CMPT 459 Data Mining
Spring 2019
Martin Ester
TA: Ruijia Mao and Ruchita Rozario

Course project: Milestone 2

In this milestone, you will train on the data that has been preprocessed in milestone 1. You can choose among the following three classifiers:

1. Decision Tree
2. Logistic Regression
3. SVM

Each group member has to choose a different classifier and train it on the dataset. If a group only has two team members, two out of the three classifiers shall be trained. **Cross-validation (5-fold or 10-fold) should be performed on the training dataset (train.json).** The evaluation metric should be multi-class logarithmic loss as mentioned in the evaluation section in Kaggle. Once the predictions for the test dataset (test.json) are generated, please upload them to Kaggle to obtain the accuracy for the test dataset. The detailed steps are in the evaluation section in Kaggle.

In your submission, please answer the following questions:

1. Which features did you select for your classifiers? Please comment on the reason for your feature selection. If you choose to work on the bonus question, you can add your features extracted from external datasets at this step. (5 points)

2. What Python or R libraries did you use for your classifiers? (5 points)

3. How did you perform cross-validation? Please describe the procedure. (10 points)

4. What performance did the first version of your classifiers achieve on the **validation dataset (in cross-validation) and on the test dataset**? Please comment on the performance of the classifier. (15 points: 5 points for performance, and 10 points for comments).

5. What actions did you take in order to improve your classifiers? You can modify your dataset or the parameters of your classifier. **Please record your modifications in your report.** (30 points: 10 points for each improvement)

6. How did you check whether any overfitting occurred during your training?Did you observe overfitting? What did you do to avoid overfitting? (10 points)

7. What performance did you achieve on the **validation dataset (in cross-validation) and on the test dataset** after your modifications? Please, try to explain the gains. (15 points: 5 points for performance, and 10 points for explanation)

8. Evaluate one additional evaluation metrics mentioned in class on the **validation dataset**. Which metric did you use? What were the results? How do these results compare to the results for multi-class logarithmic loss? (10 points)

9. **Bonus (10 points)**: You can combine your data with other, related datasets to create additional relevant features, for example, based on the nearby subway stations and malls. Which additional features did you create? By how much did these features improve the performance? If you do not train two different versions of your classifier (with and without the additional features), what evidence do you have that the additional features helped?

Total points: 110/100

Please submit the following on Coursys:

1. **Your code or a link to your code repo. If a link to your code repo is submitted, please make sure to grant proper access to TAs.**
2. **A pdf-format report named Milestone2.<GroupName>.pdf**

IMPORTANT: Your report needs to clearly specify which student developed which classifier.

Deadline: March 5, 2020