EE 660 **Homework 4** Posted: 9/11/2018
Due: Tue., 9/18/2018, 2:00 PM

1. (A) Murphy Exercise 8.1, as expanded and explained below. Note that in this dataset, $y_i = 1$ denotes spam.

   (i) This problem may be solved with PMTK in MATLAB, or with Python. All the functions and tips given below are for PMTK/MATLAB users, unless stated otherwise.

   (ii) Murphy's statistics given for features 56 and 57 don't make sense; you can ignore them because you won't be using his statistics to solve this exercise. Instead, it's best to compute your own statistics (where needed) from the current training data.

   (iii) For the classification error measure, use percent misclassified points, with $\underline{w}^T \underline{x}_i \geq 0 \Rightarrow \hat{y}_i = +1$ and $\underline{w}^T \underline{x}_i < 0 \Rightarrow \hat{y}_i = -1$ (or 0).

   (iv) After adding `pmtkData` into your Matlab path, you can call "`load spamData`" to load the training and testing data. Python users can download the csv files from the D2L homework dropbox.

   (v) For preprocessing method (a) note that you should compute the mean and variance for the training set and then use them to standardize the validation/test set.

   (vi) **Code up the cross-validation loop yourself.** MATLAB users may find PMTK functions `standardizeCols`, `logregFit` and `logregPredict` useful in the loop. Python users may find `numpy.random.shuffle` useful for shuffling the data indices before cross validation.

   (vii) For model selection (choosing your value of λ), use 5-fold cross validation, and run the cross-validation 5 times, taking average errors over the multiple runs for each value of λ. Note that at each run you should partition the given training set randomly.

   (viii) **Report your selected value of λ and explain why you chose that particular λ.** This value of λ defines your "selected model".

   (ix) **Report on the following classification errors for your selected value of λ.** Please use a table like the example in Murphy except with 4 columns of error numbers as follows:

   Column 1: average cross-validation error on cross-validation training sets in the cross-validation loop

   Column 2: average cross-validation error on the validation sets in the cross-validation loop

   Column 3: error on the full given training set (trained on the full given training set)

Column 4: error on the full given test set (trained on the full given training set)

(B) Additional Question:

This problem pertains only to the given test data, as provided with the dataset. After using preprocessing method (c) on the given test data, use **sum of features 1-48 (total count of keywords in percentage)** as x axis, and **sum of features 49-54 (total count of special characters in percentage)** as y axis, and draw the following plots:

(i) A scatter plot of all testing points, using different colors for spam and non-spam emails.

(ii) For emails labeled spam, generate a 3D histogram using function `hist3()`.

(iii) For emails labeled non-spam, generate a 3D histogram using function `hist3()`.

(iv) Do you notice any significant difference between the two histograms generated in (2) and (3)? If so, briefly describe.

Complexity of learning; generalization

Reading

AML 2.1, 2.2 (pp. 39-62). The "safe skip" part is optional reading, and the "Sketch of the proof" after Theorem 2.5 is optional reading. Note that we won't get to Sec. 2.2 until Lecture 10 (on Thur., 9/20), so if you're short on time you can take a couple extra days to read Sec. 2.2.

Problems

2. [Based on Exercise 2.1]:
   (i) Find the smallest break point k for the hypothesis set consisting of Positive Rays (defined in Example 2.2).
   (ii) Find the smallest break point k for the hypothesis set consisting of Positive Intervals (defined in Example 2.2).
3. Exercise 2.3 (p. 50).