Miniproject 1
BMI 555 IEE 520
Fall 2018
Due September 6, 2018

Please submit through the digital drop box. After you upload you must also submit.

1. Select a dataset of you **own choice** of sufficient size, at least 500 rows, (prefer over a thousand rows), and at least 10 predictor attributes (prefer at least 20), and a target attribute. The target should be categorical.

   Build a naïve Bayes classifier for this data. Consider Python packages to handle numerical versus categorical predictors.
   Evaluate the generalization error of the classifier three ways:
   - From the training data
   - From a test set of 20% of the data
   - From crossvalidation with 5 folds

a) Provide the code and the results of your analysis. Comment on any differences in these estimates on generalization error.
b) Provide the code and a confusion matrix, summary statistics, and a ROC curve calculated from the **crossvalidation only**.

2. Consider the model $y = 5 - 2x + x^2 + e$ where e is normal with mean 0 and std deviation 1.

   Replicate the following process 50 times so there are 50 predictions at each of the values specified for x below for each model specified below:

   Generate 100 values from a standard normal random variable to represent e.
   Generate 100 values for x uniformly from 0 to 10. Use the model to generate values for y.

   Fit the following four regression models and calculate the prediction from each model at $x = 5$ and $x = 15$:
   Polynomial model of degree 1 Polynomial model of degree 2
   Polynomial model of degree 5

   a) Prepare side by side box plots of the prediction error, $y - yhat$, (no squares or absolute values) for each of the four models. Make one plot for $x = 5$ and one plot for $x = 15$. Provide the output and code.
   b) Provide estimates of variance and bias for each model in a table for $x = 5$ and $x = 15$.
   c) Which model has the greatest bias, which has the greatest variance? Is this what you expected? Explain. Are there differences between the results for $x = 5$ and $x = 15$? Explain.