CPSC 392: Introduction to Data Science
Fall 2018
Assignment 3
Due: September 27, 2018 @ start of class

Overview

Now that you've had a chance to explore some of the most fundamental algorithms of machine learning at a conceptual level, let's have you dive into some real code!

You have been provided with a Python module, KNN.py, which contains an **almost** complete implementation of the KNN algorithm (for K=1). You have also been provided a sample training data set and a sample testing data set…but the training data contains missing values and outliers. ☹

Your first job is to do an exploratory data analysis on the training data sample in order to identify and fix missing data and outliers.

Your second job is to complete the Python implementation so the code correctly runs and provides the classification accuracy on the test data. This will give you an appreciation for what the implementation of KNN looks like behind the scenes, and, more importantly, why it is so freakin' slow!

Instructions

1. Download the .py file and the .csv files to a directory on your computer.
2. Using an editor of your choosing, complete the python implementation. There are comments in the code that tell you what functions need to be completed
3. Run the python code and report the classification accuracy. (The code expects all the files to be in the same directory.)

Questions to Answer

1. What missing values and outliers did you identify? How did you fix them? If you discover outliers, you must impute values to replace them, not delete them.
2. What is the accuracy of nearest neighbor on the provided data? (Copy and paste the output of your program to answer this question)
3. What lines of code are the most computationally intense? (You can copy your py file into word and highlight the lines)
4. How would your distance metric have to change if the data contained categorical variables in addition to the continuous variables?

Deliverables

Upload your complete KNN.py file (be sure to add comments with your name, etc) and cleaned training data csv to Blackboard by the deadline. Also include a word doc with the answers to the questions above.

Challenge question for extra credit: can you convert to make the K a parameter? i.e – allowing K to take on any value you wish.