Project 3: Classification Algorithms Demo

Time: Dec. 3, 2020 10:30AM
Location: Online

Please bring the following items to Demo:

- Your laptop, on which you could run and show your implementation
- Your UB card

The NEW data sets (*project3_dataset3_train*, *project3_dataset3_test*, and *project3_dataset4*) have been posted on Piazza. The data format that is described in README file, is the same as the ones we already provided.

Complete the following tasks, and show your results during demo (please do NOT include these results in your report).

1. **Nearest Neighbor** implementation:
   - ***a.*** training data: *project3_dataset3_train* file
   - ***b.*** testing data: *project3_dataset3_test* file
   - ***c.*** parameter setting: Please use **Euclidean distance** as the distance measure for data samples and **use the raw data (no preprocessing or normalization)**. You need to handle any given K and we will ask you to set K to a fixed number (for example, 9).
   - ***d.*** performance measure: report **Accuracy, Precision, Recall**, and **F-1 measure**.
2. Run your **Decision Tree** implementation with *project_dataset4* file (training dataset), and show the learnt decision tree. You will be asked to show the impurity measure in your implementation and draw your decision tree.
3. Run your **Naïve Bayes** implementation with *project_dataset4* file (training dataset), and output the following probability:

   $$p(X|H_0)p(H_0), \qquad p(X|H_1)p(H_1).$$

   For example, $X = \{sunny, cool, high, weak\}$.

   Note that there is no "zero-probability" issue for the dataset in this demo. Please **disable** you implemented components for the "zero-probability" issue.

4. Run your **Random Forests** with *project3_dataset1.txt* using 10-fold cross validation and show your performance metrics: **Accuracy, Precision, Recall**, and **F-1 measure**. The number of trees and the number of features for splitting are given.
5. Run your Kaggle competition code and explain your algorithm.

Please note:

1. The demo is divided into three parts:
   a. Check Nearest Neighbor, Decision Tree, and Naïve Bayes;
   b. Check Random Forests, and Q&A about Decision Tree and Random Forests;
   c. Check Kaggle competition, and Q&A about Nearest Neighbor, Naïve Bayes and Kaggle competition.

2. **For the first three tasks, you do not need to use cross validation.**
3. **Please remember to select your submissions for private leaderboard evaluation on Kaggle before the deadline.** If you do not do it, the system will automatically use the three submissions with best performance on the public leaderboard for private leaderboard evaluation.
4. **Please make sure that your team name displayed on the leaderboard is the same as your registered name.** The team name is the identifier for us to map your performance. If we can not map the team name on the leaderboard and your registration, you may get a **ZERO** in this part.
5. We will ask questions to test your understanding about the tested methods. Your final score is based on your report, your demo, and your answers to those questions.
6. Each group only has 15 minutes to finish the demo. Do not be late and make sure your code can run smoothly. We will NOT give extra time for debugging during the demo.
7. You should not reveal the questions you are asked during your demo to other groups. If we find this, we will give **ZERO** score for this project to your group and the groups that receive the information.