Short Problem Set (Module 7)

1. [40%] Briefly explain the following concepts in your own words and be sure to indicate how they are relevant to text classification:

(a) bias-variance tradeoff
(b) k-fold cross-validation
(c) macro-averaging vs. micro-averaging
(d) soft margin

2. [10%] Why is 3-Nearest-Neighbor (3-NN) almost always a better choice than 1-NN for a binary (i.e, two-class) text classification problem.

3. [50%] Naïve Bayes using the Binomial (also known as Bernoulli) model. First calculate estimates of P(c) and P(w|c) given the following training sentences. There are only three classes: Travel, Business, and Health. You should not use any smoothing. For features you should only use the following seven vocabulary terms: {denver, employers, florida, hospital, jobs, nurses, vacation} and you should ignore all other words and any punctuation. Next compute the probability of each class for the two test documents A & B below. Finally, indicate which is the predicted (*i.e.,* the most likely) class for each test document. Please read these directions thoroughly, count carefully, and do show all of your work. Report probabilities using scientific notation (*e.g.*, 1.563 x 10-5) with three digits after the decimal point.

Training data

1 Travel: denver hospital administrator takes vacation in florida
2 Travel: nurses plan a trip to florida
3 Travel: employers offering more jobs with vacation benefits
4 Business: employers see growth in information science
5 Business: hospital nurses in denver say high paying jobs are vanishing
6 Business: employers say florida is nice vacation spot and there are good jobs
7 Health: study: more hospital nurses need to take a vacation
8 Health: local doctors attend florida conference on diabetes
9 Health: hospital trains maternity ward nurses
10 Health: denver hospital says local employers have jobs for nurses

Test documents:

A: florida nurses take skiing vacation in denver
B: jobs available for experienced nurses at florida hospital