-20 most identifiable words that are most likely to occur in the articles over two classes using your NB classifier, and (ii) the top-20 words that maximize the following quantity $\frac{P(X_w=1|Y=y)}{P(X_w=1|Y\neq y)}$. Which list of words describe the two classes better? Briefly explain your reasoning. **[1 pts]**

(c) **SVM**. Train your SVM classification models on the training dataset. You need to report two surface plots for: (i) the soft-margin linear SVM with your choice of misclassification penalty (C), and (ii) the hard-margin RBF kernel with your choice of kernel width (σ). Explain the impact of penalty C on the soft-margin decision boundaries, as well as the kernel hyperparameter on the hard-margin decision boundaries. **[1 pts]**

(d) **NN**. Consider the neural network with the following hyperparameters: the initial weights uniformly drawn in range [0,0.1] with learning rate 0.01.

- Train a single hidden layer neural network using the hyperparameters on the training dataset, except for the number of hidden units (x) which should vary among 5, 20, and 40. Run the optimization for 100 epochs each time. Namely, the input layer consists of n features $x = [x_1, ..., x_n]^T$, the hidden layer has x nodes $z = [z_1, ..., z_x]^T$, and the output layer is a probability distribution $y = [y_1, y_2]^T$ over two classes.
- Plot the average training cross-entropy loss as shown below on the y-axis versus the number of hidden units on the x-axis. Explain the effect of numbers of hidden units.

$$CrossEntropyLoss = -\sum_{i=1}^{2} y_i \log(\hat{y}_i)$$ **[1 pts]**

Task 3: Classification Quality Evaluation [4 pts]

(a) We explore how the size of the training data set affects the test and train accuracy. For each value of m in [0.1, 0.3, 0.5, 0.7, 0.9], train your classifier on the first m portion of the training examples (that is, use the data given by XTrain[0:mN] and yTrain[0:mN]). Please report two plots: (i) training and (ii) testing accuracy for each such value of m with the x-axis referring to m and the y-axis referring to the classification accuracy in $F1$ measure as shown below. In total, there should be four curves for training accuracy and four curves for testing accuracy. Explain the general trend of the two plots in terms of training and testing accuracy if any.

$$F1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}$$

[2 pts]

(b) Let's use 5-fold cross-validation to assess model performance. Investigate the impact of key hyperparameters of your choices for each classifier using a testing dataset. Take SVM as an example, the classification accuracy may be significantly affected by the kernels and hyperparameter combination. List hyperparameters for each classifier and demonstrate how these hyperparameters impact on the testing accuracy. **[1 pts]**

(c) Report and compare your LR, NB, SVM, and NN classifiers with the best hyperparameter settings. Summarize what you have observed in the classification accuracy in $F1$ measure on the testing dataset. **[1 pts]**

Task 4: Report Writing [1 pts]

If you use jupyter notebook for this assignment, you may consider to export your notebook as an .`HTML` file and submit `HTML` and `ipynb` to Canvas.

Collaboration Policy:

- You are encouraged to do the group projects as a tutorial group of 5~6 people. Group members are responsible for dividing up the work equally and making sure that each member contributes. Please inform me early for mediations if you encounter any troubles in the group setting.
- The purpose of student collaboration is to facilitate learning. You are encouraged to seek help from each other in understanding the material needed to solve a particular homework problem. If you encounter difficulties working in group projects, feel free to consult with any of the instructors or tutors.
- Once your group lead submits the group project report and code, feedback will be provided along with the markings.
- You are encouraged to consult tutors or instructors early on your project if you intend to conduct studies beyond the project scope.
- All consultations can take place via (i) Piazza, (ii) office hours, or (iii) email (meng.chiang@auckland.ac.nz).

Grading Rubric:

Task 1(a)	1 mark for the correct output.
Task 1(b)	1 mark for the correct output.
Task 2 (a)	1 mark for the correct implementation of LR and requested output.
Task 2 (b)	1 mark for the correct implementation of NB and requested output.
Task 2 (c)	1 mark for the correct implementation of SVM and requested output.
Task 2 (d)	1 mark for the correct implementation of NN and requested output.
Task 3 (a)	1 mark for correct requested output format. 1 mark for the observations according to two requested plots with varying sizes of training data.
Task 3 (b)	1 mark for correct requested output format and the discussion of how chosen hyperparameters impact the testing accuracy.
Task 3 (c)	1 mark for discussion of comparative testing accuracy across four classifiers.
Task 4	1 mark for clarity of the report and clarity of the comments.