Name: Yuhan Shao
Student Number: 1002281327
UTORid: shaoyuha

CSC411 Assignment 1

1. (a) • X and Y are two independent univariate random variables sampled uniformly from $[0,1]$
• Therefore, we have the following:

$$\begin{aligned}
E[z] &= E[(x-y)^2] \\
&= E[x^2 - 2xy + y^2] \\
&= E[x^2] - 2E[xy] + E[y^2] && \text{\# By the linearity of expectation (*)} \\
&= E[x^2] - 2E[x]E[y] + E[y^2] && \text{\# since } X, Y \text{ are independents. (**)}
\end{aligned}$$

Since:
$E[x^2] = \int_0^1 x^2 dx = \frac{1}{3}x^3\Big|_0^1 = \frac{1}{3}$
$E[y^2] = \int_0^1 y^2 dy = \frac{1}{3}y^3\Big|_0^1 = \frac{1}{3}$
$E[x] = \int_0^1 x\,dx = \frac{1}{2}x^2\Big|_0^1 = \frac{1}{2}$
$E[y] = \int_0^1 y\,dy = \frac{1}{2}y^2\Big|_0^1 = \frac{1}{2}$

Therefore, $E[z] = E[x^2] - 2E[x]E[y] + E[y^2] = \frac{1}{3} - 2\times\frac{1}{2}\times\frac{1}{2} + \frac{1}{3} = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}$

$$\begin{aligned}
V[z] &= E[z^2] - (E[z])^2 \\
&= E[(x-y)^4] - \tfrac{1}{36} \\
&= E[x^4 - 4x^3y + 6x^2y^2 - 4xy^3 + y^4] - \tfrac{1}{36}. \\
&= E[x^4] - 4E[x^3]E[y] + 6E[x^2]E[y^2] - 4E[x]E[y^3] + E[y^4] - \tfrac{1}{36} && \text{\# By (*) and (**)}
\end{aligned}$$

Since, $E[x^3] = \int_0^1 x^3 dx = \frac{1}{4}x^4\Big|_0^1 = \frac{1}{4}$, similarly $E[y^3] = \frac{1}{4}$
$E[x^4] = \int_0^1 x^4 dx = \frac{1}{5}x^5\Big|_0^1 = \frac{1}{5}$, similarly $E[y^4] = \frac{1}{5}$.

Therefore, $V[z] = E[x^4] - 4E[x^3]E[y] + 6E[x^2]E[y^2] - 4E[x]E[y^3] + E[y^4] - \frac{1}{36}$
$= \frac{1}{5} - 4\times\frac{1}{4}\times\frac{1}{2} + 6\times\frac{1}{3}\times\frac{1}{3} - 4\times\frac{1}{2}\times\frac{1}{4} + \frac{1}{5} - \frac{1}{36}$
$= \frac{1}{5} - \frac{1}{2} + \frac{2}{3} - \frac{1}{2} + \frac{1}{5} - \frac{1}{36}$
$= \frac{7}{180}$

• Therefore, $E[z] = 1/6$, $V[z] = 7/180$

(b) • Since random variables $X_1, X_2, \dots, X_d, Y_1, Y_2, \dots, Y_d$ independently from $[0,1]$, $Z_i = (X_i - Y_i)^2$ are also independently from $[0,1]$.

• $E[R] = E[Z_1 + Z_2 + \dots + Z_d]$
$= E[Z_1] + E[Z_2] + \dots + E[Z_d]$
$= \underbrace{\frac{1}{6} + \frac{1}{6} + \dots + \frac{1}{6}}_{d \text{ times}} = d/6$

• $V[R] = V[Z_1 + Z_2 + \dots + Z_d]$
$= V[Z_1] + V[Z_2] + \dots + V[Z_d]$ # since Z_i are independents to each other.
$= \underbrace{\frac{7}{180} + \frac{7}{180} + \dots + \frac{7}{180}}_{d \text{ times}} = 7d/180$

• Therefore, $E[R] = d/6$, $V[R] = 7d/180$

(C) • Suppose we sample two points $(X_1, X_2, \dots, X_d)$ and $(Y_1, Y_2, \dots, Y_d)$ independently from a unit circle, in the higher dimension d.
• Then, the maximum possible squared Euclidean distance is when one of the point is $(\underbrace{0,0,\dots,0}_{d \text{ times}})$ and one of the points is $(\underbrace{1,1,\dots,1}_{d \text{ times}})$.
• Therefore, $R = \sum_{i=1}^{d} Z_i = \sum_{i=1}^{d} (X_i - Y_i)^2 = \sum_{i=1}^{d} (1-0)^2 = d$.
• When in high dimensions, $d \to \infty$, we have:

$E[R] = d/6 \in O(d)$
$V[R] = 7d/180 \in O(d)$
$R_{max} = d$

$\Rightarrow$ Therefore, in high dimensions, most points are far away, and approximately the same distance.

2. (b)

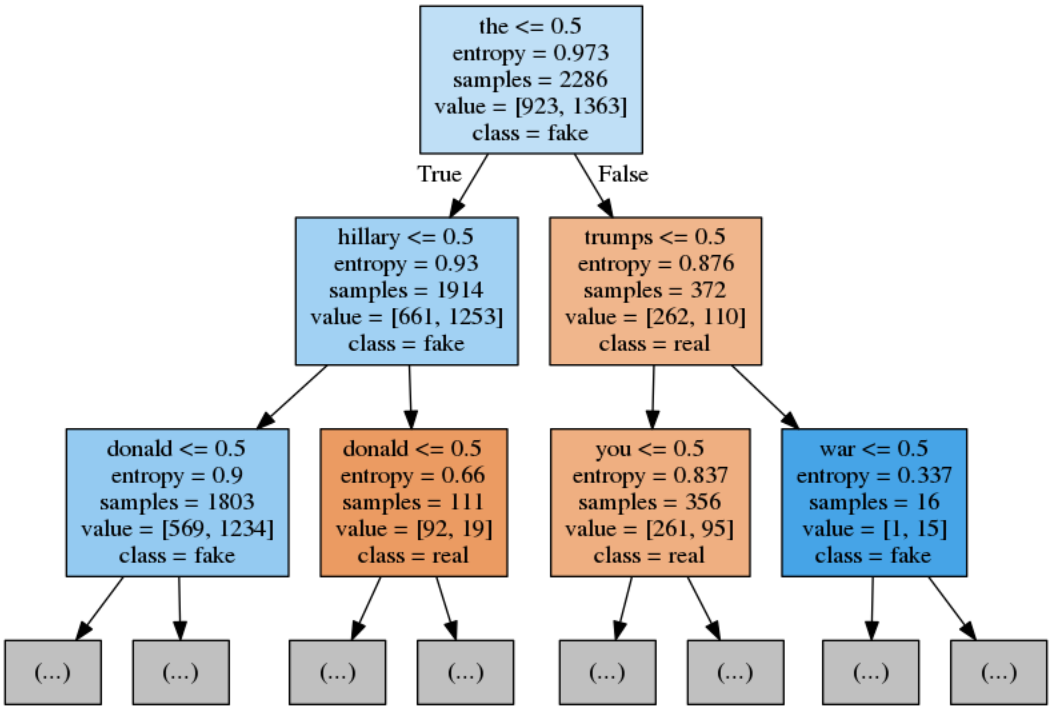
```
Decision tree: criterion=entropy, max_depth=5 has accuracy 0.6775510204081633.
Decision tree: criterion=gini, max_depth=5 has accuracy 0.673469387755102.
Decision tree: criterion=entropy, max_depth=10 has accuracy 0.7020408163265306.
Decision tree: criterion=gini, max_depth=10 has accuracy 0.7020408163265306.
Decision tree: criterion=entropy, max_depth=20 has accuracy 0.746938775510204.
Decision tree: criterion=gini, max_depth=20 has accuracy 0.7346938775510204.
Decision tree: criterion=entropy, max_depth=50 has accuracy 0.7693877551020408.
Decision tree: criterion=gini, max_depth=50 has accuracy 0.7510204081632653.
Decision tree: criterion=entropy, max_depth=100 has accuracy 0.7755102040816326.
Decision tree: criterion=gini, max_depth=100 has accuracy 0.7428571428571429.

Best decision tree: criterion=entropy, max_depth=100 has accuracy 0.7755102040816326.
```

(c) Text form of extraction and visualization of the first two layers of the tree:

```
digraph Tree {
node [shape=box, style="filled", color="black"] ;
0 [label="the <= 0.5\nentropy = 0.973\nsamples = 2286\nvalue = [923, 1363]\nclass = fake", fillcolor="#399de552"] ;
1 [label="hillary <= 0.5\nentropy = 0.93\nsamples = 1914\nvalue = [661, 1253]\nclass = fake", fillcolor="#399de578"] ;
0 -> 1 [labeldistance=2.5, labelangle=45, headlabel="True"] ;
2 [label="donald <= 0.5\nentropy = 0.9\nsamples = 1803\nvalue = [569, 1234]\nclass = fake", fillcolor="#399de589"] ;
1 -> 2 ;
3 [label="(...)", fillcolor="#C0C0C0"] ;
2 -> 3 ;
396 [label="(...)", fillcolor="#C0C0C0"] ;
2 -> 396 ;
555 [label="donald <= 0.5\nentropy = 0.66\nsamples = 111\nvalue = [92, 19]\nclass = real", fillcolor="#e58139ca"] ;
1 -> 555 ;
556 [label="(...)", fillcolor="#C0C0C0"] ;
555 -> 556 ;
565 [label="(...)", fillcolor="#C0C0C0"] ;
555 -> 565 ;
590 [label="trumps <= 0.5\nentropy = 0.876\nsamples = 372\nvalue = [262, 110]\nclass = real", fillcolor="#e5813994"] ;
0 -> 590 [labeldistance=2.5, labelangle=-45, headlabel="False"] ;
591 [label="you <= 0.5\nentropy = 0.837\nsamples = 356\nvalue = [261, 95]\nclass = real", fillcolor="#e58139a2"] ;
590 -> 591 ;
592 [label="(...)", fillcolor="#C0C0C0"] ;
591 -> 592 ;
757 [label="(...)", fillcolor="#C0C0C0"] ;
591 -> 757 ;
758 [label="war <= 0.5\nentropy = 0.337\nsamples = 16\nvalue = [1, 15]\nclass = fake", fillcolor="#399de5ee"] ;
590 -> 758 ;
759 [label="(...)", fillcolor="#C0C0C0"] ;
758 -> 759 ;
760 [label="(...)", fillcolor="#C0C0C0"] ;
758 -> 760 ;
}
```

Transfer .dot file to .png file with command: **dot -Tpng output.dot -o outfile.png**



(d)

```
The information gain on the split of word: the is 0.04706539495580031
The information gain on the split of word: hillary is 0.036271507210618226
The information gain on the split of word: donald is 0.0526295664461055
The information gain on the split of word: trumps is 0.04442019618695203
The information gain on the split of word: you is 0.01368249675448907
```