CST 383 - Intro to Data Science

Dr. Glenn Bruns

# Lab: system design

In this lab we will create some learning curves from the German credit data set.

1. Create a file ‘system-design.py’, and insert [this code](https://drive.google.com/file/d/1ZyCdzPWHlzmhmEfwDyvUGp6arPc7HdbV/view?usp=sharing) into it.
2. Read and understand the code. It trains a kNN classifier on increasingly large training sets, and then makes predictions from both the training set and a fixed size test set.
3. Add code to plot the learning curve. On the x axis you will have the training set size; on the y axis you will have one line for classification error with the training set, and another line for classification with the test set. Use different colors for the training and test error lines, and include a legend.
4. Extend the code by adding a loop so that you produce learning curves for k = 1, 3, 5, and 9. Put all the plots on one page. Put the value of k in the title of the learning curve plot.
5. **Explain the curves you get**. For example, what’s with the low training error when k = 1?
6. What do the learning curves tell you? Write some sentences to explain what the learning curves tell you about bias and variance.
7. Add some more features, and produce the 4 learning curve plots again. Explain how and why the learning curves changed.
8. If you still have time, experiment with different feature sets.
9. If you still have time, try again but using logistic regression.