Fall 2016 OMS DAVA HW 3 - Classification and Logistic Regression
Submission Deadline: 2016-10-31 (see T-Square for specific time)

- This HW counts as two homework assignments since it is longer than usual and will be the last hw assignment in this class.
- More information about the dataset mentioned below can be found in the documentation of the mlbench package (https://cran.r-project.org/web/packages/mlbench/mlbench.pdf).
- Please include all code (documented) and motivate your answers. If you use markdown or notebook format as your PDF submission document also include a .R file with the code.

Assignments:

1. Write down detailed formulas for the gradient of the loss function in the case of logistic regression, and write detailed pseudo code for training a LR model based on gradient descent. Count how many operations are done per each gradient descent iteration and explain how you computed your answer (use the following variables in your answer: n for the number of examples and d for the dimensionality).
2. Implement in R logistic regression based on gradient descent. To avoid unnecessary slow-down use vectorized code when computing the gradient (avoid loops).
3. Train and evaluate your code on the BreastCancer data from the mlbench R package. Specifically, randomly divide the dataset into 70% for training and 30% for testing and train on the training set and report your accuracy (fraction of times the model made a mistake) on the train set and on the test set. Repeat the random partition of 70% and 30% 10 times and average the test accuracy results over the 10 repetitions. Try several different selections of starting positions - did this change the parameter value that the model learned? Try to play with different convergence criteria to get better accuracy.
4. Repeat (3) but this time using logistic regression training code from an R package such as glm2. How did the accuracy in (4) compare to the accuracy in (3).
5. Repeat (4), but replace the 70%-30% train-test split with each of the following splits: 5%-95%, 10%-90%, …, 95%-5%. Graph the accuracy over the training set and over the testing set as a function of the size of the train set. Remember to average the accuracy over 10 random divisions of the data into train and test sets of the above sizes so the graphs will be less noisy.
6. Repeat (5) but instead of graphing the train and test accuracy, graph the logistic regression loss function (negative log likelihood) over the train set and over the test set as a function of the train set size.