Project 2

due: 10/27/2015

Naive Bayes Classifier (50 points)
For this problem, you will implement a Naive Bayes classifier for text categorizations. Given a document $\mathbf{x} = (x_1, x_2, \ldots, x_m)$, where x_i is the number of occurrences of word w_i in the document. We compute $p(\mathbf{x}|\mathcal{C}_k)$ as

$$p(\mathbf{x}|\mathcal{C}_k) \propto \prod_{i=1}^{m} [p(w_i|\mathcal{C}_k)]^{x_i},$$

where $p(w_i|\mathcal{C}_k)$ is the probability of observing word w_i in documents from class $\mathcal{C}_k$. Given a collection of training documents $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^{n_k}$ in category $\mathcal{C}_k$, we can compute $p(w_i|\mathcal{C}_k)$ as

$$p(w_i|\mathcal{C}_k) = (1-\delta)\frac{\sum_{j=1}^{n_k} x_i^j}{\sum_{j=1}^{n_k}\sum_{l=1}^{m} x_l^j} + \frac{\delta}{m},$$

where $\delta = 0.1$ is a smoothing parameter to avoid zero probabilities and m is the size of the vocabulary.

Download the text data set, `20newsgroups.zip`. You will find six files in this data set: `train.data`, `train.label`, `train.map`, `test.data`, `test.label`, and `test.map`, where the first three files are used for training and the last three are for testing. The `train.data` file contains word histograms of all documents; each row is a tuple of the format (`document-id, word-id, word-occurrence`). The class assignment information of training documents can be found in `train.label`, and the topic of each class can be found in `train.map`. Similarly, the word histograms and the class assignments of the test documents can be found in `test.data` and `test.label`, respectively. For this problem, you need to train a Naive Bayes classifier using the training data and apply the learned classifier to predict the class labels for the test documents. Submit the following in a single PDF document:

(a) Commented Matlab code for your Naive Bayes classifier. (10 points)

(b) Your classification accuracy of the test documents. (10 points)

(c) A short description of the results. Some questions you might want to address include: What classes are most likely to be misclassified? If you look at the topics, do the classification probabilities make sense? etc. (30 points)

Your score on each part will be based on how well you present the results and how insightful your analysis is.