CS722/822 MACHINE LEARNING

Homework #3

Due: Nov. 7th, end of the day

1. Please **derive** the gradient of the log likelihood function in logistic regression model with respect to **w**, which is a vector and the model parameter. The log likelihood function is as follows:

$$LL(\mathbf{w}) = \sum_{i=1}^{N} y_i \mathbf{w}^T \mathbf{x}_i - \log(1 + \exp(\mathbf{w}^T \mathbf{x}_i))$$

where $(\mathbf{x}_i,\ y_i)$ represents the i-th example, $\mathbf{x}_i$ being data vector for input variables and y_i being the label. N is the total number of examples in the data.

2. **Program** your own logistic regression classifier (Python is preferred) by implementing a gradient decent algorithm to find the optimal **w** that maximizes $LL(\mathbf{w})$ in Problem 1. That is to find solution to:

$$\max_{\mathbf{w}} \left(\sum_{i=1}^{n} y_i \mathbf{w}^T \mathbf{x}_i - \log(1 + \exp(\mathbf{w}^T \mathbf{x}_i)) \right)$$

3. **Program** (with python preferred) a function that plots an ROC curve with input of a vector containing the true label and another vector containing the predicted probabilities of class membership for a set of examples.

4. **Apply** your logistic regression classifier to the breast cancer Wisconsin dataset, which can either be loaded with python by following instructions from

   http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html

   or downloaded from

   https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic) .

   Randomly splitting the data into two subsets with one having 2/3 of the examples and the other one having the rest 1/3. Use the 2/3 subset to train a logistic regression model and the 1/3 subset to test the model. Plot the ROC curve on the testing set with your ROC plotting function.