Assignment 1

Statistical Methods in AI - CSE471
Deadline at 11:55pm on 25th August, 2017

- **All questions are compulsory. Follow the instructions carefully.**
- **Assignment can be implemented in Python (strongly recommended) or Matlab**
- **Ensure that submitted assignment is your original work. Please do not copy any part from any source including your friends, seniors and/or the internet. If any such attempt is caught then serious action will be taken.**
- **The submission - Roll_number.zip - should contain a directory roll_number with only 5 files : q1.py, q2.py, q3.py, q4.py, Report.pdf**
- **Output Format:** `precision, recall` **should be printed on running the q*.py files. You have to code these metrics yourself. Use of any other external libraries is strictly not allowed. Print the values of the metrics upto 2 decimal places.**
- **To test your code, split the dataset into train and development dataset where required.**
- **Report should contain details of algorithm implementation, results and observations.**
- **The data must be read from a subdirectory named** `datasets` **in the uploaded directory. Before uploading, delete this directory.**
- **Sample evaluation script:** `python q2.py breast_cancer.test`
- **Scoring will automated, will be done on the basis of precision and recall values of the classifiers. Therefore it is mandatory to follow upload format. A zero will be given otherwise.**
- **Manual evaluation will be held for other aspects of implementation.**

Problem 1

Implement the following algorithms

1. Single sample perceptron without margin
2. Single sample perceptron with margin
3. Batch perceptron with and without margin.

Report should contain accuracies, reasoning and observations.
Use the MNIST dataset for training and testing. `mnist_train.csv` and `mnist_test.csv` are the corresponding csv files having the following format:

```
label, pixel-11, pixel-12,..., pixel-2828
```

Dataset:https://researchweb.iiit.ac.in/ pinkesh.badjatiya/smai_assignment1/q1_mnist_train.csv

Problem 2

Consider a dataset with a large number of sample data points. After about 100 iterations, say an online perceptron has learned a formidable classifier which is so good that it goes over all samples with no updates. On reaching the very last sample, it faults and updates. Now, there is a

chance that this ruins the hyperplane's weight vector that does really well for the rest of the data. Design an algorithm that ensures that a weight vector that does well for most of the data plays an important role in the final classifier. Use the `breast_cancer.train` and `breast_cancer.test` csv files. Since the data need not be linearly separable, run the perceptron algorithm for some arbitrary number of epochs (trial and error). One Epoch is a complete single pass of the training set from the algorithm.

Implement a single sample perceptron with

1. Relaxation algorithm + margin (Refer to pdf attached or Pattern Classification by R. Duda)
2. The modified perceptron algorithm

Mention in the report the accuracies for different epochs. Also explain how algorithm was implemented with observations. The data has the following format:

`id-number, attr-1, attr-2, ..., attr-9, class-label`

Class Labels : 2 for benign, 4 for malignant

Dataset:https://researchweb.iiit.ac.in/ pinkesh.badjatiya/smai_assignment1/q2_breast_cancer.train

Problem 3

Tanmay is the HR of a company and suddenly he observes an increase in the number of employees leaving the company. Being afraid, he starts by awarding senior employees with huge incentives & stay bonuses. He even talks to senior employees to get an inside view of the situation following the increase in the numbers. Seeing all efforts go in vain, he hires you, a senior data analyst, to help him in understanding the reason for the employees leaving prematurely.

You being an expert in Machine Learning algorithms take the offer for a huge amount of money (Yes! Data analysts get paid a lot). On your request for statistics for analysis, the HR hands over to you some data of the employees taking during regular employee feedbacks, hoping it might be useful someday.

Design a Decision Tree classifier to predict which valuable employees will leave next. You are tasked with helping in reducing the number of senior employees leaving the company by predicting the next bunch. The fate of the company rests in your hands.

For each employee you are provided the following attributes:

1. Satisfaction Level
2. Last evaluation
3. Number of projects
4. Average monthly hours
5. Time spent at the company
6. Whether they have had a work accident
7. Whether they have had a promotion in the last 5 years
8. Departments
9. Salary
10. Whether the employee has left the company

The data has the following format:

`attr-1, attr-2, ..., attr-9`

Class Labels(column name *left*) : 1 for employee left the company, 0 for not.

Dataset:https://researchweb.iiit.ac.in/ pinkesh.badjatiya/smai_assignment1/decision_tree_train.csv

Problem 4

Privacy or No Privacy? Authorship Identification

Individuals data can be used for potentially harmful purposes by using machine learning- Privacy activists warn us. How true is this? We investigate the same in this question, attempting to prove that given just a few lines of someones writing, we can figure out the author by using past writings of that person- commonly termed as Author Identification. Surprisingly, as we would find in this question, it can be achieved pretty accurately using very simplistic data representations and classifiers. This has a lot of usecases- it has the potential to deanonymize a significant amount of aggregated information. We explore this on a toy dataset as a question.

The question details is as follows:
You are given a few excerpts of past documents written by 10 different organizations/people including CIA communications, novelists and others. Using these pieces of documents, you have to predict using a K-NN classifier which author/organization does this new document fragment belong to?

The data format is as follows:
You are given 10 folders, each folder is a class. It contains text which has been preprocessed already. You need to implement the following which is left blank in the starter code:

1. A simple BoW representation.
   Details are as follows:
   (a) Normalization or weighting techniques like tf-idf are NOT required. Simply count the words, giving a vector representation.
   (b) No further NLP based preprocessing is required.
   (c) Implement some form of feature engineering (selection/addition/deletion) and specify the effects of the same in the report.
2. Implement a KNN based classifier to take these features as input, and predict the author (class) as the output.
   (a) Run for a minimum of 5 different k and report variance with k in the report.
   (b) Try to achieve a 95% accuracy on the test set.
   (c) Specify the F1-score, Accuracy and Confusion Matrix in report for every experiment.

To aid the assignment, we have provided you with the starter code for this question, since it is the first assignment. Do read it carefully. It should give you a fair idea so as to how datasets are generally organized, code written and accuracies reported.

Dataset:https://researchweb.iiit.ac.in/ pinkesh.badjatiya/smai_assignment1/knn_question/knn_train/

Starter Code:https://researchweb.iiit.ac.in/ pinkesh.badjatiya/smai_assignment1/knn_question/starter_code.py