PROJECT PLAN

1. Student names. (The project is to be done in groups of 3 students.)

   Xinru Wang, Mian Yang, Shupei Wang

2. [Up to 3 lines] Definition of the problem, possibly relevant to your interests.

   Using machine learning algorithm to fit review data from Amazon and use LIME (LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS), a model interpretation package to evaluate the model interpretability.

3. [Up to 3 lines] Description of the dataset (or datasets) to be used. Datasets should be already publicly available, since there is not enough time for you to collect data. For possible datasets, see the course webpage.

   The Multi-Domain Sentiment Dataset contains product reviews taken from Amazon.com from many product types (domains). Some domains (books and dvds) have hundreds of thousands of reviews. Others (musical instruments) have only a few hundred. Reviews contain star ratings (1 to 5 stars) that can be converted into binary labels if needed.

4. URL where the above dataset(s) is(are) available.

   http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

5. [Up to 5 lines] Which 3 machine learning algorithms are going to be used? (You should list 3 algorithms, e.g., SVM, Prank, Adaboost, etc.) **You are allowed to either implement this from scratch or use third-party code, e.g., liblinear for SVM.**

   SVM
   Neural network
   Logistic Regression
   LIME.

6. [Up to 5 lines] Cross-validation technique (e.g., training/validation/testing, k-fold cross-validation, bootstrapping). **You MUST implement this from scratch.**

   5-fold cross validation
   Bootstrapping (number of bootstraps = 100).

7. [Up to 10 lines] Which hyperparameter(s) is(are) going to be tuned. **You MUST implement this from scratch.**

   For SVM: e.g. kernel function, threshold
   For neural network: e.g. number of layers, nodes for layers
   For logistic regression: regularization, learning rates

8. [Up to 15 lines] Description of the experimental results, e.g., plots of number of samples versus accuracy (you can use different subsets of the same dataset), regularization parameter versus accuracy, ROC curves, plots of different datasets, etc. **You MUST implement this from scratch.**

   Plots of accuracy vs. hyperparameters.
   Train and test error for different number of samples.
   Permutation feature importance.

9. Which programming language are you going to use? (Only MATLAB, C++, Java and Python are allowed.)

   Python.

Advice: Do not spend too much time on things such as "understanding the data", "memory problems because your data is too big", etc. Only if you are already familiar with computer vision, brain data, natural language processing, big data, parallelism, etc. then you can make use of those things, but this will not imply that you will get a higher grade just based on that fact. In general, I would recommend to use easy-to-understand datasets, and smaller subsets of the data, for instance.