005230642 Rui Xu

1.(a). $n=4$, as shown in the table, 2 mistakes are made over the 16 training examples.

$n \geq 4$.

$Y=0$ if $X_1=0$ & $X_2=0$ & $X_3=0$ $\Rightarrow$ There are $1\times1\times1\times2^{n-3}$ examples labelled 0 by $f$ because there are two choices for $X_i$ $(i=4,\cdots,n)$

but these $Y$ are assigned 1 by the best 1-leaf decision tree.

$\Rightarrow 2^{n-3}$ mistakes are made over the $2^n$ training examples.

(b). No.

① If a feature $X_i$, $i=1,2,$ or $3$, is used to create a decision tree.

$X_i$ has two choices $\Rightarrow X_i=1$, $2^{n-1}$ examples. $X_i=0$, $2^{n-1}$ examples

When $X_i=1$, $Y=1$ by $f$ $\Rightarrow 2^{n-1}$ examples are assigned 1

When $X_i=0$, $Y=0$ if $X_1=0$ & $X_2=0$ & $X_3=0$, $2^{n-3}$ examples

$Y=1$ otherwise, $2^{n-1}-2^{n-3}=4\times2^{n-3}-2^{n-3}=3\times2^{n-3}>2^{n-3}$ examples

In order to make fewer mistakes. $Y$ of this part should be assigned 1

$\Rightarrow 2^{n-3}$ mistakes are made $\Rightarrow$ mistakes is not reduced.

② If a feature $X_i$, $i=4,\cdots$ or $n$, is used to create a decision tree.

$X_i$ has two choices $\Rightarrow X_i=1$, $2^{n-1}$ examples, $X_i=0$, $2^{n-1}$ examples.

When $X_i=1$, $Y=0$ by $f$ has $2^{(n-1)-3}=2^{n-4}$ examples

$\Rightarrow Y=1$ has $2^{n-1}-2^{n-4}=8\times2^{n-4}-2^{n-4}=7\times2^{n-4}>2^{n-4}$ examples

In order to make fewer mistakes. $Y$ of this part should be assigned 1

$\Rightarrow 2^{n-4}$ mistakes are made.

When $X_i=0$, similarly, $2^{n-4}$ mistakes are made

$\Rightarrow 2^{n-4}+2^{n-4}=2^{n-3}$ mistakes are made $\Rightarrow$ mistakes is not reduced

⇒ There is no single split that reduces the number of mistakes by at least one.

(c). $H[Y] = -P(Y=1)\log_2 P(Y=1) - P(Y=0)\log_2 P(Y=0)$

$= -\frac{14}{16}\log_2\frac{14}{16} - \frac{2}{16}\log_2\frac{2}{16} \approx 0.544$

(d). $X_4 = 0$, $P(Y=1|X_4=0) = \frac{7}{8}$, $P(Y=0|X_4=0) = \frac{1}{8}$

$H[Y|X_4=0] = -\frac{7}{8}\log_2\frac{7}{8} - \frac{1}{8}\log_2\frac{1}{8} \approx 0.544$

$X_4 = 1$, $P(Y=1|X_4=1) = \frac{7}{8}$, $P(Y=0|X_4=1) = \frac{1}{8}$

$H[Y|X_4=1] = -\frac{7}{8}\log_2\frac{7}{8} - \frac{1}{8}\log_2\frac{1}{8} \approx 0.544$

conditional entropy: $\frac{8}{16} \times 0.544 + \frac{8}{16} \times 0.544 = 0.544$

information gain: $0.544 - 0.544 = 0$

(e). Yes. A split on $X_i$ ($i = 1, 2, \text{or } 3$) reduces the entropy of the output Y in Table 1 by a non-zero amount.

$X_i = 0$, $P(Y=1|X_i=0) = \frac{6}{8}$, $P(Y=0|X_i=0) = \frac{2}{8}$

$H[Y|X_i=0] = -\frac{6}{8}\log_2\frac{6}{8} - \frac{2}{8}\log_2\frac{2}{8} \approx 0.811$

$X_i = 1$, $P(Y=1|X_i=1) = \frac{8}{8}$, $P(Y=0|X_i=1) = \frac{0}{8}$

$H[Y|X_i=1] = -\frac{8}{8}\log_2\frac{8}{8} - \frac{0}{8}\log_2\frac{0}{8} = 0$

conditional entropy: $\frac{8}{16} \times 0.811 + \frac{8}{16} \times 0 = 0.4055$

information gain: $0.544 - 0.4055 = 0.1385 > 0$

⇒ Any split on feature $X_1$, $X_2$ or $X_3$ reduces the entropy of the output Y in Table 1. The resulting conditional entropy of Y is 0.4055.

2. (a). $B'(q) = -q\log q - (1-q)\log(1-q)$

$= -\log q - q\frac{1}{q\ln 10} - (-1)\log(1-q) - (1-q)\frac{-1}{(1-q)\ln 10}$

$= -\log q - \frac{1}{\ln 10} + \log(1-q) + \frac{1}{\ln 10}$

$= -\log q + \log(1-q) = \log\frac{1-q}{q} = 0$

$\frac{1-q}{q} = 1 \Rightarrow q = 1-q \Rightarrow q = 0.5$

$q < 0.5$ $B'(q) > 0$ $B(q)$ increases on $[0, 0.5)$

$q > 0.5$ $B'(q) < 0$ $B(q)$ decreases on $(0.5, 1]$

$\Rightarrow q = 0.5$ maximizes $B(q)$

(b). let $p = \Sigma_k p_k$, $n = \Sigma_k n_k$,

$\frac{p_k}{p_k + n_k}$ is the same for all $k \Rightarrow$ let $q = \frac{p_k}{p_k + n_k}$

$\Rightarrow \frac{p}{p+n} = \frac{\Sigma_k p_k}{\Sigma_k p_k + \Sigma_k n_k} = \frac{\Sigma_k p_k}{\Sigma_k (p_k + n_k)} = \frac{p_k}{p_k + n_k} = q$

The entropy of S: $H(S) = B\left(\frac{p}{p+n}\right)$

After spliting:

$\Sigma_k \frac{p_k + n_k}{p+n} H(S_k) = \Sigma_k \frac{p_k + n_k}{p+n} B\left(\frac{p_k}{p_k + n_k}\right) = \Sigma_k \frac{p_k + n_k}{p+n} B\left(\frac{p}{p+n}\right)$

$= \frac{\Sigma_k (p_k + n_k)}{p+n} B\left(\frac{p}{p+n}\right) = \frac{\Sigma_k p_k + \Sigma_k n_k}{p+n} B\left(\frac{p}{p+n}\right) = \frac{p+n}{p+n} B\left(\frac{p}{p+n}\right) = B\left(\frac{p}{p+n}\right) = H(S)$

information gain: $H(S) - H(S) = 0$

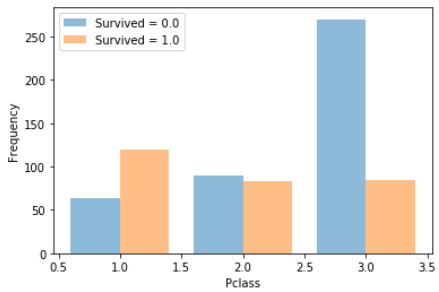
3. (a). $k=1$ minimizes the training set error for this dataset.
The resulting training error is 0.

(b) ① If $k$ is too large, then the samples shouldn't have impact on the prediction will be considered.
Underfitting may occur. Particularly, if $k$ equals the amount of the data. Then the output will always be the most common value of the dataset. For example, $k=13$. The leave-one-out cross-validation error is 1.

② If $k$ is too small, then the disturbance will influence the prediction.
Overfitting may occur. For example, $k=1$. The leave-one-out cross-validation error is 0.714.

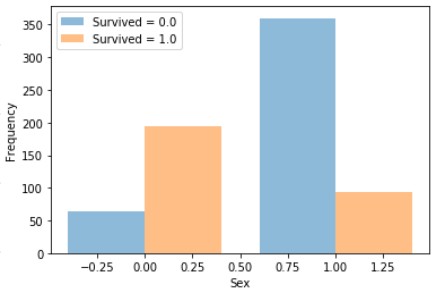
(c). $k=5$ or $k=7$ minimizes leave-one-out cross-validation error for this dataset.
The resulting error is 0.286.

4.1(a).①



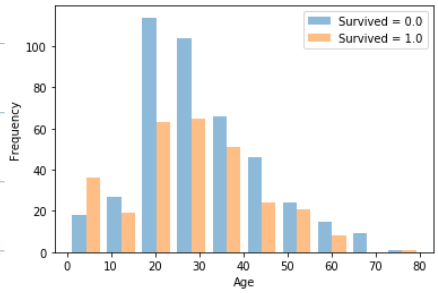

Pclass: People in the upper class are more likely to survive.

②



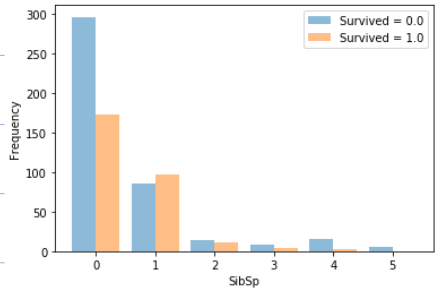

Sex: Female are more likely to survive than male.

③




Age: Children younger than 10 have the highest survival rate.
People whose age around 20, 45 and 78 tend to die.

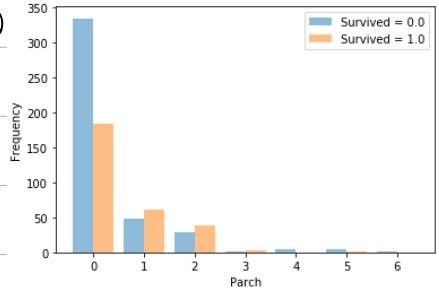
④




Sibsp: People with one sibling or spouse abroad have the highest survival rate.

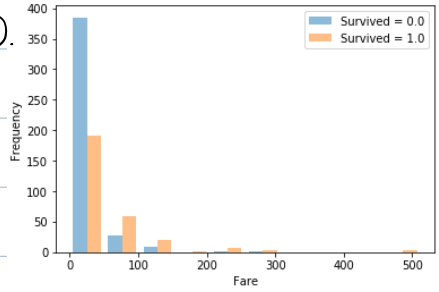
People without siblings or spouses abroad and with too much tend to die.

⑤



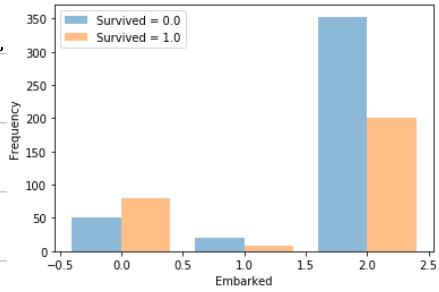

Parch: People with 1-3 parents or children abroad are more likey to survive.

⑥.




Fare: People paid more for the passenger fare are more likely to survive.

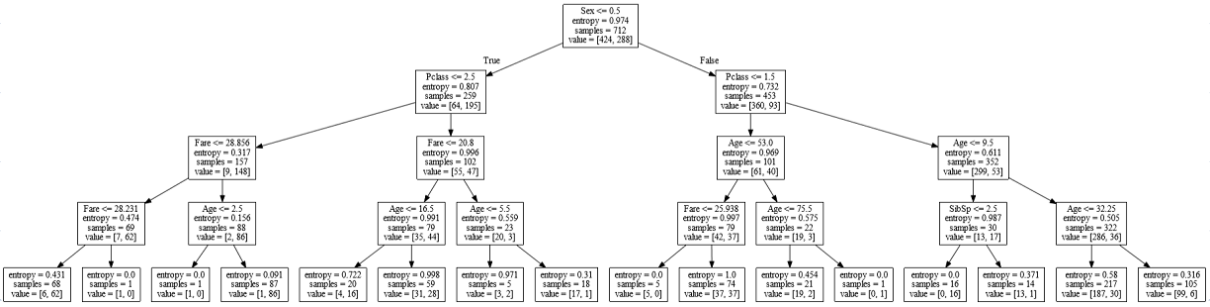
⑦.




Embarked: People embarked at Cherbourg have the highest survival rate.

4.2(b). The training error of RandomClassifier is 0.473.

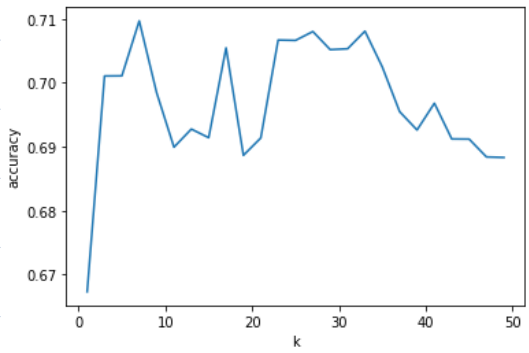
(c). The training error of DecisionTreeClassifier is 0.166. (depth limit=4)

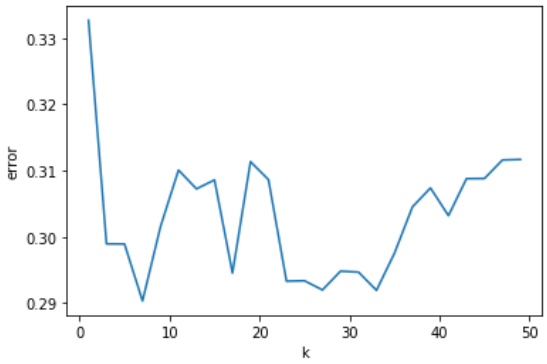



(d). The training error of KNeighborsClassifier with k=3 is 0.167.
The training error of KNeighborsClassifier with k=5 is 0.201.
The training error of KNeighborsClassifier with k=7 is 0.240.

(e). The average training error of MajorityVoteClassifier is 0.404.
The average test error of MajorityVoteClassifier is 0.407.
The average training error of RandomClassifier is 0.477.
The average test error of RandomClassifier is 0.476.
The average training error of DecisionTreeClassifier is 0.163. (depth=4)
The average test error of DecisionTreeClassifier is 0.205. (depth=4)
The average training error of KNeighborsClassifier with k=5 is 0.213.
The average test error of KNeighborsClassifier with k=5 is 0.315.

(f).



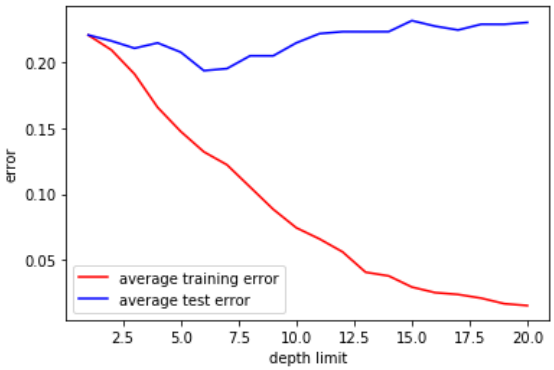




When k increases, the error first decreases rapidly, and then it fluctuates wildly.

The least validation error corresponds to $k=7$.

$\Rightarrow$ The best value of k is 7.

(g).

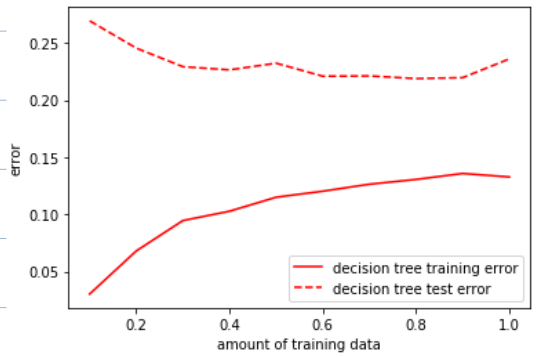


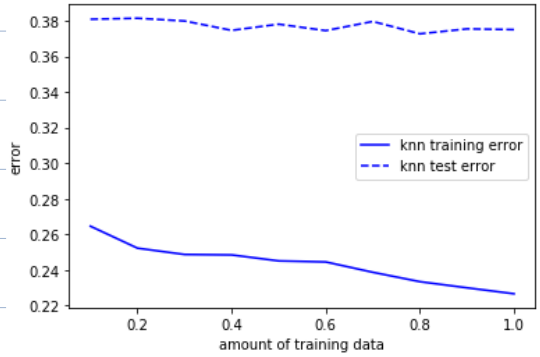

When depth limit increases, the average training error decrease; the average test error decreases and then increases.

Overfitting happens when depth limit is larger than 6, which means the classifiers do not generalize well to new data, thus causing test error increasing.

$\Rightarrow$ The best depth limit to use for this data is 6.

(h).




① The learning curve of decision tree:

When the amount of training data increases, the training error increases and ends at a low level, while the test error decreases but ends at a relatively high level.

② The learning curve of K Nearest Neighbors:

When the amount of training data increases, both the training error and test error decrease. The training error is at a low level while the test error is at a relatively high level.