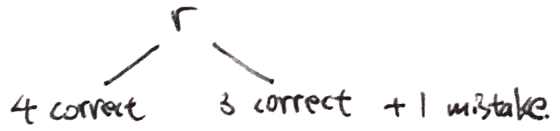
Tianyang Zhang

404-743-024

#1 a). $\pi_i (i \geq 4)$ don't effect value of Y, only π_1, π_2, π_3 effect value of Y.
There are 8 combinations of π_1, π_2, π_3, and there is one mistake among these 8 options.
Therefore, over 2^n training examples, there should be $\frac{2^n}{8}$ mistakes.

b). There is no such split that reduces the number of mistakes by at least one.

Split on π_i, $i \geq 4$ makes no different since it's always $\frac{1}{8}$ mistake.
Split on π_1, π_2, or π_3 would be:



c). $H[Y] = -\frac{1}{8}\log(\frac{1}{8}) - \frac{7}{8}\log(\frac{7}{8}) = 0.543$

d). Do the split in b) on π_1, π_2, π_3, the resulting entropy would be decreased:

$$H[Y|\pi_1] = \left(-\frac{1}{4}\log(\frac{1}{4}) - \frac{3}{4}\log(\frac{3}{4}) - 0\right) \times \frac{1}{2}$$

$$= 0.406$$

#2 a). $S_1, S_2, S_3, \ldots, S_k$ are disjoint dividents of X_j, therefore,

$$\frac{P}{P+n} = \frac{P_k}{P_k+n_k} \quad \text{for all } k=1,\ k=2,\ \ldots,\ k=k$$

$$\therefore H(S_k) = B\left(\frac{P}{P+n}\right) \text{ for all } k$$

$$H(S|X_j) = H(S_1)\frac{P_1+n_1}{P+n} + H(S_2)\frac{P_2+n_2}{P+n} + \cdots + H(S_k)\frac{P_k+n_k}{P+n}$$

$$= H(S) \cdot \frac{P_1+P_2+\cdots+P_k+n_1+n_2+\cdots+n_k}{P+n}$$

$$= B\left(\frac{P}{P+n}\right)$$

Therefore, $H(S) - H(S|X_j) = B\left(\frac{P}{P+n}\right) - B\left(\frac{P}{P+n}\right) = 0$
Thus, the information gain of this attribute is 0.

#3 a) Since a point can be it's own neighbor, so $k=0$ minimizes the training set error, the resulting training error is 0.

b) Too big k may lead to misclassify on datapoints, too small k may leads datapoints not fit in graph.

c). $k=5$ or $k=7$ minimizes the LOO-CV error for this dataset. The resulting error is $\frac{4}{14}$.

#4. a).

Pclass: This plot shows that survive rate for first-class passengers is the highest, and it's extremly low on survive rate of third-class passengers. which is: $P(S=1|x=1) > P(S=1|x=2) > P(S=1|x=3)$

sex: Shows that Passenger with sex = 0 are most likely survive, and sex = 1 are most likely not survive. Expect that 0 for female and 1 for male, this feature is a good one to reduce entropy in dataset.

Age: Only children below 10 years old has higher survive rate than dead rate. Most people are in range 20 - 40 years old, but they have lower survive rate than othe age.

SibSp: This plot shows that passengers with 1 or 2 Siblings or spouses have higher survive rate.

Parch: This plot shows that most people don't have parent/child on board, but peple with 1 or 2 parent/child on board have higher survive rate.

Fare: This plot shows most people has no fare. People with fare have much higher survive rate.

Embarked: People embarked at 0.0 has highest survive rate.

b). The error I got after implementing the RandomClassifier is 0.485 as expected.

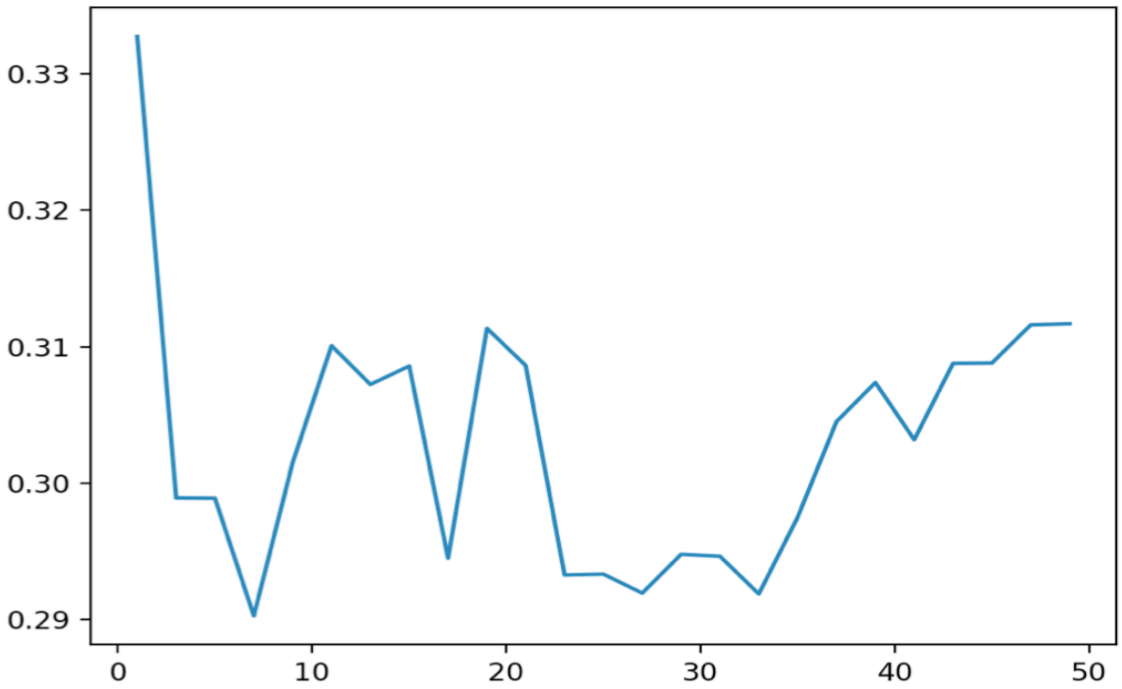
c). The training error for DecisionTreeClassifier is 0.014.

d). The training error for KNeighborsClassifier for k=3, 5, 7 are 0.167, 0.201, 0.240

e). Avg training and testing errors:

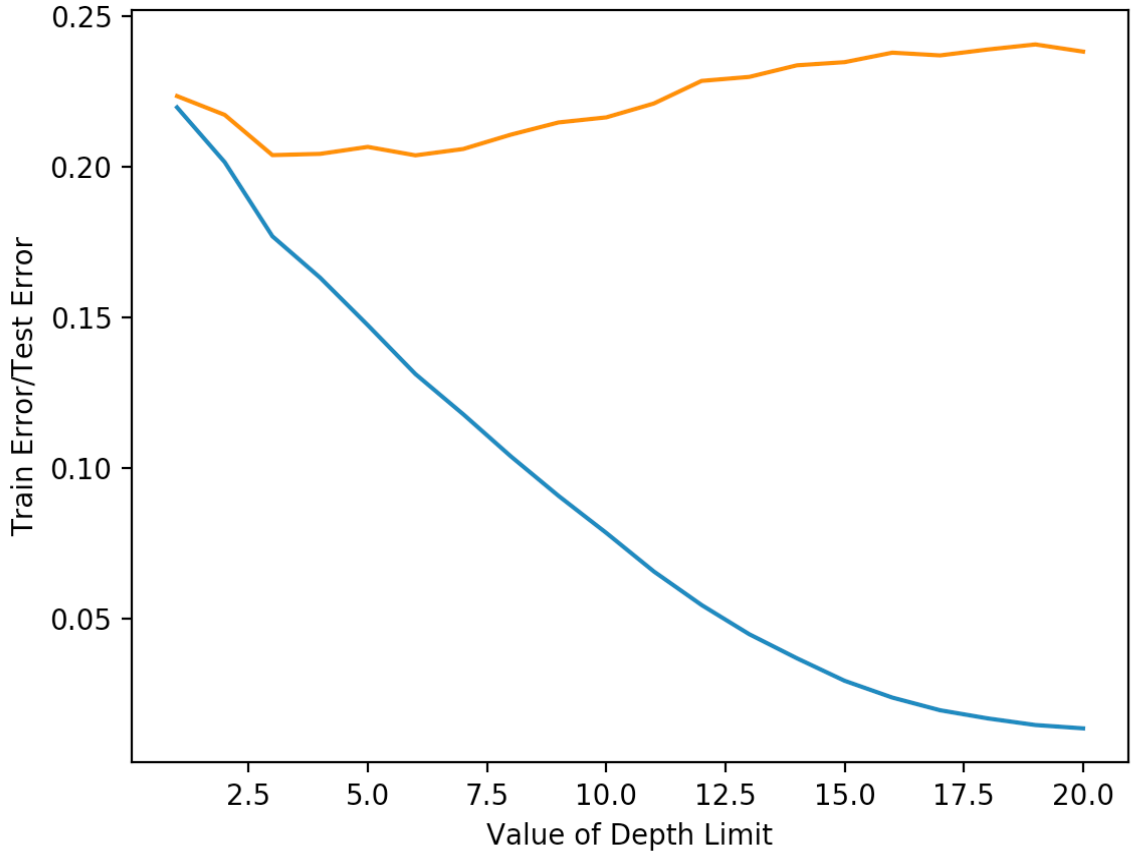
	Training error	Testing error
MajorityClassifier:	0.4038	0.4073
RandomClassifier:	0.4890	0.4866
DecisionTreeClassifier:	0.0115	0.2408
KNeighborsClassifier:	0.2124	0.3149

f). The best k value is 7, which has about 30% cross-validated error:



9). The best depth is 3, which has 20.4% testing error. I also see overfitting, since the error percentage is increasing after depth = 6.

* Yellow line is testing error, Blue line is training error.



h) The distribution of testing error are about same between decision tree and KNN. However for decision tree, training error is increasing while size of training data increasing. For both graph, testing error is always larger than training error.

* Yellow line is testing error, Blue line is training error.