DATA 221
Homework 6 (rev 2)

W. Trimble

Due: Wednesday 2022-05-25 - 11:59pm (Middle of 9th week)

1. (Bishop CH 14.11) Classification tree evaluation Consider a data set comprising 400 data points from class C1 and 400 data points from class C2. Suppose that a tree model A splits these into (300,100) at the first leaf node and (100, 300) at the second leaf node, where (n, m) denotes that n points are assigned to C1 and m points are assigned to C2. Similarly, suppose that a second tree model B splits them into (200, 400) and (200, 0). Evaluate the misclassification rates for the two trees and hence show that they are equal. Similarly, evaluate the cross-entropy and Gini index for the two trees and show that they are both lower for tree B than for tree A.

2. k-means The "Online Retail II" dataset, credited to Daqing Chen at London South Bank University, is hosted by the UCI machine learning repository contains one million items ordered from an online UK retailer; the sales date from 2009-2011.

   `https://archive.ics.uci.edu/ml/datasets/Online+Retail+II`

   Apply reasonable cleaning to this dataset (or a subset of it) and apply k-means clustering for a range of k to a subset of the features. There are 4000 customers, 4000 items in inventory, the customers range from 40 countries, but there are few other numerical values usable for clustering. One-hot encoding for the categorical variables should work; our computers can handle 4000x4000 matrices. This is also a good case for dimension reduction.

   Make a plot of the Bayesian Information Criterion and Akaike Information criterion for the fits as a function of k and choose the best k.

3. Produce a visualization of some sort (scatterplot, histogram, or violin plots are reasonable choices) that explains how some of the clusters differ. If you don't know where to start, you can look at correlations between the features and indicator variables for the class identity (the output of the k-means classification).

4. Plot the receiver operating characteristic curve for identifying "7" vs all other digits for the logistic classifier / single layer perceptron classifier for one of your MNIST digit classifiers. You want to extract a numerical score for the classification of number 7 vs other, and use this number to find relationships between FPR and FNR that are different from the standard classifier output. Plot the ROC for "0" vs all other digits on the same graph.