Rio Sonoyama
Homework 3

1.

(a) For IsGoodRestaurant, there are five 1s and three 0s. Thus,

$$H(\text{IsGoodRestaurant}) = -\frac{5}{8}\log\frac{5}{8} - \frac{3}{8}\log\frac{3}{8} = \underline{0.9544340029}.$$

(b) First, for HasOutdoorSeating, there are four 1s and four 0s. So,

$$P(\text{HasOutdoorSeating} = 0) = P(\text{HasOutdoorSeating} = 1) = \frac{1}{2}.$$

Now, for those samples whose HasOutdoorSeating = 1, two of them have IsGoodRestaurant = 1 and two of them have IsGoodRestaurant = 0. So,

$$H(\text{IsGoodRestaurant} \mid \text{HasOutdoorSeating} = 1) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1.$$

For those samples whose HasOutdoorSeating = 0, three of them have IsGoodRestaurant = 1 and one of them has IsGoodRestaurant = 0. So,

$$H(\text{IsGoodRestaurant} \mid \text{HasOutdoorSeating} = 1) = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} = 0.8112781245.$$

Hence, by putting the values calculated above together, we have

$$\begin{aligned}
&H(\text{IsGoodRestaurant} \mid \text{HasOutdoorSeating}) \\
&= P(\text{HasOutdoorSeating} = 0)H(\text{IsGoodRestaurant} \mid \text{HasOutdoorSeating} = 0) \\
&\quad + P(\text{HasOutdoorSeating} = 1)H(\text{IsGoodRestaurant} \mid \text{HasOutdoorSeating} = 1) \\
&= \frac{1}{2} \times 1 + \frac{1}{2} \times 0.8112781245 \\
&= \underline{0.9056390623}\,.
\end{aligned}$$

(c) Similarly, by reading the table, we can compute the followings:

$H(\text{IsGoodRestaurant} \mid \text{HasBar})$:

- $P(\text{HasBar} = 1) = \frac{1}{4}$ and $P(\text{HasBar} = 0) = \frac{3}{4}$
- $H(\text{IsGoodRestaurant} \mid \text{HasBar} = 1) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$
- $H(\text{IsGoodRestaurant} \mid \text{HasBar} = 0) = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} = 0.9182958341$

So,

$$\begin{aligned}
H(\text{IsGoodRestaurant} \mid \text{HasBar}) &= P(\text{HasBar} = 0)H(\text{IsGoodRestaurant} \mid \text{HasBar} = 0) \\
&\quad + P(\text{HasBar} = 1)H(\text{IsGoodRestaurant} \mid \text{HasBar} = 1) \\
&= \frac{1}{4} \times 1 + \frac{3}{4} \times 0.9182958341 \\
&= \underline{0.9387218756}\,.
\end{aligned}$$

$H(\text{IsGoodRestaurant} \mid \text{IsClean})$:

- $P(\text{IsClean} = 1) = P(\text{IsClean} = 0) = \frac{1}{2}$
- $H(\text{IsGoodRestaurant} \mid \text{IsClean} = 1) = -1 \log 1 - 0 \log 0 = 0$
- $H(\text{IsGoodRestaurant} \mid \text{IsClean} = 0) = -\frac{1}{4} \log \frac{1}{4} - \frac{3}{4} \log \frac{3}{4} = 0.8112781245$

So,

$$\begin{aligned} H(\text{IsGoodRestaurant} \mid \text{IsClean}) &= P(\text{IsClean} = 0) H(\text{IsGoodRestaurant} \mid \text{IsClean} = 0) \\ &\quad + P(\text{IsClean} = 1) H(\text{IsGoodRestaurant} \mid \text{IsClean} = 1) \\ &= \frac{1}{2} \times 0 + \frac{1}{2} \times 0.8112781245 \\ &= \underline{0.4056390623}\,. \end{aligned}$$

$H(\text{IsGoodRestaurant} \mid \text{HasGoodAtmosphere})$:

- $P(\text{HasGoodAtmosphere} = 1) = P(\text{HasGoodAtmosphere} = 0) = \frac{1}{2}$
- $H(\text{IsGoodRestaurant} \mid \text{HasGoodAtmosphere} = 1) = -1 \log 1 - 0 \log 0 = 0$
- $H(\text{IsGoodRestaurant} \mid \text{HasGoodAtmosphere} = 0) = -\frac{1}{4} \log \frac{1}{4} - \frac{3}{4} \log \frac{3}{4} = 0.8112781245$

So,

$$\begin{aligned} &H(\text{IsGoodRestaurant} \mid \text{HasGoodAtmosphere}) \\ &= P(\text{HasGoodAtmosphere} = 0) H(\text{IsGoodRestaurant} \mid \text{HasGoodAtmosphere} = 0) \\ &\quad + P(\text{HasGoodAtmosphere} = 1) H(\text{IsGoodRestaurant} \mid \text{HasGoodAtmosphere} = 1) \\ &= \frac{1}{2} \times 0 + \frac{1}{2} \times 0.8112781245 \\ &= \underline{0.4056390623}\,. \end{aligned}$$

(d) Using the conditional entropies found in part (c), we can compute the information gains:

$$\begin{aligned} I(\text{IsGoodRestaurant}; \text{HasOutdoorSeating}) &= H(\text{IsGoodRestaurant}) - H(\text{IsGoodRestaurant} \mid \text{HasOutdoorSeating}) \\ &= 0.9544340029 - 0.9056390623 \\ &= \underline{0.0487949406} \end{aligned}$$

$$\begin{aligned} I(\text{IsGoodRestaurant}; \text{HasBar}) &= H(\text{IsGoodRestaurant}) - H(\text{IsGoodRestaurant} \mid \text{HasBar}) \\ &= 0.9544340029 - 0.9387218756 \\ &= \underline{0.0157121273} \end{aligned}$$

$$\begin{aligned} I(\text{IsGoodRestaurant}; \text{IsClean}) &= H(\text{IsGoodRestaurant}) - H(\text{IsGoodRestaurant} \mid \text{IsClean}) \\ &= 0.9544340029 - 0.4056390623 \\ &= \underline{0.5487949406} \end{aligned}$$

$$\begin{aligned} I(\text{IsGoodRestaurant}; \text{HasGoodAtmosphere}) &= H(\text{IsGoodRestaurant}) - H(\text{IsGoodRestaurant} \mid \text{HasGoodAtmosphere}) \\ &= 0.9544340029 - 0.4056390623 \\ &= \underline{0.5487949406} \end{aligned}$$

(e) Since IsClean and HasGoodAtmosphere give the same information gain, we choose the one on the left, namely IsClean.

(f) After splitting the dataset into two using IsClean, we have the following dataset D_1 and D_2:

D_1 (IsClean = 1)				
Sample#	HasOutdoorSeating	HasBar	HasGoodAtmosphere	IsGoodRestaurant
1	0	0	1	1
3	0	1	1	1
6	1	0	0	1
8	0	0	1	1

D_2 (IsClean = 0)				
Sample#	HasOutdoorSeating	HasBar	HasGoodAtmosphere	IsGoodRestaurant
2	1	0	0	0
4	0	0	0	0
5	1	1	0	0
7	1	0	1	1

Because $H(\text{IsGoodRestaurant}) = 0$ on D_1 (i.e. $H(\text{IsGoodRestaurant} \mid \text{IsClean} = 1) = 0$ on the original dataset), we stop splitting this branch.

For D_2, we know the entropy of IsGoodRestaurant from previous calculations. Namely,

$$H(\text{IsGoodRestaurant}) = 0.8112781245.$$

Now compute the conditional entropy of HasOutdoorSeating, HasBar, and HasGoodAtmosphere.

$H(\text{IsGoodRestaurant} \mid \text{HasOutdoorSeating})$:

- $P(\text{HasOutdoorSeating} = 1) = \frac{3}{4}$ and $P(\text{HasOutdoorSeating} = 0) = \frac{1}{4}$
- $H(\text{IsGoodRestaurant} \mid \text{HasOutdoorSeating} = 1) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.9182958341$
- $H(\text{IsGoodRestaurant} \mid \text{HasOutdoorSeating} = 0) = -0\log 0 - 1\log 1 = 0$

So,

$$\begin{aligned} & H(\text{IsGoodRestaurant} \mid \text{HasOutdoorSeating}) \\ &= P(\text{HasOutdoorSeating} = 1) H(\text{IsGoodRestaurant} \mid \text{HasOutdoorSeating} = 0) \\ &\quad + P(\text{HasOutdoorSeating} = 0) H(\text{IsGoodRestaurant} \mid \text{HasOutdoorSeating} = 1) \\ &= \frac{3}{4} \times 0.9182958341 + \frac{1}{4} \times 0 \\ &= \underline{0.6887218756}. \end{aligned}$$

$H(\text{IsGoodRestaurant} \mid \text{HasBar})$:

- $P(\text{HasBar} = 1) = \frac{1}{4}$ and $P(\text{HasBar} = 0) = \frac{3}{4}$
- $H(\text{IsGoodRestaurant} \mid \text{HasBar} = 1) = -0 \log 0 - 1 \log 1 = 0$
- $H(\text{IsGoodRestaurant} \mid \text{HasBar} = 0) = -\frac{1}{3} \log \frac{1}{3} - \frac{2}{3} \log \frac{2}{3} = 0.9182958341$

So,

$$\begin{aligned} H(\text{IsGoodRestaurant} \mid \text{HasBar}) &= P(\text{HasBar} = 1) H(\text{IsGoodRestaurant} \mid \text{HasBar} = 1) \\ &\quad + P(\text{HasBar} = 0) H(\text{IsGoodRestaurant} \mid \text{HasBar} = 0) \\ &= \frac{1}{4} \times 0 + \frac{3}{4} \times 0.9182958341 \\ &= \underline{0.6887218756}. \end{aligned}$$

$H(\text{IsGoodRestaurant} \mid \text{HasGoodAtmosphere})$:

- $P(\text{HasGoodAtmosphere} = 1) = \frac{1}{4}$ and $P(\text{HasGoodAtmosphere} = 0) = \frac{3}{4}$
- $H(\text{IsGoodRestaurant} \mid \text{HasBar} = 1) = -1 \log 1 - 0 \log 0 = 0$
- $H(\text{IsGoodRestaurant} \mid \text{HasBar} = 0) = -0 \log 0 - 1 \log 1 = 0$

So,

$$\begin{aligned} H(\text{IsGoodRestaurant} \mid \text{HasBar}) &= P(\text{HasBar} = 1) H(\text{IsGoodRestaurant} \mid \text{HasBar} = 1) \\ &\quad + P(\text{HasBar} = 0) H(\text{IsGoodRestaurant} \mid \text{HasBar} = 0) \\ &= \frac{1}{4} \times 0 + \frac{3}{4} \times 0 \\ &= \underline{0}. \end{aligned}$$

Now, using the conditional entropies above, we can compute the information gains:

$$\begin{aligned} I(\text{IsGoodRestaurant}; \text{HasOutdoorSeating}) &= H(\text{IsGoodRestaurant}) - H(\text{IsGoodRestaurant} \mid \text{HasOutdoorSeating}) \\ &= 0.8112781245 - 0.6887218756 \\ &= \underline{0.1225562489} \end{aligned}$$

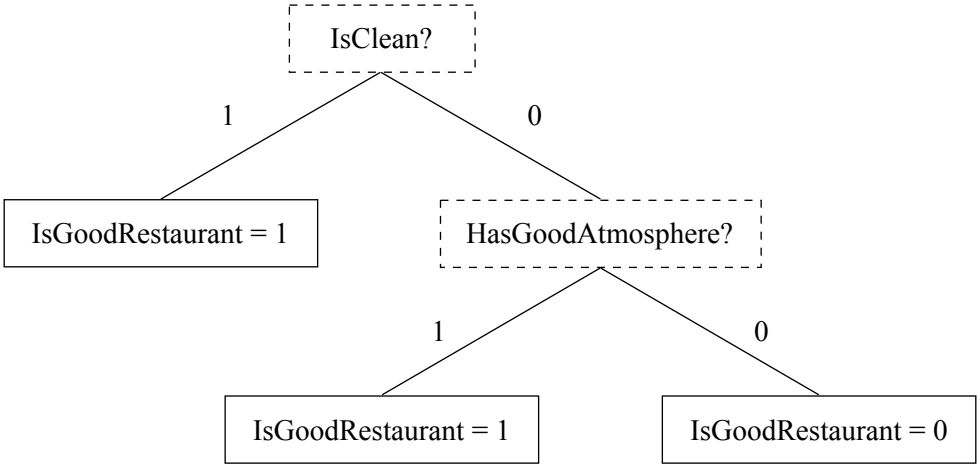
$$\begin{aligned} I(\text{IsGoodRestaurant}; \text{HasBar}) &= H(\text{IsGoodRestaurant}) - H(\text{IsGoodRestaurant} \mid \text{HasBar}) \\ &= 0.8112781245 - 0.6887218756 \\ &= \underline{0.1225562489} \end{aligned}$$

$$\begin{aligned} I(\text{IsGoodRestaurant}; \text{HasGoodAtmosphere}) &= H(\text{IsGoodRestaurant}) - H(\text{IsGoodRestaurant} \mid \text{HasGoodAtmosphere}) \\ &= 0.8112781245 - 0 \\ &= \underline{0.8112781245} \end{aligned}$$

So we split D_2 by HasGoodAtmosphere.

Notice that $H(\text{IsGoodRestaurant} \mid \text{HasBar} = 1) = H(\text{IsGoodRestaurant} \mid \text{HasBar} = 0) = 0$ on D_2, we do not farther split the tree.

Since there is no branch to split, we have constructed a decision tree, which looks like the following:



(g) Based on the tree above, both sample# 9, 10 are good restaurant.

2. Proof:

In order to show that $J(w_0, w_1) = \sum_{n=1}^{N} \alpha_n (w_0 + w_1 x_{n,1} - y_n)^2$ has a global optimal solution, it suffices to show the Hessian matrix of J is positive semi-definite.

Let

$$\mathbf{d} = \begin{pmatrix} \frac{\partial}{\partial w_0} \\ \frac{\partial}{\partial w_1} \end{pmatrix}.$$

Then,

$$\mathbf{d}\mathbf{d}^T = \begin{pmatrix} \frac{\partial^2}{\partial w_0^2} & \frac{\partial^2}{\partial w_0 \partial w_1} \\ \frac{\partial^2}{\partial w_1 \partial w_0} & \frac{\partial^2}{\partial w_1^2} \end{pmatrix}.$$

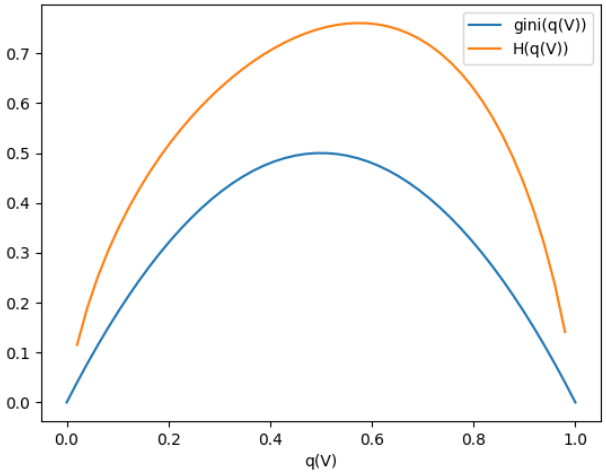
But since $\frac{\partial^2}{\partial w_0 \partial w_1} = \frac{\partial^2}{\partial w_1 \partial w_0}$, $\mathbf{d}\mathbf{d}^T$ is a symmetric matrix.

Notice that we can write the Hessian matrix of J in terms of $\mathbf{d}\mathbf{d}^T$. Namely, $H = \mathbf{d}\mathbf{d}^T J$. Since J is a scalar and $\mathbf{d}\mathbf{d}^T$ is a symmetric matrix, H of J is also a symmetric matrix. So, it's enough to show that $\mathbf{z}H\mathbf{z}^T = \mathbf{z}\mathbf{d}\mathbf{d}^T J\mathbf{z}^T$ is nonnegative for every non-zero column vector $\mathbf{z}$. Now using properties of matrix multiplication, we can rewrite $\mathbf{z}H\mathbf{z}^T$ as follows:

$$\begin{aligned} \mathbf{z}^T H\mathbf{z} &= \mathbf{z}^T \mathbf{d}\mathbf{d}^T J\mathbf{z} \\ &= \mathbf{z}^T \mathbf{d}\mathbf{d}^T \mathbf{z} J && \text{we can move scalar } J \\ &= (\mathbf{d}^T\mathbf{z})^T (\mathbf{d}^T\mathbf{z}) J && \text{matrix multiplication is associative} \end{aligned}$$

Notice that $J \geq 0$ because $\alpha_n > 0$ and $(w_0 + w_1 x_{n,1} - y_n)^2 \geq 0$ for any n.
Also, by the positivity axiom of inner product, namely $\mathbf{v}^T\mathbf{v} \geq 0$ for any $\mathbf{v}$, $(\mathbf{d}^T\mathbf{z})^T(\mathbf{d}^T\mathbf{z}) \geq 0$.
Thus, $\mathbf{z}^T H\mathbf{z} = (\mathbf{d}^T\mathbf{z})^T(\mathbf{d}^T\mathbf{z})J \geq 0$ for any non-zero column vector $\mathbf{z} \Longrightarrow$ the Hessian matrix of J is positive semi-definite. Therefore, J has a global optimal solution. ⍁

3.
(a) The plot is shown below:



The shape of gini($q(V)$) and $H(q(V))$ are similar in the sense that both functions are concave and have local maximum around $q(V) = 0.5$ (i.e. the middle point).

(b) Proof:
Given $i(q(V))$ is concave, we want to show that

$$I(V_1, V_2, V) = i(q(V)) - \big(p(V_1, V)i(q(V_1)) + p(V_2, V)i(q(V_2))\big) \geq 0$$

for all $V_1, V_2 \subseteq V$ that forms a partition of V.

First, we evaluate $i(q(V))$:
By using the definition of $q(.)$, we can rewrite $q(V)$ as follows:

$$\begin{aligned}
q(V) &= \frac{|\{i : i \in V, y_i = 1\}|}{|V|} \\
&= \frac{|\{i : i \in V_1, y_i = 1\}| + |\{i : i \in V_2, y_i = 1\}|}{|V|} && \text{because } V_1, V_2 \text{ form a partition of } V \\
&= \frac{|V_1|}{|V_1|}\frac{|\{i : i \in V_1, y_i = 1\}|}{|V|} + \frac{|V_2|}{|V_2|}\frac{|\{i : i \in V_2, y_i = 1\}|}{|V|} \\
&= \frac{|V_1|}{|V|}\frac{|\{i : i \in V_1, y_i = 1\}|}{|V_1|} + \frac{|V_2|}{|V|}\frac{|\{i : i \in V_2, y_i = 1\}|}{|V_2|} \\
&= p(V_1, V)q(V_1) + p(V_2, V)q(V_2) && \text{by definition of } p \text{ and } q
\end{aligned}$$

Since since V is finite and V_1, V_2 form a partition of V, $|V_1| + |V_2| = |V|$.
In particular, $p(V_1, V) = 1 - p(V_2, V)$ and $0 \leq p(V_1, V) \leq 1$.

Let $\lambda = p(V_1, V)$, $q_1 = q(V_1)$, and $q_2 = q(V_2)$. Notice that because $0 \le p(V_1, V) \le 1,\ \lambda \in [0,1]$.
Then, we can rewrite $i(q(V))$ as:

$$i(q(V)) = i(\lambda q_1 + (1-\lambda)q_2) \qquad -(1).$$

Now we evaluate $p(V_1, V)i(q(V_1)) + p(V_2, V)i(q(V_2))$:
Using λ, q_1, q_2 defined above, we can write it as:

$$p(V_1, V)i(q(V_1)) + p(V_2, V)i(q(V_2)) = \lambda i(q_1) + (1-\lambda)i(q_2) \qquad -(2).$$

By putting (1) and (2) together, we have

$$I(V_1, V_2, V) = i(\lambda q_1 + (1-\lambda)q_2) - \big(\lambda i(q_1) + (1-\lambda)i(q_2)\big).$$

But since $i(q(V))$ is concave, $i(\lambda q_1 + (1-\lambda)q_2) \ge \lambda i(q_1) + (1-\lambda)i(q_2)$.
Hence, $I(V_1, V_2, V) = i(\lambda q_1 + (1-\lambda)q_2) - \big(\lambda i(q_1) + (1-\lambda)i(q_2)\big) \ge 0 \implies I(V_1, V_2, V) \ge 0.$ ◪

(c)

$$\begin{aligned}\frac{d^2}{dx^2}H(x) &= \frac{1}{\ln 2}\frac{d^2}{dx^2}\big(-x\ln x - (1-x)\ln(1-x)\big)\\ &= \frac{1}{\ln 2}\frac{d}{dx}\big(-1 - \ln x + 1 + \ln(1-x)\big)\\ &= \frac{1}{\ln 2}\left(-\frac{1}{x} + \frac{1}{1-x}\right)\\ &= \frac{1}{\ln 2\, x(x-1)}\end{aligned}$$

Thus, $\frac{d^2H(x)}{dx^2} \le 0$ for $x \in [0,1]$. But since $q(V) \in [0,1]$, $\frac{d^2H(q(x))}{dx^2} \le 0$ for all $q(V)$.Hence, $H(q(V))$ is concave.

(d) $\frac{d^2}{dx^2}\text{gini}(x) = \frac{d}{dx}(-4x + 2) = -4 \le 0$. Thus, $\text{gini}(q(V))$ is concave.

4.

(a) The average error for each *k* is shown below:

k	Example 1	Example 2
1	10/14	2/14
3	8/14	4/14
5	4/14	6/14
7	4/14	12/14

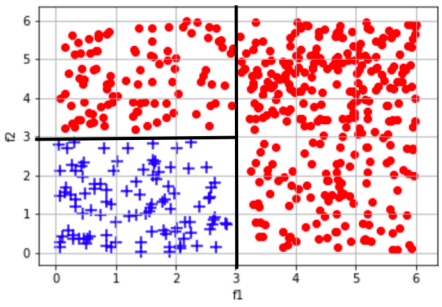
On Example1, $k = 5,7$ minimized the error with error = 2/7 and on Example2, $k = 1$ minimizes the error with error = 1/7.

(b) When $k = 1$, error = 10/14. When $k = 14$, the error = 1.
In general, using too small k causes an overfitting because the model becomes very sensitive to how the train datapoint are distributed. Also, if k is too small, outliers play a role in assigning a label for an unseen datapoint, which is not desirable. On the other hand, when k is too big, the model is biased towards the majority of the training dataset. So, for example, if you use $k = 13$ for Example1, the leave-out validation gives an error for each datapoint because majority becomes the opposite of the datapoint in interest. Hence, it is important to tune k to obtain an optimal model.

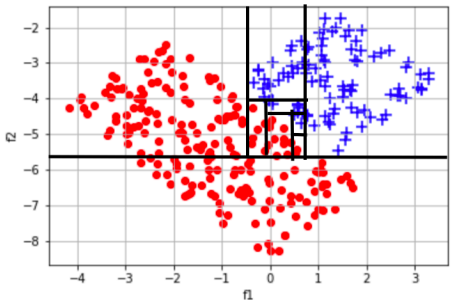
5.

(a) Example1 and Example3. For Example1, you can ask $f_1 < 3$ first and then $f_2 > 3$ for those points whose answer to $f_1 < 3$ is yes. For Example2, first you can ask $f_2 > 3$. If yes, then ask $f_1 < 4$. If no, then ask $f_1 < 2$.

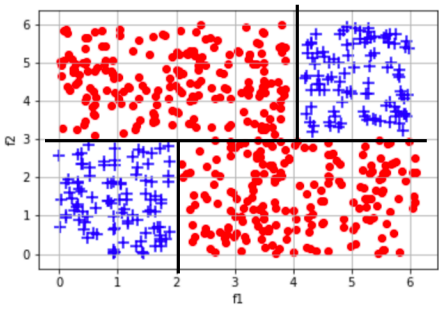
(b) Example2
Because the decision boundary for Example2 has a slope, dividing the data points by horizontal or vertical line would take long time (i.e. the decision tree becomes deep). Other Examples are relatively easy to separate by horizontal or vertical lines. Sample decision boundaries are shown below:



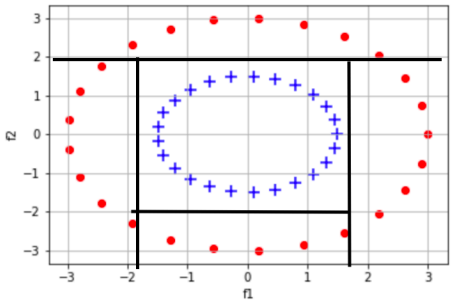
(1) Decision Tree Example 1



(2) Decision Tree Example 2

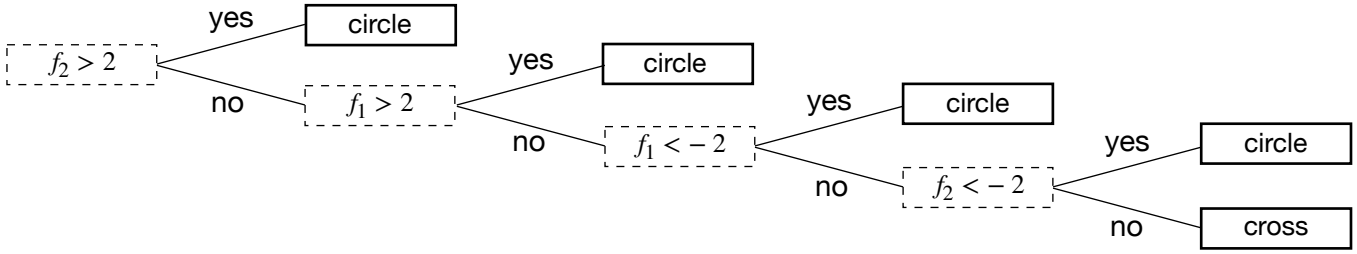


(3) Decision Tree Example 3



(4) Decision Tree Example 4

(c) Yes, Example4 is separable by a depth 4 decision tree. A sample tree is shown below:



6.

(a)

```
Accuracy on the training set:  0.6154929577464788
Accuracy on the test set:  0.6101694915254238
```

(b)

```
Accuracy on the training set:  0.9859154929577465
Accuracy on the test set:  0.7005649717514124
```

(c) $k = 1$:

```
Accuracy on the training set:  0.9845070422535211
Accuracy on the test set:  0.7288135593220338
```

$k = 3$:

```
Accuracy on the training set:  0.8971830985915493
Accuracy on the test set:  0.768361581920904
```

$k = 5$:

```
Accuracy on the training set:  0.8690140845070422
Accuracy on the test set:  0.7796610169491526
```

(d) Decision Tree:

```
Cross validation accuracy (10fold): 0.8018013637379834
```

KNN:

$k = 1$:

```
Cross validation accuracy (10fold): 0.774896042924212
```

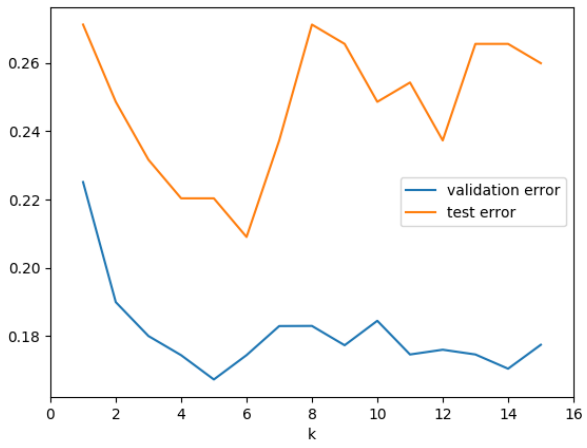
$k = 3$:

```
Cross validation accuracy (10fold): 0.8199524927341828
```

$k = 5$:

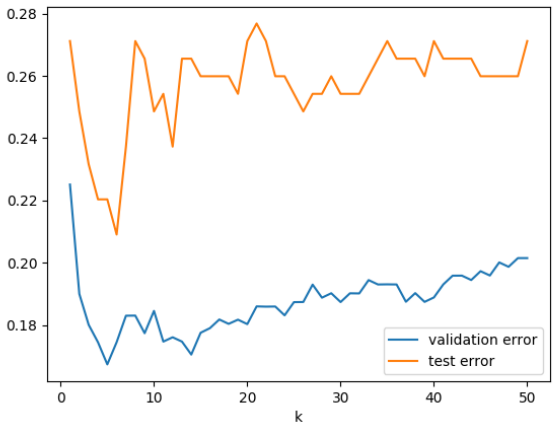
```
Cross validation accuracy (10fold): 0.8326699083389224
```

(e)



k	Validation Error	Test Error
1	0.225103957075788	0.2711864406779660
2	0.1899670243684330	0.248587570621469
3	0.1800475072658170	0.231638418079096
4	0.1744321484462330	0.2203389830508470
5	0.1673300916610780	0.2203389830508470
6	0.1744533869885980	0.2090395480225990
7	0.182963894477979	0.2372881355932200
8	0.1830035770176610	0.2711864406779660
9	0.1773485356583950	0.2655367231638420
10	0.1845098367985690	0.248587570621469
11	0.1746490051419630	0.2542372881355930
12	0.1760384529398610	0.2372881355932200
13	0.1746495640509720	0.2655367231638420
14	0.1704828973843060	0.2655367231638420
15	0.1774865861837690	0.2598870056497180

The data shows that our knn classifier achieves the smallest error on the test dataset when $k = 6$.
The validation errors roughly decreases as k increases. On the other hand, the test error decreases as k increases only up to around $k = 6$, after which the test error increases and somewhat stays sound 0.25 for the rest of ks.



In order to understand how k relates to the test and validation errors, I have tested for even larger ks (the plot is shown on the left). The validation error seems to have a certain relationship with k, namely, it decreases rapidly at the beginning and after a certain point, it increasing gradually. The test error also decreases at the beginning; however, it looks somewhat randomly taking values between 0.24 - 0.28 after a certain k.

To summarize, it seems to me that both test and validation errors are high for very small k (e.g. $k = 1$) because the model is overfitted. So, by increasing k, the model becomes less overfitted and can achieve the best prediction task at a certain k (e.g. $k = 6$). However, if you increases k even further, the model now becomes not optimal, resulting higher errors.

(f) I would choose knn classifier because knn with optimal k can achieve better accuracy than that of a decision tree model.