cs5600_6600_project2_proposal

Erick Stone

October 2019

Problem 1:

1. I will determine whether someone should get a mammogram or not.

2. I will use the mammogram data from https://archive.ics.uci.edu/ml/datasets/Mammographic+Mass . The data is labeled. The data does have some missing values which will need to be handled which will simply be excluded.

3. I will use several techniques covered in class and ensemble the results to predict if someone should get a mammogram.

4. This will be a python project with a run.py file that simply gets run to train.

5. Bench marks:

- 1. clean up data: remove any rows with missing data or try to fill them in based on other information available.
- 2. build linear classifier: train on data until optimal performance is reached
- 3. build svm: train on data until optimal performance is reached
- 4. build tree: train on data until optimal performance is reached
- 5. build random forest: train on data until optimal performance is reached
- 6. ensemble results

6. The biggest risk of this project is I have such little time between work, this class, and my statistics machine learning class. I also have family obligations. If anything prevents me from completion, it will likely be insufficient time.