Machine Learning with Networks *ECEN765.600, Fall 2018*

Assignment #1: Naive Bayes & K-Nearest Neighbor

Due date: September 28th, 2018 (Friday)

"On my honor, as an Aggie, I have neither given nor received unauthorized aid on this academic work."

Signature:

Name:

1. Programming assignment (You are encouraged to program the specific machine learning algorithms by yourself. In case you are using the ones in existing packages, you would at least need to read the code!)

Implement the naive Bayes classifier (NBC) and k-nearest-neighbors (KNN) for digit recognition:

1. (10pts) Get familiar with the data set on eCampus, which was originally provided in "Machine Learning in Action" (`digits.zip`, in which it contains two folders with training and testing images. Each image has been transformed into texts with '0' and '1' representing the original foreground and background of digit images. From the training data, pick one sample for each digit, display them as binary images to visualize how they look like. Convert all the data (training and testing) to vectors so that they can be used later on.

2. (10pts) Implement naive Bayes classifier for digit recognition: Use the training data to learn conditional probabilities. Compute both training and testing error rates.

3. (10pts) Implement k-nearest-neighbor classifier for digit recognition: Use the data in the training folder as training samples, for which the actual digit can be obtained by the first character of the file names. Evaluate and plot the testing error rates of k-nearest-neighbors using the data in the testing folder with varying k from 1 to 10.

   Discuss the trend of the above testing error rates. Compare them with the corresponding error rates from NBC. Briefly discuss pros and cons of KNN and NBC. Discuss possible ways to improve the performance of NBC.

4. (10pts) Repeat the previous experiments for both NBC and 1-nearest-neighbor with different numbers of training samples randomly drawn from the original training folder. Plot the training and testing error rates when the number of training samples increase from 100 to 1,800 (Typically similar numbers of samples from each class are drawn). Discuss the trends.

5. (10 bonus points) As discussed in class, scikit-learn tutorial has the MNIST dataset. Can we use MNIST data to help training to improve the prediction performance for this testing data?

2. Math assignment

1. (10pts) Suppose that a study shows that 90% of people who have contracted Creutzfeldt–Jakob disease ("mad cow disease") ate hamburgers prior to contracting the disease. Creutzfeldt–Jakob disease is incredibly rare; suppose only one in a million people have the disease.

   If you eat hamburgers, should you be worried? Does this depend on how many other people eat hamburgers?

2. (10pts) Under zero-one loss, the Bayes' decision rule chooses class k if $p(\text{class k}|x) > p(\text{class j}|x)$ for all $j \neq k$. Imagine instead we use a randomized decision rule, choosing class j with probability $q(\text{class j}|x)$. Calculate the error for this decision rule, and show that the error is minimized by using Bayes' decision rule.

3. (10 bonus pts) **[Error bound]** Given p, $p' \in [0,1]$ and $y' \in \{0,1\}$. Show that $Pr_{y\sim p}(y \neq y') \leq Pr_{y\sim p'}(y \neq y') + |p - p'|$.