Homework#1

CS576 Machine Learning, Fall 2023

Due: September 20

Instruction:

- Please submit one file (YourLastName_YourFirstName_CS576_HW1.zip)) for the homework.
- You may organize your submission file with the directory of each part, Part I, Part II, Part III, etc.
- Clear number your answer, e.g., Q1 (1), Q1 (2), …, Q3, ….
- Submit your homework to the assignment link in the course website.

Part I. ML - The learning problem

1. Machine learning is often referred to as an ill-posed problem. **(1)** What does **ill-posed problem** mean? **(2)** How do machine learning algorithms deal with the fact that machine learning is an ill-posed problem?

2. In machine learning, **(1)** what is meant by the term **inductive bias**? **(2)** What can go wrong when an inappropriate inductive bias is used?

3. (1) What is meant by the term **consistent model**? **(2)** Why might a consistent model not **generalize** well?

4. Consider the following set of training examples:

ID	High Blood Pressure	Smoker	Diabetes	Heart Disease	Stroke Risk
1	true	false	true	true	high
2	true	true	true	true	high
3	true	false	false	true	medium
4	false	false	false	false	low
5	true	true	true	false	high

This dataset shows a set of stroke risk factors and their probability of suffering a stroke in the next five years (*low*, *medium*, and *high*). All the descriptive features are Boolean, taking two levels: *true* or *false*.

How many possible models exist for the scenario described by the features in this dataset? Explain your answer.

5. Briefly explain **mathematical optimization** for machine learning.

Part II. Data Exploration

In Part II, you will have hands-on data exploration. You can use any machine learning API or tool, or any program language. Answer to each of the following questions and also submit your program codes used for Part II.

0. First, download ***boston.csv*** provided. This is a Boston house-price dataset. The data is from StatLab – Carnegie Mellon University. For the data description, refer to http://lib.stat.cmu.edu/datasets/boston. In the data, MEDV (Median value of owner-occupied homes in $1000's) is a target feature for prediction.

1. Check there are any missing values in the dataset. If any, report it.

2. Report the summary statistics for the TAX feature (full-value property-tax rate per $10,000) with **(1)** minimum, maximum, and range, **(2)** mean and media, **(3)** variance and standard deviation, **(4)** 1st quartile and 3rd quartile, **(5)** inter-quartile range, **(6)**12th percentile.

3. Show (1) a **histogram** for each numerical feature including the target feature, MEDV. **(2)** Is there any features which show "bimodal" distributions?

4. Show (1) a **scatter plot matrix** of numeric features from the dataset to check for correlation between features. (2) Which feature pairs show positive correlation? (3) Which feature pairs show negative correlation?

5. Conduction a **heatmap** which shows correlation values for every feature paris.

6. Conduct **standardization** on all numeric features except the target feature, MEDV.

Part III. Dimensionality Reduction

In Part III, you will have hands-on dimensionality reduction using PCA.

0. First, download a Breast Cancer dataset (***wdbc.data***) and its description ***wdbc.names*** provided. The dataset is from UCI Machine Learning Repository. For the data description, see *wdbc.names*. The Attribute #2 is the class feature with M = malignant and B = benign.

1. One of the most common applications of PCA (Principal Component Analysis) is visualizing high-dimensional datset. The Breast Cancer dataset has 30 descriptive features. It is impossible to visualize the data set with all the 30 features.

Use PCA to reduce the dataset's dimensionality from 30 to 2, and then generate the two-dimensional scatter plot of the breast cancer dataset using the first two principal components, where each data point is shaped and colored according to its class label, i.e., blue dot for M and red triangle for B.

Submission: **(1)** Your program code. You can use any machine learning API or tool for PCA and the scatter plot generation
(2) The dataset reduced after PCA (named with **wdbc_2D.data**). The dataset is supposed to have three attributes, Class, PC1, and PC2.
(3) The generated scatter plot.

Part IV. Decision Tree Learning

In Part IV, you will play with a program which implements decision tree learning.

0. First, download two programs *dtree.py* and *party.py* and a data set *party.data* provided. The *dtree.py* program implements the ID3 algorithm from scratch. The *party.py* program is a drive program to run the decision tree induction algorithm. You may need to modify these programs for properly running.

1. Build a decision tree model from *party.data,* using the programs provided.
Report **(1)** the **true class label** and **predicted class label** of each (training) instance, and **(2)** the **training error rate**

2. The program uses a given dataset for both training and validation purpose. (1) Revise the program code to prepare a training dataset and a validation dataset from a user input dataset using ***bootstrap***.
(2) Revise the program code for computing the **test error rate**.

Submit the revised program code with your explanation of the changed/added program parts.

3. For impurity measures, three metrics are popular used: *entropy*, *gini index* and *misclassification error*
The *gini index* of a node *t* is computed as $1 - \sum_{i=1}^{c} p_i(t)^2$, where *c* is the number of distinct class labels and $p_i(t)$. *Information gain* can be also computed with the gini index.

Revise the program so that the program can chooses a test attribute based on **the *information gain* with *gini index*** instead of entropy-based information gain.

Submit the revised program code with your explanation of the changed/added program parts.

4. Rebuild a decision tree model from *party.data,* using the revised program.
Report **(1)** the **training error rate** and **(2)** the **test error rate**.