Homework 1 CM146

Stephanie Chen

October 21 2019

1. a) An input X will be labeled 0 if and only if X_1 = X2 = X3 = 0. Over the 2^n examples, there are 2^{n-3} possible binary vectors with X1 = X2 = X3 = 0. Since 2^n is much larger than the 2^{n-3} samples that should be labeled as 0, the best 1-leaf decision tree should predict 1 for every input and make 2^{n-3} errors. The error rate is then $\frac{2^{n-3}}{2^n} = 2^{-3} = \frac{1}{8}$

   b) There is no feature that you can put at the root that will reduce the error rate to below $\frac{1}{8}$. Splitting the data on X_i for i $\geq$ 4 will split the training data so $\frac{7}{8}$ of the data split in the two resulting leaves are labeled 1. In both leaves, the tree will predict 1, so the single split tree makes the same amount of error as a single leaf tree always predicting 1. Similarly, splitting on X_1, X_2, or X_3 will produce one leaf containing only 1s as labels and one leaf where the proportion of data labeled 1 is $\frac{3}{4}$. Like the tree splitting on X_i, these trees predict 1 on both leaves, so it makes the same amount of error as does a 1-leaf tree.

   c)

$$
\begin{aligned}
Entropy = H(Y) &= -(\frac{1}{8}log(\frac{1}{8}) + \frac{7}{8}log(\frac{7}{8})) \\
&= 0.375 + 0.16 = 0.54
\end{aligned}
$$

   d) H($X_4 = 0$) $= -(\frac{1}{8}log(\frac{1}{8}) + \frac{7}{8}log(\frac{7}{8})) = 0.54$

   H($X_4 = 1$) $= -(\frac{1}{8}log(\frac{1}{8}) + \frac{7}{8}log(\frac{7}{8})) = 0.54$

   H(X_4) $= \frac{1}{2}(0.54) + \frac{1}{2}(0.54) = 0.54$

   From 1c), We calculate that H(Y) = 0.54, so information gain from X_4 is 0.54 - 0.54 = 0

   e) We can split the data by any of the three critical features X_1, X_2, or X_3 to reduce entropy by a nonzero amount. If we split by X_3,

$$H(X_3 = 0) = -(\frac{1}{4}log(\frac{1}{4}) + \frac{3}{4}log(\frac{3}{4})) = 0.81125$$

H(X_3 = 1) = 0

H(X_3) = $\frac{1}{2}(0) + \frac{1}{2}(0.81125) = 0.405$

H(Y) - H(X_3) = 0.54 - 0.405 = 0.135

2. a) To find q that maximizes B(q), we solve $\frac{d}{dq}(B(q)) = 0$.

$$\frac{d}{dq}(B(q)) = -log_2 q + log_2(1-q) = 0$$

We can rewrite the above equation as:

$$log(q) = log(1-q)$$

This is true when q = $\frac{1}{2}$.

b) To show the information gain is 0, we need to show

$$Gain(S, A) = H(S) - \sum_{v \epsilon Values(A)} \frac{|S_v|}{|S|} H(S_v) = 0$$

If we calculate entropy of a Bernoulli variable B(q), the information gain equation becomes:

$$H(S) = B(\frac{p}{p+n}) - \sum_{i=1}^{k} \frac{|p_i + n_i|}{|p+n|} B(\frac{p_i}{p_i + n_i})$$

Note since we are just splitting all p and n examples into k disjoint subsets, $\sum_{i=1}^{k} p_i$ = p and $\sum_{i=1}^{k} n_i = n$. Then,

$$B(\frac{p}{p+n}) - (1)B(\frac{p}{p+n}) = 0$$

3. a) If a point can be its own neighbor, we can set k=0 to minimize the training error and just use the label of the training data.

b) Using a large k is bad can easily lead to wrong classification because there aren't many data points in this data set, and considering many neighbors would be considering points that really should not be considered as neighbors to a data point. Using a small k is not ideal either because this can easily lead to over-fitting, as the positive and negative points do not have a simple segregation using Euclidean distance.

c) Using k=5 or k=7 would minimize the error using leave-one out cross validation. Using these values for k gives an error rate of 2/7. Furthermore, there is no value of k that can correctly predict the 2 +'s at the upper left corner and 2 -'s in the bottom right corner using leave one out cross validation.

4. a) For the pclass feature, there are more people in pclass = 1 who survived than died, and the number survived is also higher than those survived in pclass = 2 and pclass = 3. In pclass = 3, there are significantly more (almost 3 times) people who died compared to those who survived.

   For the sex feature, there are more who survived in sex = 0 than sex = 1, and significantly more died than survived in sex = 1.

   For the age feature, there are the most death in the 20-30 age group. All age groups have more who died than survive except the 0-10 year old age group.

   For the sibsp (number of siblings/spouses aboard) feature, there is most death in group 0 (sibsp = 0). Only group 1 had more who survived than died.

   For the parch (number of parents/children aboard) feature, there is also most death in group 0 (parch = 0), indicating more passengers with parents/children survived.

   For the fare feature, we can see that passengers who pay a higher fare had more who survived than died, as more people who payed less than 100 died.

   For the embarked feature, most passenger embarked from location 2, and there are most passengers from group 2 who died. Only location 0 had more people who survived than died.

   b) The random classifier has a training error of 0.485.

   c) The decision tree classifier has a training error of 0.014.

   d) For k-Nearest Neighbors classifier, the training error for 3-Nearest Neighbors is 0.167, the training error for 5-Nearest Neighbors is 0.201, and the training error for 7-Nearest Neighbors is 0.240.

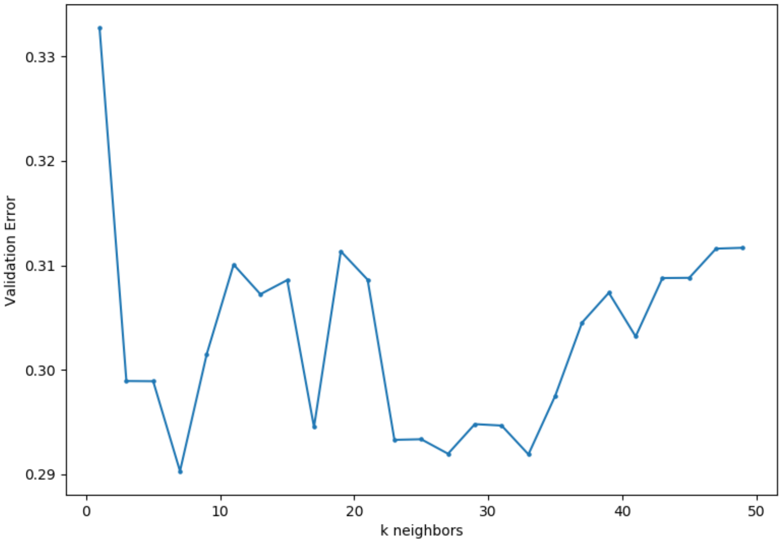
   e) For Majority Vote Classifier: – training error (average): 0.404, testing error (average): 0.407

   For Random classifier: – training error (average): 0.489, testing error (average): 0.487

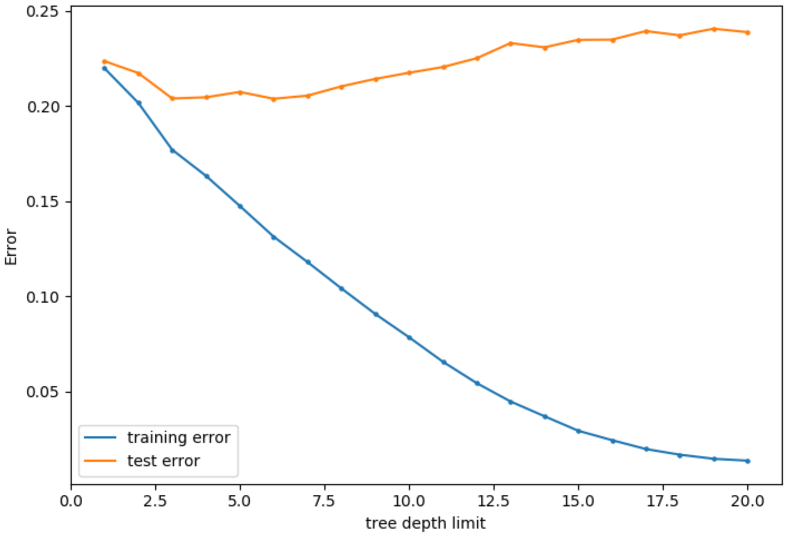
   For Decision Tree classifier: – training error (average): 0.012, testing error (average): 0.241

   For 5-Nearest Neighbors classifier: – training error (average): 0.212, testing error (average): 0.315

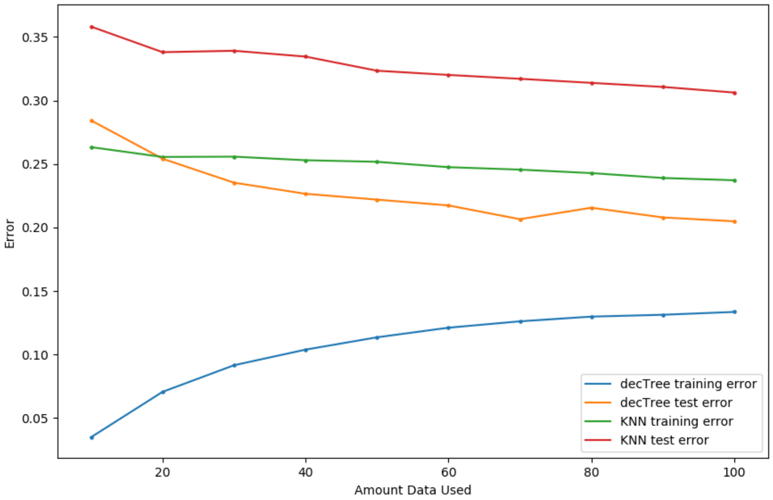
f) The optimal choice with lowest validation error is when k = 7 with validation error of 0.290. In addition, when k = 1 to 21, the error changes dramatically and when k = 23 to 33, fluctuation in validation error is more stable.



g) From the graph, we can see the best depth to use is 6, which is when test error is around 0.2 and before test error starts to increase. We can observe over-fitting from how training error decreased dramatically as test error increased steadily.



h) From 4f), we know the best k to use for KNN classifier is k=7. From 4g), we use tree depth of 6 for the decision tree classifier.



As seen in the plot, test error for both classifiers decrease as more data are used. Training error for decision tree increased while that of KNN decreased as more data are used.