Homework 1

Soon Chee Loong
999295793
cheeloong.soon@mail.utoronto.ca
Last Name: Soon
First Name: Chee Loong
cdf Markus: soonchee

1. Nearest Neighbour & Curse of Dimensionality

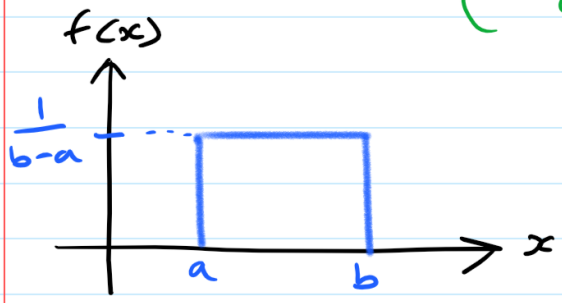
a) 2 independent univariate random variables X, Y uniformly sampled from interval $[0,1]$.

Determine the expectation & variance of the random variable Z.

$Z = (X-Y)^2$

Uniform Distribution $X \sim U(a,b)$

$$f(x) = \lim_{h \to 0} P(x < X \leq x+h) = \begin{cases} \frac{1}{b-a}, & x \in [a,b], \alpha < \beta, \beta > a, \alpha < b \\ 0, & \text{otherwise} \end{cases}$$



Moments of Uniform Distribution

$$E[x^n] = \int_{-\infty}^{\infty} x^n f(x)\,dx$$

$$= \int_{-\infty}^{a} x^n f(x)\,dx + \int_a^b x^n f(x)\,dx + \int_b^{\infty} x^n f(x)\,dx$$

$$= \int_a^b x^n \left(\frac{1}{b-a}\right) dx$$

$$= \left[\frac{x^{n+1}}{(n+1)(b-a)}\right]_a^b = \frac{b^{n+1}-a^{n+1}}{(n+1)(b-a)} = \frac{(1)^{n+1}-(0)^{n+1}}{(n+1)(1-0)}$$

$$= \frac{1}{(n+1)}, \quad a=0, b=1$$

Expectation of uniform distribution $X \sim U(0,1)$

$$E[X] = \frac{1}{2}$$

Variance of uniform distribution

$$\text{Var}(x) = E[x^2] - (E[x])^2 = \frac{1}{3} - \left(\frac{1}{2}\right)^2 = \frac{1}{12}$$

Expectation of random variable $Z = E[Z] = \frac{1}{6}$

$$E[Z] = E[(x-y)^2] = E[x^2 - 2xy + y^2]$$

$$= E[x^2] - 2E[xy] + E[y^2]$$

$$= E[x^2] - 2E[x]E[y] + E[y^2]$$

$$= \frac{1}{3} - 2\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) + \frac{1}{3}$$

Expectation of random variable Z $= E[Z] = \frac{1}{6}$

$E[Z] = E[(X-Y)^2] = E[X^2 - 2XY + Y^2]$

$= E[X^2] - 2E[XY] + E[Y^2]$

$= E[X^2] - 2E[X]E[Y] + E[Y^2]$

$= \frac{1}{3} - 2\left(\frac{1}{2}\right)\left(\frac{1}{2}\right) + \frac{1}{3}$

$= \frac{2}{3} - \frac{1}{2} = \frac{1}{6}$

Variance of random variable Z $= Var(Z) = \frac{7}{180}$

$Var(Z) = E[Z^2] - (E[Z])^2$

$E[Z^2] = E[((X-Y)^2)^2]$

$= E[(X-Y)^4]$

$= E[(X^4 - 4X^3Y + 6X^2Y^2 - 4XY^3 + Y^4)]$

$= E[X^4] - 4E[X^3]E[Y] + 6E[X^2]E[Y^2] - 4E[X]E[Y^3] + E[Y^4]$

$= \frac{1}{5} - 4\left(\frac{1}{4}\right)\left(\frac{1}{2}\right) + 6\left(\frac{1}{3}\right)\left(\frac{1}{3}\right) - 4\left(\frac{1}{2}\right)\left(\frac{1}{4}\right) + \frac{1}{5}$

$= \frac{1}{5} - \frac{1}{2} + \frac{2}{3} - \frac{1}{2} + \frac{1}{5}$

$= \frac{2}{5} - 1 + \frac{2}{3} = \frac{6}{15} - \frac{15}{15} + \frac{10}{15} = \frac{1}{15}$

$\therefore$

$Var(Z) = E[Z^2] - (E[Z])^2$

$= \frac{1}{15} - \left(\frac{1}{6}\right)^2 = \frac{12}{180} - \frac{5}{180} = \frac{7}{180}$

b) Suppose we sample 2 points independently from a unit cube in d dimensions.

Can view as sampling each coordinate independently as random variables

$X_1, X_2, \ldots, X_d$

$Y_1, Y_2, \ldots, Y_d$

independently from $[0,1]$.

$Z_i = (X_i - Y_i)^2$, $i \in \{1, 2, \ldots, d\}$

$R = \sum_{i=1}^{d} Z_i$, the total squared distance is simply the sum of each dimension.

Calculate $E[R]$, $Var(R)$ in terms of d, $E[Z]$, $Var(Z)$

$E[R] = E\left[\sum_{i=1}^{d} Z_i\right]$

$= \sum_{i=1}^{d} E[Z_i]$, since $\forall$ dimension is independent, and stationary

$= \sum_{i=1}^{d} E[Z]$

$= d\,E[Z]$

$\Rightarrow$ as $d \to \infty \Rightarrow E[R] \to \infty \Rightarrow$ expected distance increases to ∞

and will be very far away

b) Suppose we sample 2 points independently from a unit cube in d dimensions.
Can view as sampling each coordinate independently as random variables

$X_1, X_2, \ldots, X_d$

$Y_1, Y_2, \ldots, Y_d$

independently from $[0,1]$.

$Z_i = (X_i - Y_i)^2$, $i \in \{1, 2, \ldots, d\}$

$R = \sum_{i=1}^{d} Z_i$, the total squared distance is simply the sum of each dimension.

Calculate $E[R]$, $Var(R)$ in terms of d, $E[Z]$, $Var(Z)$

$E[R] = E[\sum_{i=1}^{d} Z_i]$

$= \sum_{i=1}^{d} E[Z_i]$, since $\forall$ dimension is independent, and stationary

$= \sum_{i=1}^{d} E[Z]$

$= d\,E[Z]$

$\Rightarrow$ as $d \to \infty \Rightarrow E[R] \to \infty \Rightarrow$ expected distance increases to ∞ and will be very far away.

$Var(R) = Var(\sum_{i=1}^{d} Z_i)$

$= \sum_{i=1}^{d} Var(Z_i)$, since $\forall$ dimension is independent

$\Rightarrow Cov(Z_i, Z_j) = 0 \quad \forall\, i \neq j,\ i, j \in \{1, 2, \ldots, d\}$

$= \sum_{i=1}^{d} Var(Z)$

$= d\,Var(Z)$

Soon Chee Loong, Last Name: Soon, First Name: Chee Loong
cheeloong.soon@mail.utoronto.ca
CSC2515 Fall 2018
Homework 1 Writeup

Question 2 b)

```
root@soon:~/Github/CSC411Fall2018Assignments/Homework1# bash runTree.sh | tee runOutput.txt
criterion = gini, depth = 5, validationAccuracy = 0.7034764826175869
criterion = gini, depth = 10, validationAccuracy = 0.6993865030674846
criterion = gini, depth = 100, validationAccuracy = 0.7239263803680982
criterion = gini, depth = 200, validationAccuracy = 0.7239263803680982
criterion = gini, depth = None, validationAccuracy = 0.7239263803680982
criterion = entropy, depth = 5, validationAccuracy = 0.7014314928425358
criterion = entropy, depth = 10, validationAccuracy = 0.6973415132924335
criterion = entropy, depth = 100, validationAccuracy = 0.7198364008179959
criterion = entropy, depth = 200, validationAccuracy = 0.7321063394683026
criterion = entropy, depth = None, validationAccuracy = 0.7259713701431493
bestCriterion = entropy, bestDepth = 200, maxValidAccuracy = 0.7321063394683026, testAccuracy = 0.740816326530612
Word = trump, informationGain = 0.03604646543849388
Word = donald, informationGain = 0.047632623717864586
Word = hillary, informationGain = 0.0353197213188684
Word = the, informationGain = 0.050187537139538274
Word = trumps, informationGain = 0.04591824908230524
Word = here, informationGain = 0.005107066410498073
```

Figure 1: Actual output of real run, which includes 2 b) and 2 d)

The picture of real run in Figure 1 above is a little small, so I pasted the outputs below.
Depth of None means unlimited depth in scikit-learn's DecisionTreeClassifier.
The best criterion that was selected using the highest validation accuracy is pasted below.
I also evaluated on the test accuracy to see if it generalizes well on the test set.

criterion = gini, depth = 5, validationAccuracy = 0.7034764826175869
criterion = gini, depth = 10, validationAccuracy = 0.6993865030674846
criterion = gini, depth = 100, validationAccuracy = 0.7239263803680982
criterion = gini, depth = 200, validationAccuracy = 0.7239263803680982
criterion = gini, depth = None, validationAccuracy = 0.7239263803680982
criterion = entropy, depth = 5, validationAccuracy = 0.7014314928425358
criterion = entropy, depth = 10, validationAccuracy = 0.6973415132924335
criterion = entropy, depth = 100, validationAccuracy = 0.7198364008179959
criterion = entropy, depth = 200, validationAccuracy = 0.7321063394683026
criterion = entropy, depth = None, validationAccuracy = 0.7259713701431493
bestCriterion = entropy, bestDepth = 200, maxValidAccuracy = 0.7321063394683026, testAccuracy = 0.7408163265306122

Soon Chee Loong, Last Name: Soon, First Name: Chee Loong
cheeloong.soon@mail.utoronto.ca
CSC2515 Fall 2018
Homework 1 Writeup

Question 2 c)

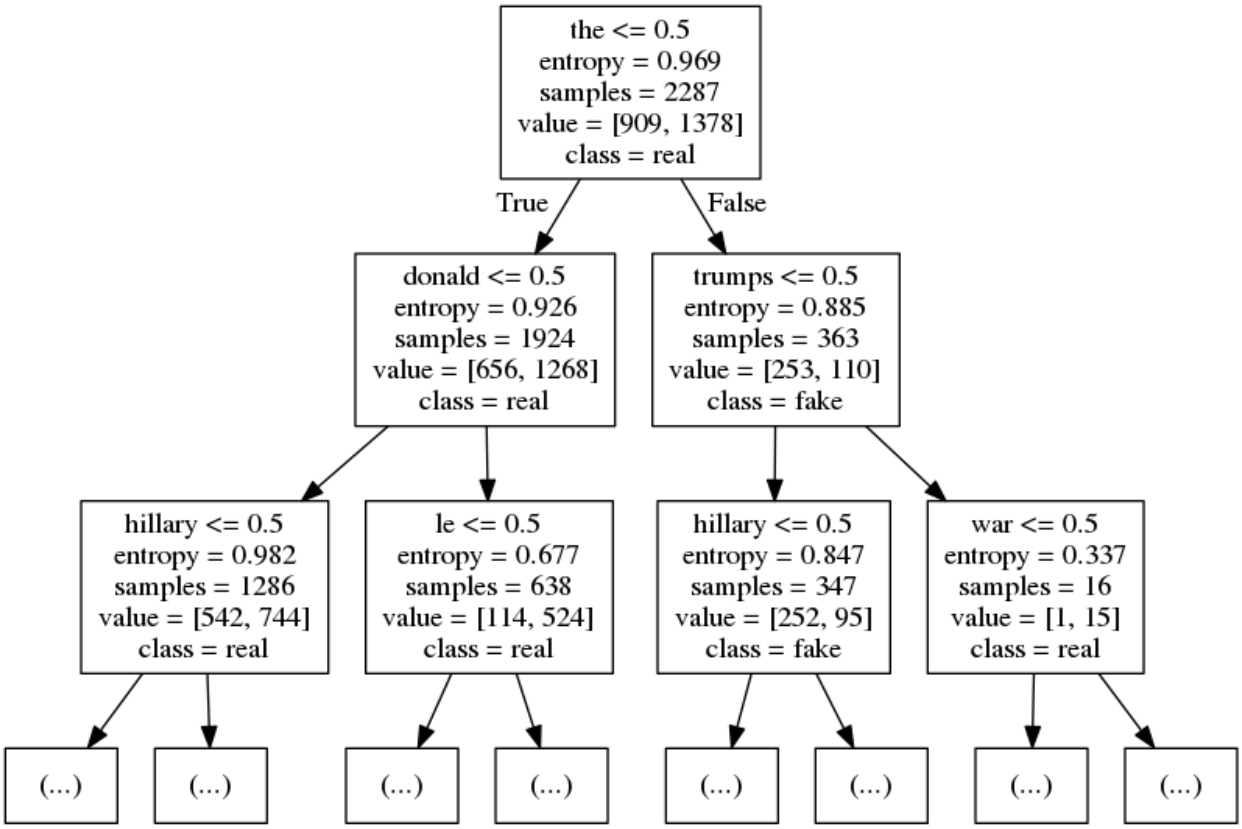


Figure 2: Output of tree with depth of 2 using graphviz.

The tree above was plotted using GraphViz.

Soon Chee Loong, Last Name: Soon, First Name: Chee Loong
cheeloong.soon@mail.utoronto.ca
CSC2515 Fall 2018
Homework 1 Writeup

Question 2 d)

The results on information gain is also pasted below.
As you can see, the word **'the'** (bolded below) had the highest information gain, and hence was used greedily as the root of the split.

Pasted from Figure 1 above.

Word = trump, informationGain = 0.03604646543849388
Word = donald, informationGain = 0.047632623717864586
Word = hillary, informationGain = 0.0353197213188684
Word = the, informationGain = 0.050187537139538274
Word = trumps, informationGain = 0.04591824908230524
Word = here, informationGain = 0.005107066410498073