# CSCI 567: Theory Assignment 1

Anurima Anil Padwal
USC ID: 4348819703

September 2019

## Solutions

1. For $\mathbf{x_i}, \mathbf{x_o}, \mathbf{x_j}$ normalized to the unit norm, $\|\mathbf{x_i}\|_2 = \|\mathbf{x_o}\|_2 = \|\mathbf{x_j}\|_2 = 1$. The cosine distance between $\mathbf{x_i}$ and $\mathbf{x_j}$ is

$$\begin{aligned} C(\mathbf{x_i}, \mathbf{x_j}) &= 1 - \frac{\mathbf{x_i^T} \cdot \mathbf{x_j}}{\|\mathbf{x_i}\|_2 \|\mathbf{x_j}\|_2} \\ &= 1 - \mathbf{x_i^T} \cdot \mathbf{x_j} \qquad (1) \end{aligned}$$

Similarly, the cosine distance between $\mathbf{x_i}$ and $\mathbf{x_o}$

$$\begin{aligned} C(\mathbf{x_i}, \mathbf{x_o}) &= 1 - \frac{\mathbf{x_i^T} \cdot \mathbf{x_o}}{\|\mathbf{x_i}\|_2 \|\mathbf{x_o}\|_2} \\ &= 1 - \mathbf{x_i^T} \cdot \mathbf{x_o} \qquad (2) \end{aligned}$$

Euclidian distance between $\mathbf{x_i}$ and $\mathbf{x_j}$ is given by:

$$\begin{aligned} E(\mathbf{x_i}, \mathbf{x_j}) &= \|\mathbf{x_i} - \mathbf{x_j}\|_2^2 \\ &= (\mathbf{x_i} - \mathbf{x_j})^T \cdot (\mathbf{x_i} - \mathbf{x_j}) \\ &= \mathbf{x_i}^T \mathbf{x_j} - 2 \cdot \mathbf{x_i}^T \mathbf{x_j}^T + \mathbf{x_j}^T \mathbf{x_j} \\ &= 1 + 1 - 2 \cdot \mathbf{x_i}^T \mathbf{x_j} \\ &= 2 - 2 \cdot \mathbf{x_i}^T \mathbf{x_j} \\ &= 2(1 - \mathbf{x_i}^T \mathbf{x_j}) \qquad (3) \end{aligned}$$

Similarly, Euclidian distance between $\mathbf{x_i}$ and $\mathbf{x_o}$ is given by:

$$\begin{aligned} E(\mathbf{x_i}, \mathbf{x_o}) &= \|\mathbf{x_i} - \mathbf{x_o}\|_2^2 \\ &= (\mathbf{x_i} - \mathbf{x_o})^T \cdot (\mathbf{x_i} - \mathbf{x_o}) \\ &= \mathbf{x_i}^T \mathbf{x_o} - 2 \cdot \mathbf{x_i}^T \mathbf{x_o}^T + \mathbf{x_j}^T \mathbf{x_o} \\ &= 1 + 1 - 2 \cdot \mathbf{x_i}^T \mathbf{x_o} \\ &= 2 - 2 \cdot \mathbf{x_i}^T \mathbf{x_o} \\ &= 2(1 - \mathbf{x_i}^T \mathbf{x_o}) \qquad (4) \end{aligned}$$

Given,

$$C(\mathbf{x_i} \cdot \mathbf{x_j}) <= C(\mathbf{x_i} \cdot \mathbf{x_o})$$

`Multiplying on both sides by 2,`

$$2C(\mathbf{x_i} \cdot \mathbf{x_j}) <= 2C(\mathbf{x_i} \cdot \mathbf{x_o})$$

`From (1) and (2),`

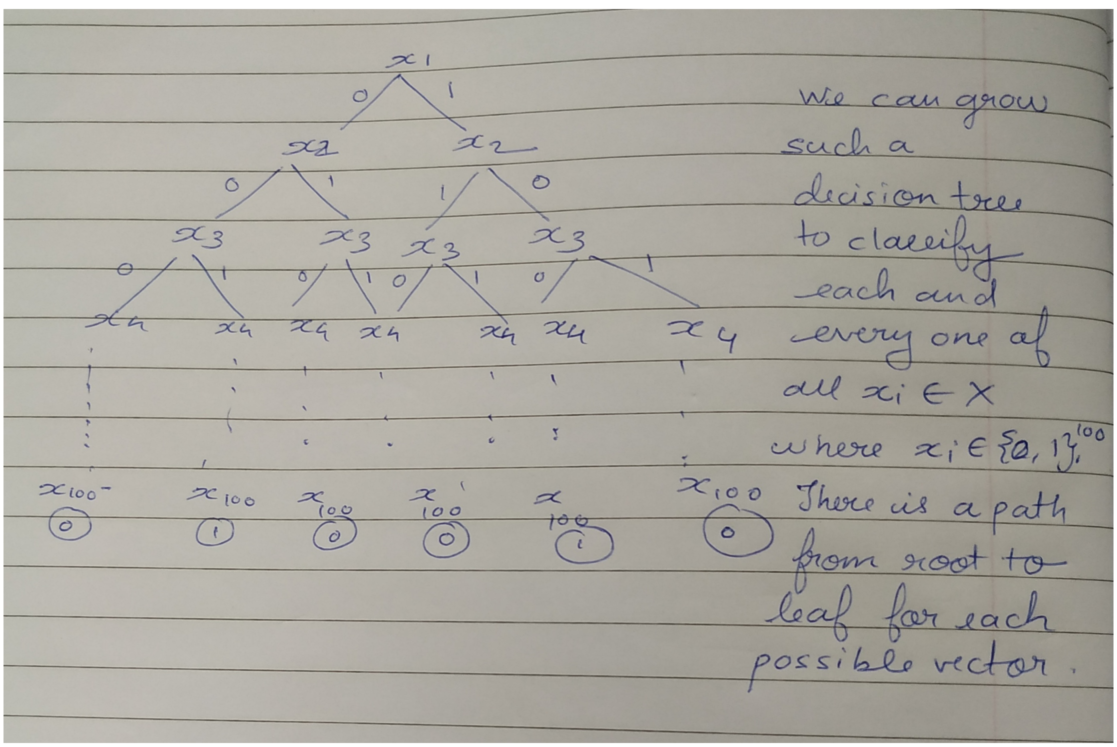
$$2(1 - \mathbf{x_i^T} \cdot \mathbf{x_j}) <= 2(1 - \mathbf{x_i^T} \cdot \mathbf{x_o})$$

`From (3) and (4),`

$$E(\mathbf{x_i}, \mathbf{x_j}) <= E(\mathbf{x_i}, \mathbf{x_o})$$

Hence, proved.

2.1 Yes, we can have a decision tree to classify a dataset of 100 dimensional vectors, such that each vector is a binary vector $\mathbf{x} \in \{0,1\}^{100}$, and the tree has zero classification error. First we generate all possible combinations for each such $\mathbf{x_i}$. There are $2^{100}$ such possibilities since the dataset has binary vectors of 100 dimensions. Here, we split on any feature in the vector for the root node, and continue building the tree top down, such that we have a path from root to leaf node for each $\mathbf{x_i}$.



2.2 Yes. First we generate all possible combinations for each such $\mathbf{x_i}$. There are $2^{100}$ such possibilities since the dataset has binary vectors of 100 dimensions. We can classify this entire set using our decision tree, and use the resulting set of vectors and labels as our 1-NN classifier. This will give the same result as the decision tree since each point is its own neighbor.

3. Yes, the given decision tree can be implemented as a 1 NN classifier. The values are as follows:

| $(x_1, x_2)$ | Label |
|---|---|
| (A + 1, B + 1) | 1 |
| (A + 1, B - 1) | 0 |
| (A - 1, B + 1) | 0 |
| (A - 1, B - 1) | 1 |

4.1 Based on the decision tree in Figure 3, the number of mis-classifications is 0. Hence the test error $= \frac{\texttt{number of misclassifications}}{\texttt{number of total test samples}} = \frac{0}{2} = 0$

4.2 Based on the decision tree in Figure 4, the number of misclassifications on the test data is 1. The point $(x_1 = 0.2, x_2 = 0.8)$ is misclassified. Hence the test error $= \frac{\texttt{number of misclassifications}}{\texttt{number of total test samples}} = \frac{1}{2} = 0.5$

4.3 Yes, the decision tree in Figure 4 is a linear classifier in terms of $(x_1, x_2)$. The classifier may be represented as y = 1 when z > 0.5 and y = 0 for all other cases. Here z represents the weighted combination $z = w_1 \cdot x_1 + w_2 \cdot x_2$ where $w_2 = 0$ and $w_1 \geq 1$. No, we cannot classify the data and get zero classification error by drawing a depth 1 decision tree.

4.4 No.
Justification:
Putting the data in the rectangle:

$$ax_1 + bx_2 \geq c$$

$$ax_1 + bx_2 < c$$

we get,

$$a \geq 1 \qquad (1)$$

$$b \geq 1 \qquad (2)$$

$$0 < 1 \qquad (3)$$

$$a + b < 1 \qquad (4)$$

Equations (1), (2), and (4) are contradictory. Hence, the data in Table 1 is not linearly separable. Hence, we cannot classify the data using a depth 1 decision tree to get a zero classification error.

5. The structure of $T_1$ is:
Left child:
class A: 150 samples
class B: 50 samples
Label: class A

Right child:
class A: 50 samples
class B: 150 samples
Label: class B

The structure of $T_2$ is:
Left child:
class A: 0 samples
class B: 100 samples
Label: class B

Right child:
class A: 200 samples
class B: 100 samples
Label: class A

5.1 For tree $T_1$,
Left child:
50 samples are misclassified as class B.
Classification error $= \frac{50}{200} = 0.25$

Right child:
50 samples are misclassified as class A.
Classification error $= \frac{50}{200} = 0.25$

For tree $T_2$,
Left child:
0 samples are misclassified
Classification error $= \frac{0}{100} = 0$

Right child:
100 samples are misclassified as class B.
Classification error $= \frac{100}{300} = 0.33$

Entropy is given by,

$$H(P) = -\sum_{k=1}^{C} P(Y = k) \log P(Y = k)$$

For tree $T_1$,

Left child:

$$\begin{aligned}
\texttt{Entropy} &= -\frac{150}{150+50}\log_e\frac{150}{150+50} - \frac{50}{150+50}\log_e\frac{50}{150+50} \\
&= -\frac{150}{200}\log_e\frac{150}{200} - \frac{50}{200}\log_e\frac{50}{200} \\
&= 0.56
\end{aligned}$$

Right child:

$$\begin{aligned}
\texttt{Entropy} &= -\frac{150}{150+50}\log_e\frac{150}{150+50} - \frac{50}{150+50}\log_e\frac{50}{150+50} \\
&= -\frac{150}{200}\log_e\frac{150}{200} - \frac{50}{200}\log_e\frac{50}{200} \\
&= 0.56
\end{aligned}$$

For tree $T_2$,
Left child:

$$\begin{aligned}
\texttt{Entropy} &= -\frac{0}{100+0}\log_e\frac{0}{100+0} - \frac{100}{100+0}\log_e\frac{100}{100+0} \\
&= 0
\end{aligned}$$

Right child:

$$\begin{aligned}
\texttt{Entropy} &= -\frac{200}{200+100}\log_e\frac{200}{200+100} - \frac{100}{200+100}\log_e\frac{100}{200+100} \\
&= -\frac{200}{300}\log_e\frac{200}{300} - \frac{100}{300}\log_e\frac{100}{300} \\
&= 0.64
\end{aligned}$$

Gini Impurity
For tree $T_1$,
Left child,

$$\begin{aligned}
\texttt{Gini Impurity} &= 1 - \sum p_i^2 \\
&= 1 - (\frac{150}{200})^2 - (\frac{50}{200})^2 \\
&= 0.38
\end{aligned}$$

Right child,

$$\begin{aligned}
\texttt{Gini Impurity} &= 1 - \sum p_i^2 \\
&= 1 - (\frac{150}{200})^2 - (\frac{50}{200})^2 \\
&= 0.38
\end{aligned}$$

For tree $T_2$, Left child,

$$\begin{aligned}\texttt{Gini Impurity} &= 1 - \sum p_i^2 \\ &= 1 - (\frac{0}{100})^2 - (\frac{100}{100})^2 \\ &= 0\end{aligned}$$

Right child,

$$\begin{aligned}\texttt{Gini Impurity} &= 1 - \sum p_i^2 \\ &= 1 - (\frac{200}{300})^2 - (\frac{100}{300})^2 \\ &= 0.44\end{aligned}$$

5.2 $T_1$ misclassifies 100 samples in all and $T_2$ also misclassifies 100 samples in all. Hence we cannot say one is better than the other in terms of classification error as it is the same of both, i.e. 0.25.

The conditional entropy of $T_1$ is:

$$\begin{aligned}&H_1 * P_1 + H_2 * P_2 \\ &= 0.56 * \frac{200}{400} + 0.56 * \frac{200}{400} = 0.56\end{aligned}$$

The conditional entropy of $T_2$ is:

$$\begin{aligned}&H_1 * P_1 + H_2 * P_2 \\ &= 0 * \frac{100}{400} + 0.63 * \frac{300}{400} = 0.48\end{aligned}$$

The conditional entropy of $T_2$ is less than that of $T_1$. Hence, the split at $T_2$ is more pure than at $T_1$.

Now, weighted Gini for $T_1 = \frac{200}{400} * 0.37 + \frac{200}{400} * 0.37 = 0.37$
Weighted Gini for $T_2 = 0 * \frac{100}{400} + 0.44 * \frac{300}{400} = 0.33$
Weighted Gini of $T_2$ is less than that of $T_1$. Hence $T_2$ is more pure and has a better quality of split than $T_1$.

6.1. P(PlayTennis = Yes) $= \frac{4}{6} = \frac{2}{3}$ P(PlayTennis = No) $= \frac{2}{6} = \frac{1}{3}$

6.2 P(Weather = Sunny | PlayTennis = Yes) $= \frac{2}{4} = \frac{1}{2}$
P(Emotion = Normal | PlayTennis = Yes) $= \frac{1}{4}$

P(HomeWork = Much | PlayTennis = Yes) = $\frac{1}{4}$

6.3

$$
\begin{aligned}
&P(\texttt{PlayTennis = Yes | x}) \\
&= \frac{P(\texttt{x | PlayTennis = Yes}) * P(\texttt{PlayTennis = Yes})}{P(x)} \\
&= \frac{P(\frac{\texttt{Weather = Sunny}}{Yes}) * P(\frac{\texttt{Emotion = Normal}}{Yes}) * P(\frac{\texttt{Homework = Much}}{Yes})}{P(x)} \\
&= \frac{\frac{1}{2} * \frac{1}{4} * \frac{1}{4} * \frac{2}{3}}{P(x)} \\
&= \frac{1/48}{P(x)} \\
&P(\texttt{PlayTennis = No | x}) \\
&= \frac{P(\texttt{x | PlayTennis = No}) * P(\texttt{PlayTennis = No})}{P(x)} \\
&= \frac{P(\frac{\texttt{Weather = Sunny}}{No}) * P(\frac{\texttt{Emotion = Normal}}{No}) * P(\frac{\texttt{Homework = Much}}{No})}{P(x)} \\
&= \frac{\frac{1}{2} * \frac{1}{2} * \frac{1}{2} * \frac{1}{3}}{P(x)} \\
&= \frac{1/24}{P(x)}
\end{aligned}
$$

The numerator of $P(\texttt{PlayTennis = No}|x)$ is greater than that of $P(\texttt{PlayTennis = Yes}|x)$. Hence $P(\texttt{PlayTennis = No}|x)$ has a larger value.