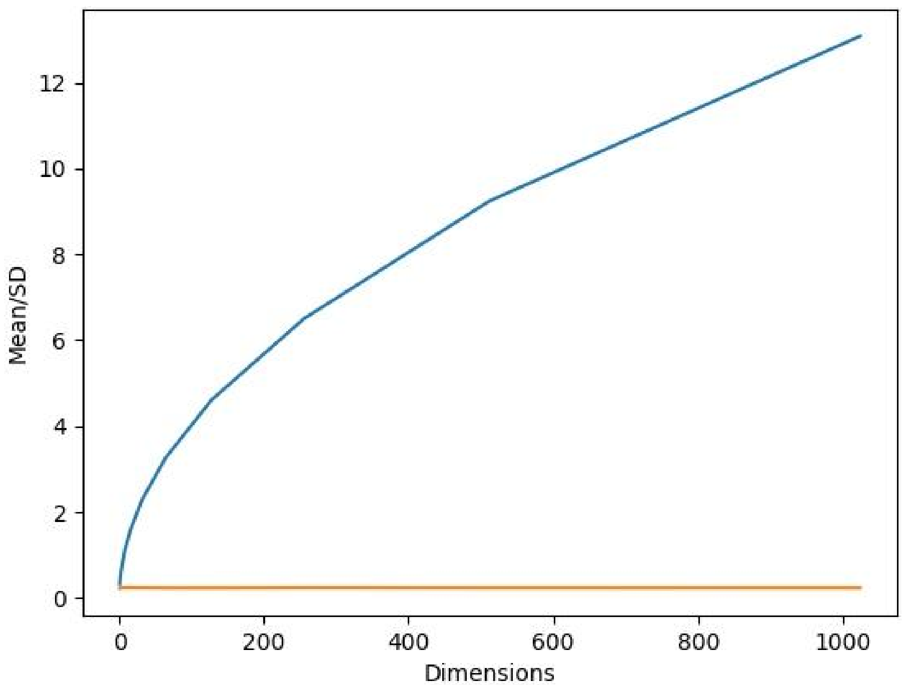
Q1.
(a)

Figure 1: Plot of the Average and Standard Deviation of the Squared Euclidean Distance



Legend
Blue Line = Average Squared Euclidean Distance
Orange Line = Standard Deviation of the Squared Euclidean Distance

(b)

1)

(b) $E[R] = E[Z_1 + \dots + Z_d]$

$= E\left[\sum_{i=1}^{d} z_i\right]$ where $z_i = (x_i - y_i)^2$

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

$= d\,E[z_i]$

$= \frac{d}{6}$

$Var[R] = Var[Z_1 + \dots + Z_d]$

$= Var\left[\sum_{i=1}^{d} z_i\right]$

$= \sum_{i=1}^{d} Var[z_i]$

$= \frac{7d}{180}$

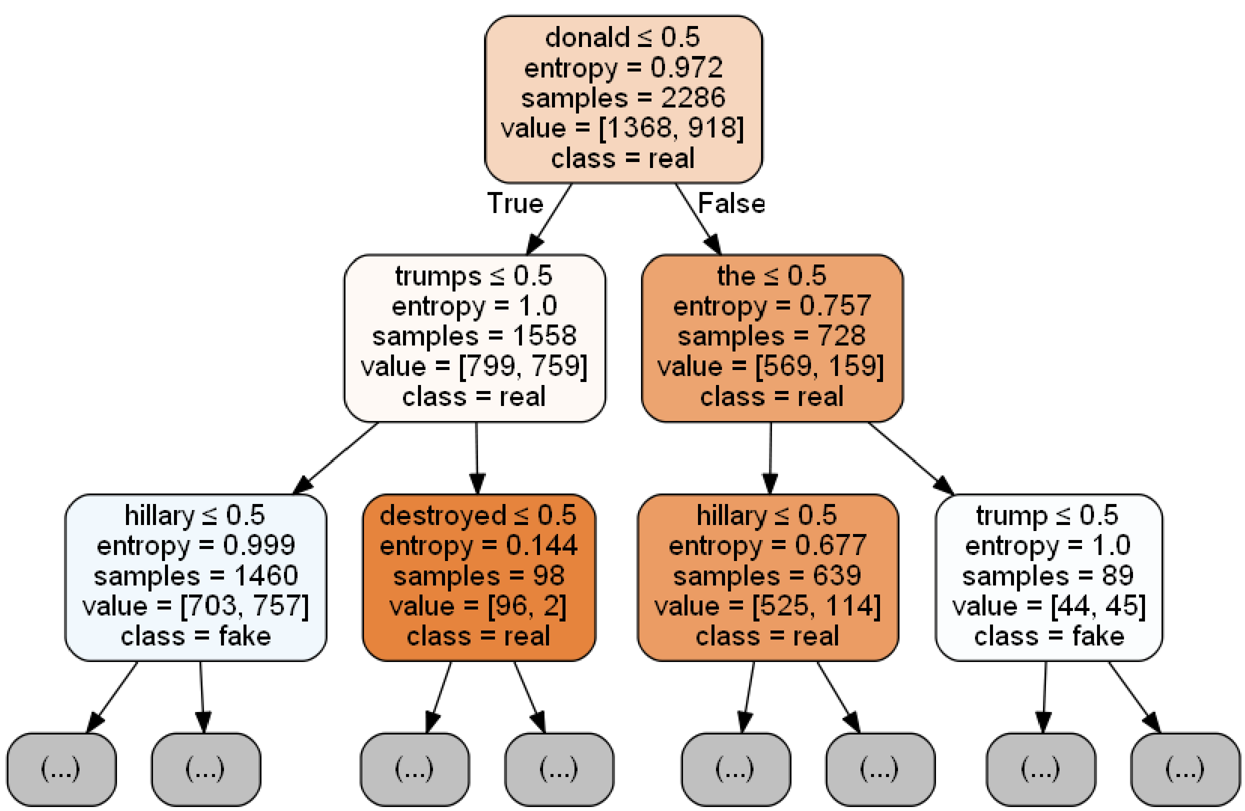
$\therefore E[R] = \frac{d}{6} \qquad Var[R] = \frac{7d}{180}$

2.

(b)

```
Depth: 5
Gini Accuracy: 0.7183673469387755
Information Gain Accuracy: 0.7163265306122449
Depth: 10
Gini Accuracy: 0.7306122448979592
Information Gain Accuracy: 0.7224489795918367
Depth: 20
Gini Accuracy: 0.7551020408163265
Information Gain Accuracy: 0.7448979591836735
Depth: 40
Gini Accuracy: 0.753061224489796
Information Gain Accuracy: 0.7448979591836735
Depth: 80
Gini Accuracy: 0.7755102040816326
Information Gain Accuracy: 0.7653061224489796
```

(c)



(d)

Top most split: ‘donald’

```
Keyword chosen for the split: donald
Information Gain on respective keyword: 0.037168762239267766
Keyword chosen for the split: trumps
Information Gain on respective keyword: 0.027569459011552694
Keyword chosen for the split: the
Information Gain on respective keyword: 0.03246994945859035
Keyword chosen for the split: hillary
Information Gain on respective keyword: 0.0246722560190763
Keyword chosen for the split: destroyed
Information Gain on respective keyword: 0.0006164502599837612
Keyword chosen for the split: trump
Information Gain on respective keyword: 0.030112168541829715
```

3.
(a)

3.
(a)

$$w_j \leftarrow w_j - \alpha \frac{\partial J_{reg}^{\beta}}{\partial w_j}$$

$$= w_j - \alpha \frac{\partial}{\partial w_j}(J + R)$$

$$= w_j - \alpha \left(\frac{\partial J}{\partial w_j} + \frac{\partial R}{\partial w_j} \right)$$

$$= w_j - \alpha \frac{\partial J}{\partial w_j} - \alpha \frac{\partial R}{\partial w_j}$$

$$= w_j - \alpha \frac{\partial J}{\partial w_j} - \alpha \beta_j w_j$$

$$= (1 - \alpha\beta_j) w_j - \alpha \frac{\partial J}{\partial w_j}$$

$$= (1 - \alpha\beta_j) w_j - \frac{\alpha}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)}) x_j^{(i)}$$

3.

(a) cont.

$$b \leftarrow b - \frac{\partial J_{reg}^{\beta}}{\partial b}$$

$$= b - \alpha \frac{\partial}{\partial b}(J + R)$$

$$= b - \alpha \frac{\partial J}{\partial b} - \alpha \frac{\partial R}{\partial b}$$

$$= b - \alpha \frac{\partial J}{\partial b}$$

$$= b - \frac{\alpha}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)})$$

If we ignore the gradient of the cost function, $\frac{\delta J}{dw_j}$, and consider the gradient from the L^2 regularization, we shrink the weight at each iteration by a factor smaller than 1 which is $(1 - \alpha\beta_j)w_j$ thus, giving the name 'weight decay'.

3.
(b)

3.

(b) $J_{reg}^{\beta} = \frac{1}{2N} \sum_{i=1}^{N} \left(y^{(i)} - t^{(i)} \right)^2 + \frac{1}{2} \sum_{j=1}^{D} \beta_j w_j^2$

$$\frac{\partial J_{reg}^{\beta}}{\partial w_j} = \frac{\partial J}{\partial w_j} + \frac{\partial R}{\partial w_j}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j'=1}^{D} w_{j'} x_{j'}^{(i)} - t^{(i)} \right) x_j^{(i)} + \beta_j w_j$$

$$= \frac{1}{N} \sum_{i=1}^{N} \sum_{j'=1}^{D} x_j^{(i)} x_{j'}^{(i)} w_{j'} + \beta_j w_j - \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)} t^{(i)}$$

$$= \sum_{j'=1}^{D} \frac{1}{N} \sum_{i=1}^{N} \left(x_j^{(i)} x_{j'}^{(i)} + \beta_j \right) w_{j'} - \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)} t^{(i)} \quad \text{iff } j = j'$$

$$\therefore A_{jj'} = \begin{cases} \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)} + \beta_j & \text{if } j = j' \\ \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)} & \text{if } j \neq j' \end{cases}$$

$$c_j = \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)} t^{(i)}$$

3.
(c)

3.

(c) from part (b), we have:

$$\left(\frac{1}{N}X^TX + \operatorname{diag}(\vec{\beta})\right)\vec{w} - \frac{1}{N}X^Tt = 0$$

where $\vec{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_D \end{pmatrix}$

Solving for $\vec{w}$, we have

$$\left(X^TX + N\operatorname{diag}(\vec{\beta})\right)\vec{w} = X^Tt$$

$$\vec{w} = \frac{X^Tt}{\left(X^TX + N\operatorname{diag}(\vec{\beta})\right)}$$

$$\vec{w} = \left(X^TX + N\operatorname{diag}(\vec{\beta})\right)^{-1}X^Tt$$