CM146, Winter 2018
Problem Set 1
Due Jan, 30, 2018

1 Problem 1

(a) **Solution:** The best case is 2^{n-3}. When the machine only guesses 1, the only time it will be wrong is when all X_1 X_2 X_3 are 0. Otherwise, it will be right.

(b) **Solution:** No. The only attributes that will have impact on the results are X_1 X_2 X_3. However, if the tree relies on either one of the three, the number of mistakes will increase, because the tree will guess 0 whenever $X_1 = 0$, while X_2 or X_3 could be 1 and the result should be 1.

(c) **Solution:** $-\frac{2}{16}log(\frac{2}{16}) - \frac{14}{16}log(\frac{14}{16}) = 0.54$

(d) **Solution:** We can split upon X_1.

$$H(Y) = 0 + (-\frac{6}{8}log(\frac{6}{8}) - \frac{2}{8}log(\frac{2}{8})) * \frac{1}{2} = 0.41$$

2 Problem 2

Solution:

$$H(S) = B(\frac{p}{p+n}) = -\frac{p}{p+n} * log\frac{p}{p+n} - (1 - \frac{p}{p+n})log(1 - \frac{p}{p+n})$$

$$G = H(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|}(S_i)$$

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

$$= 1 * (-\frac{p_i}{p_i+n_i} * log\frac{p_i}{p_i+n_i} - (1 - \frac{p_i}{p_i+n_i})log(1 - \frac{p_i}{p_i+n_i}))$$

Assume $\frac{p}{p+n} = x$ and $\frac{p_i}{p_i+n_i=y}$, $\sum_{i=1}^{k} p_i = p$, $\sum_{i=1}^{k} n_i = n$
We want to prove $x = y$. Suppose $x \neq y$

$$\frac{p}{p+n} = x$$

$$n = \frac{p}{x} - p$$

$$\frac{p_i}{p_i+n_i} = y$$

$$n_i = \frac{p_i}{y} - p_i$$

From $\sum_{i=1}^{k} p_i = p$,

$$\sum_{i=1}^{k} n_i = \sum_{i=1}^{k} \frac{p_i}{y} - p_i = \frac{p}{y} - p$$

Since $y \neq x$, $\frac{p}{y} - p \neq \frac{p}{x} - p$. Therefore, $\sum_{i=1}^{k} n_i \neq n$, which contradicts the premise that $\sum_{i=1}^{k} n_i = n$. Thus, $x = y$. Hence, $\sum_{i=1}^{k} \frac{|S_i|}{|S|}H(S_i) = 1 * (-\frac{p_i}{p_i+n_i} * log\frac{p_i}{p_i+n_i} - (1 - \frac{p_i}{p_i+n_i})log(1 - \frac{p_i}{p_i+n_i})) = -\frac{p}{p+n} * log\frac{p}{p+n} - (1 - \frac{p}{p+n})log(1 - \frac{p}{p+n}) = B(\frac{p}{p+n}) = H(S)$.

$$G = 0$$

3 Problem 3

(a) **Solution:** $k = 1$. The resulting error in this case is 0.

(b) **Solution:** If k is too large, it will smooth out some important point that might overlook some regions. Too-small k will cause over-fit, as shown in part 1.

(c) **Solution:** $k = 7$. When k is 7, the error will be $\frac{4}{14} = \frac{2}{7}$.

4 Problem 4

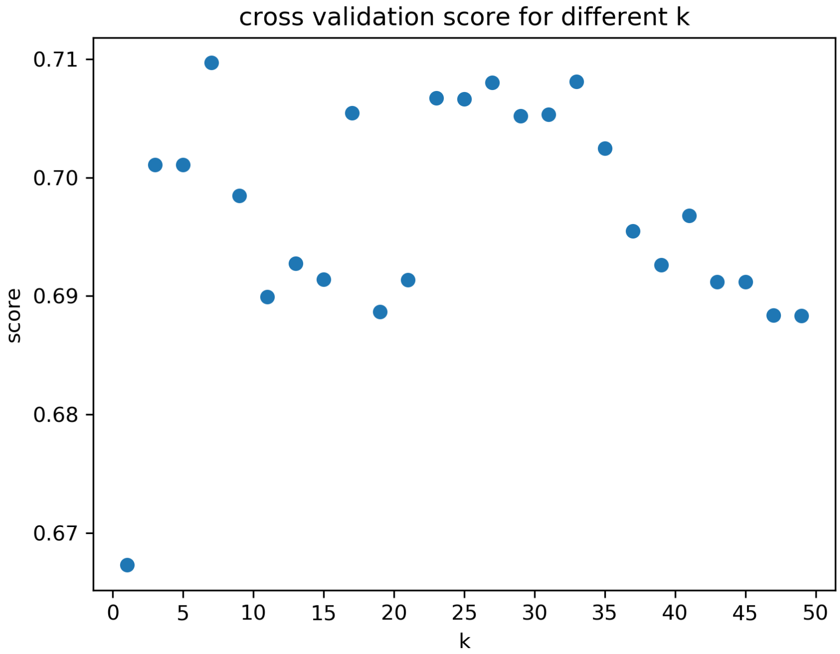
(a) **Solution:** When Pclass increases, the frequency of survived decreases and stays almost unchanged eventually, while the frequency of not-survived increases dramatically.
The female have higher frequency of survival while the male have much higher frequency of non-survival.
The frequencies of survival and non-survival show skewed bell curve shapes. The non-survival frequency peaks at age 20 and 30. And the ratio of non-survival vs survival peaks at age 40.
The frequencies of both survival and non-survival decreases when sibsp decreases with one exception that at sibsp = 4, the frequency of non-survival increases a little.
The frequencies of both survival and non-survival decreases with parch.
The frequencies of both survival and non-survival decreases dramatically with fare. And fare equal to 0 has the highest non-survival ratio.
Embark equal to 2 has the highest non-survival ratio.

(b) **Solution:** The error is 0.485.

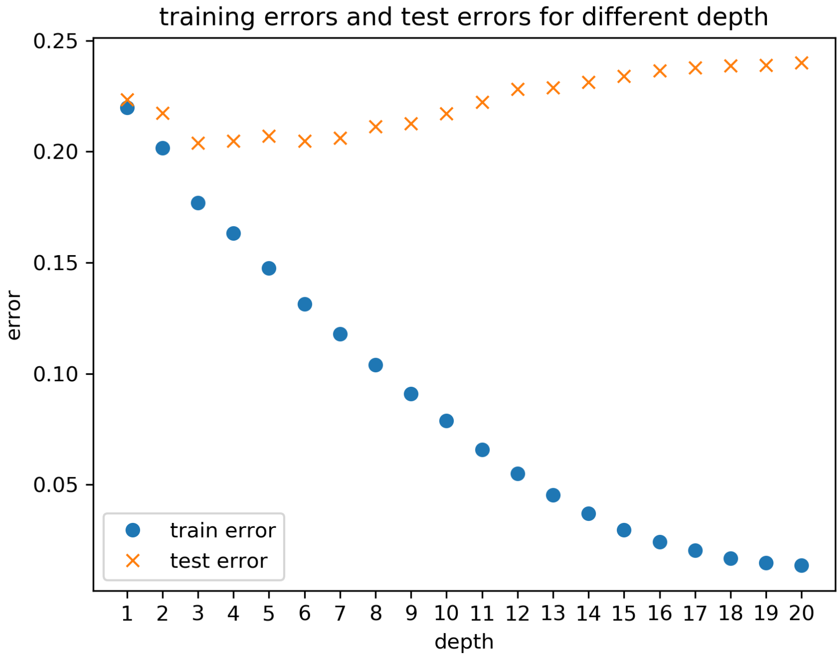
(c) **Solution:** The error is 0.014.

(d) **Solution:** For k = 3, the error is 0.167. For k = 5, the error is 0.201. For k = 7, the error is 0.240.

(e) **Solution:** The train error for Majority Vote Classifier is 0.404. The test error for Majority Vote Classifier is 0.407.
The train error for Random Classifier is 0.489. The test error for Random Classifier is 0.487.
The train error for Decision Tree Classifier is 0.012. The test error for Decision Tree Classifier is 0.240.
The train error for 5 Neighbor Classifier is 0.212. The test error for 5 Neighbor Classifier is 0.315.

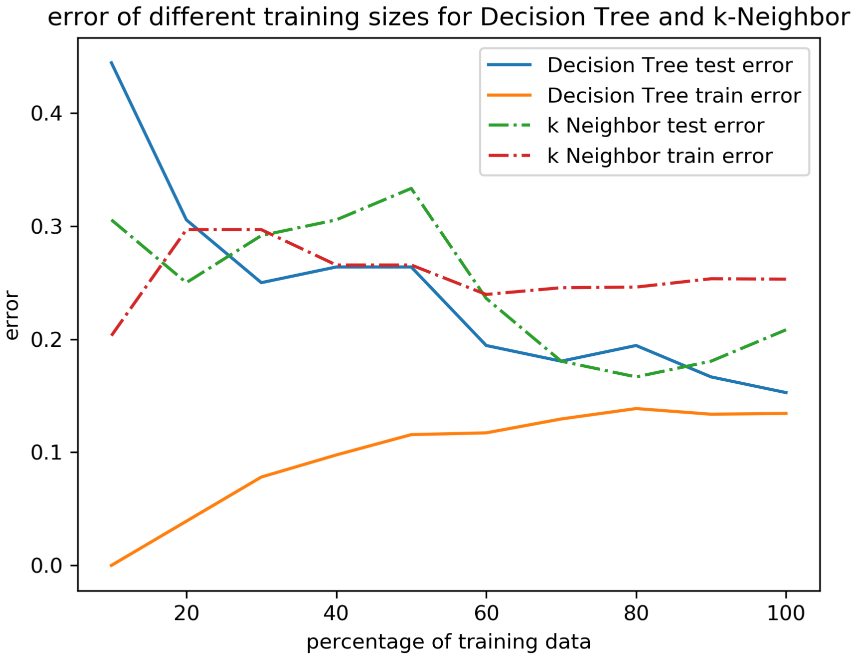


(f) **Solution:**
The k value for best score is 7.



(g) **Solution:**

The depth for best score is 6. The over fitting clearly happens after depth bigger than 6. The error increases with increasing depth.



(h) **Solution:**
With increasing training data size, the error of test error for decision tree classifier is decreasing while the training data error increasing. For k neighbor classifier, with training data size increasing, the error of both test error and training error fluctuate and keep almost the same as the beginning, eventually.