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 apply it to the dataset. If a group only has two team members, two out of the three classifiers shall be applied. The classifiers shall be trained on the training dataset and be evaluated on the test dataset. The evaluation metric should be multi-class logarithmic loss as mentioned 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. (10 points)

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

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

4. 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)

5. What did you do to avoid overfitting? How did you make sure that your classifiers do not overfit? Do you inspect any overfitting during your training or testing? How do you deal with overfitting? (10 points)

6. What performance did you achieve after your modifications? Please, try to explain the gains. (15 points: 5 points for performance, and 10 points for explanation)

7. Evaluate one additional evaluation metrics mentioned in class. Which metric did you use? What were the results? How do these results compare to the results for multi-class logarithmic loss? (15 points)

8. **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?

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