CM146, Fall 2017
Problem Set 1: Decision trees and k-Nearest Neighbors
Due Jan 29, 2018

1 Problem 1

(a) It makes 2^{n-3} mistakes.

(b) No. Splitting by one of X_1, X_2, X_3 makes us choose to predict 1 for the $X_1 = 1$ branch, and 0 otherwise. This means the number of errors increases to $3 \cdot 2^{n-1}$! If we decide to split based on $X_4, X_5, ...$ then we will have an equal split of $Y = 1$ go to either side and whichever branch we predict to be 1 will end up causing 2^{n-1} errors in total.

(c) The entropy of the output label will be

$$H[Y] = -\left(\frac{7}{8}\ln\frac{7}{8} + \frac{1}{8}\ln\frac{1}{8}\right)$$

(d) Yes, we can split over the value of X_1:

$$\begin{aligned} Gain[Y, X_1] &= H[Y] - \sum_{v \in Values(X_1)} \frac{|S_v|}{|S|} H(S_v) \\ &= -\left(\frac{7}{8}\ln\frac{7}{8} + \frac{1}{8}\ln\frac{1}{8}\right) + \frac{1}{2}\left(0 + \frac{3}{4}\ln\frac{3}{4}\right) \approx 0.27 > 0. \end{aligned}$$

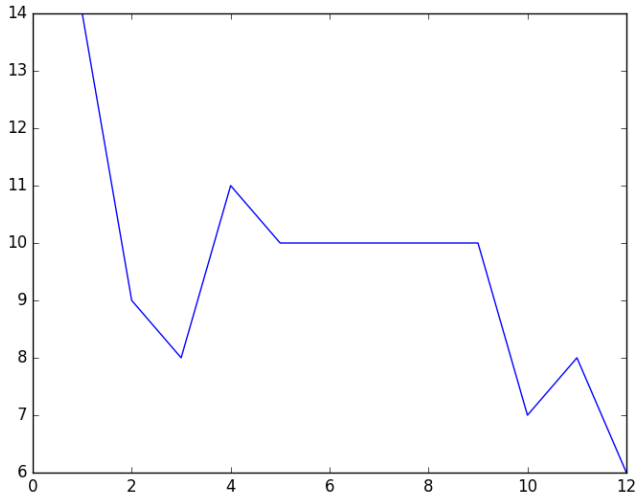
2 Problem 2

(a) By definition

$$\begin{aligned} Gain(S, X_j) &= H(S) - \sum_{v \in Values(X_j)} \frac{|S_v|}{|S|} H(S_v) \\ &= B\left(\frac{p}{p+n}\right) - k\frac{p_k + n_k}{p+n} B\left(\frac{p_k}{p_k + n_k}\right) \\ &= B\left(\frac{p}{p+n}\right) - B\left(\frac{p}{p+n}\right) \\ &= 0 \end{aligned}$$

because $k(p_k + n_k) = p + n$ and $\frac{p_k}{n_k + p_k} = \frac{p}{p+k}$.

3 Problem 3



(a)

Figure 1: Number of correct matches vs. number of neighbors considered in model.

$k = 1$. Error is 0 because the training data points are their own neighbors exclusively and they are correctly classified by definition.

(b) Using large values of k might toss too long of a hook for sampling the surrounding points. Namely, if we have n clustered data points on one side, and n clustered data points on another, choosing $k > n$ would sample data from the cluster that the new point clearly does not belong to. If just one of the two clusteres has more training points than the other, and if k is really large, we can have all new points classified as belonging to the larger group even if they clearly belong to the smaller one.

Having small k is also not good. Imagine our training data has a cluster of a certain class clearly separated from all other training data points. Imagine that there is one outlier in the cluster that is classified differently than the entire cluster. If $k = 2$ all points that have the outlier as the nearest neighbor will become incorrectly classified ($k = 2$ means we look at nearest neighbor only and self, but self is not pre-classified).

(c) $k = 5$ and $k = 7$. I ran one-out-cross-validation in Python and got the result.

4 Programming exercise

4.1 Visualization

(a) The graph of number of survivors vs. class obviously shows that the greater the class the greater the ratio of survivors. In fact, the first class has more survivors than killed, second about equal, while the third class has an overwhelming majority killed. Note also that the number of survivors in second and third class is nearly identical, however the fact that there are so many more people in the third class reduces that class's survival rate drastically. Notice also that the number of survivors in first class is only about 25% greater than that of the other two classes. We can also observe that the first and second class were about the same in size.

The graph of number of survivors vs. sex tells us about twice as many women survived than men. Also, there were significantly less women than man. Due to these two facts, about three times as many women survived than died, while for man this ratio is about exactly opposite.

The graph of survivors vs. age tells us that the majority of passangers were 20-30 years old. Also, the survivor rate of these two groups is very low compared to others (but most of the survivors were also from these two groups). Ages 40-50 have very bad survival rates (worst than 20-30). 50 year olds had really good survival rates which is probably related to that fact that people over 50 on the ship were more probable to be rich and less probable to be working on the ship. Also children ages 10 and under have a significantly higher survival rate than any other group.

The graph of survivors vs. siblings/spouses tells us that most people were traveling without siblings or spouses and that they had the worst survival rate out of any other group. The group that had one sibling/spouse had a significantly better survival rate than 0. Other groups number too few passangers to be statistically significant.

The graph of survivors vs. parent/children tells us a very similar story as the siblings/spouses graph. However, there are even more people without a single parent or child aboard. The survival rates for 0, 1 parent/child is about the same as that for 0, 1 sibling/spouse respectively.

The graph survivors vs. fare graph tells us that the number of passangers paying the lowest price for the ticket is by far in the majority

compared to other groups. Also, all but the lowest fare price group had more people survive than die. The survival rate for the lowest fare was about 1/3.

Graph of survivors vs. embarked port, tells us that most people joined the ship on the last (second) port. Only the first port had more survivors than killed and the last and port 1 had about twice as many people die as did survive.

(b)

(c) The training error is 0.014.

(d)

Model	Error
KNN3	0.167
KNN5	0.201
KNN7	0.240

(e)

Model	Training error	Test error
KNN3	0.169	0.309
KNN5	0.212	0.315
KNN7	0.237	0.311
Decision Tree	0.011	0.240
Random Sample	0.489	0.486

(f) The best value of k is 7.

(g) Optimal max depth is 7. Obviously, the bigger the depth of the tree, the better the fit to the Training data. However, we see that the test error barely moves. Specifically, it loses precision after max depth 7. After that, the test error increases while the training data keeps decreasing. This implies we are fitting the data too well to the specific training set, and the tree is not general and robust enough to cope with data outside the training set.

(h) We note that all of the error rates decrease except the Decision Tree training error, which steadily rises with the ratio of training data actually used for training. We notice the KNN errors converge to one another. This is a good indication that the data is neither over- nor underfitted. However, with the Decision Tree model, we can notice overfitting, since the Training error is so low compared to the test error (about 5 times smaller in the best case), *i.e.* it fits too well to the training data and does not generalize well.

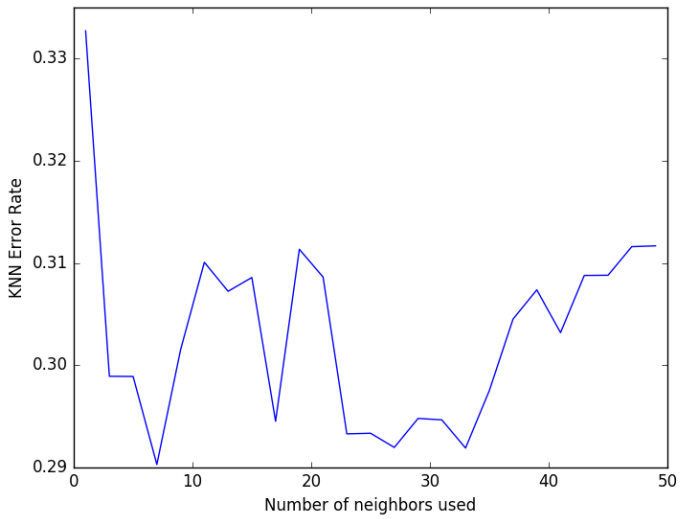


Figure 2: Error rate of test data vs. number of neighbours in KNN.

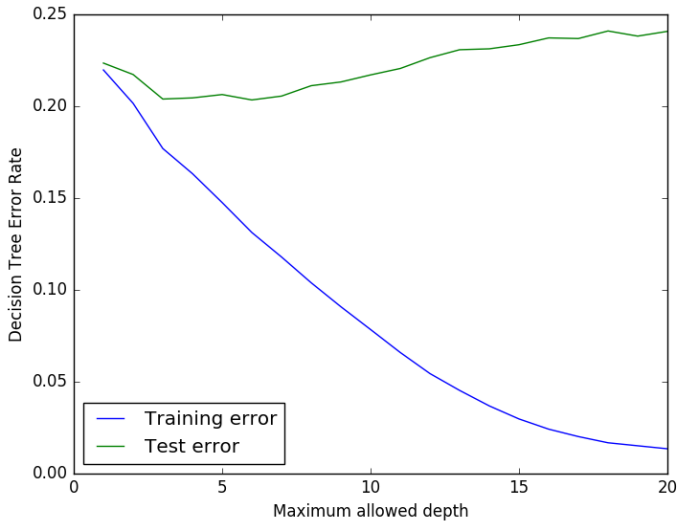


Figure 3: Error vs. max depth

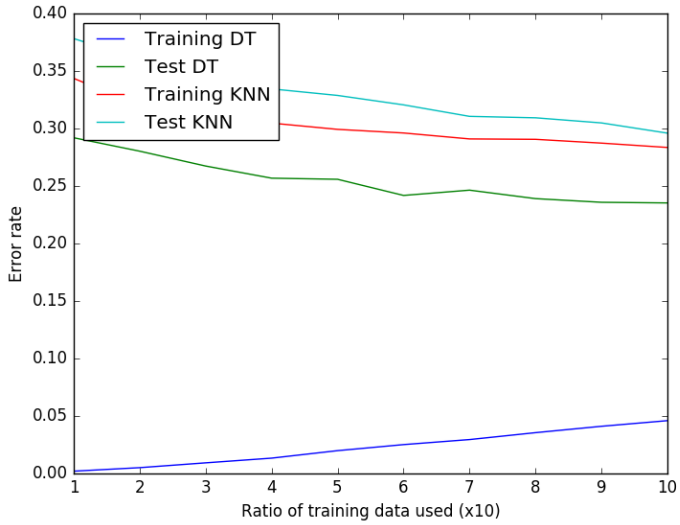


Figure 4: The Learning Curve