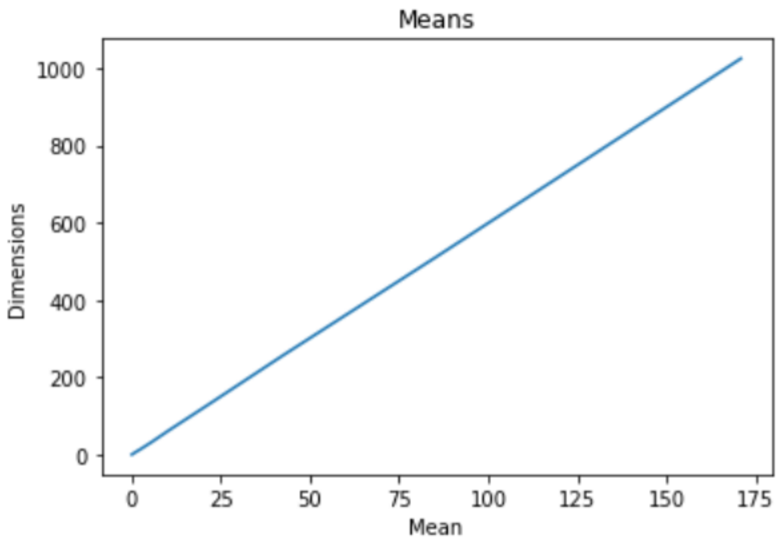
CSC311 HW1

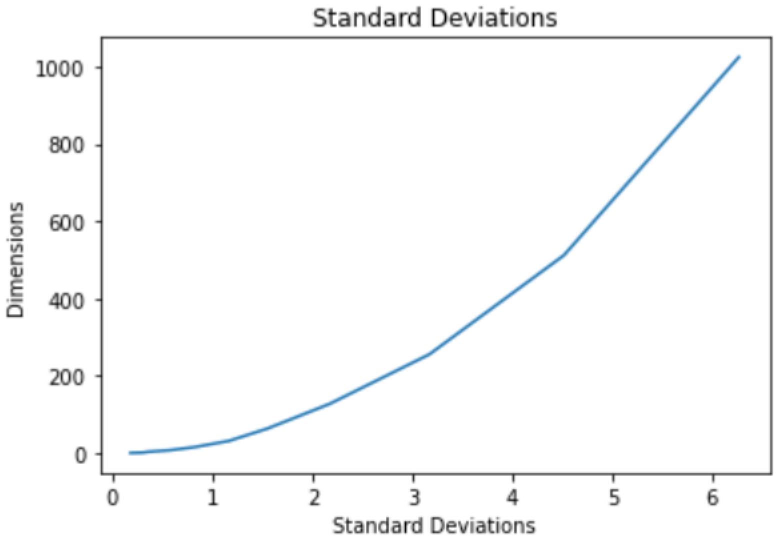
Shiven Taneja (1005871013)

September 2021

1 Nearest Neighbours and the Curse of Dimensionality

1.1 a)





1.2 b)

$$E[R] = E[Z_1 + ... + Z_d] = E[Z_1] + ... + E[Z_d] = \frac{1}{6} + ... + \frac{1}{6} = \frac{d}{6}$$

$$Var[R] = Var[Z_1 + ... + Z_d] = Var[Z_1] + ... + Var[Z_d] = \frac{7}{180} + ... + \frac{7}{180} = \frac{7d}{180}$$

2 Decision Trees

2.1 b)

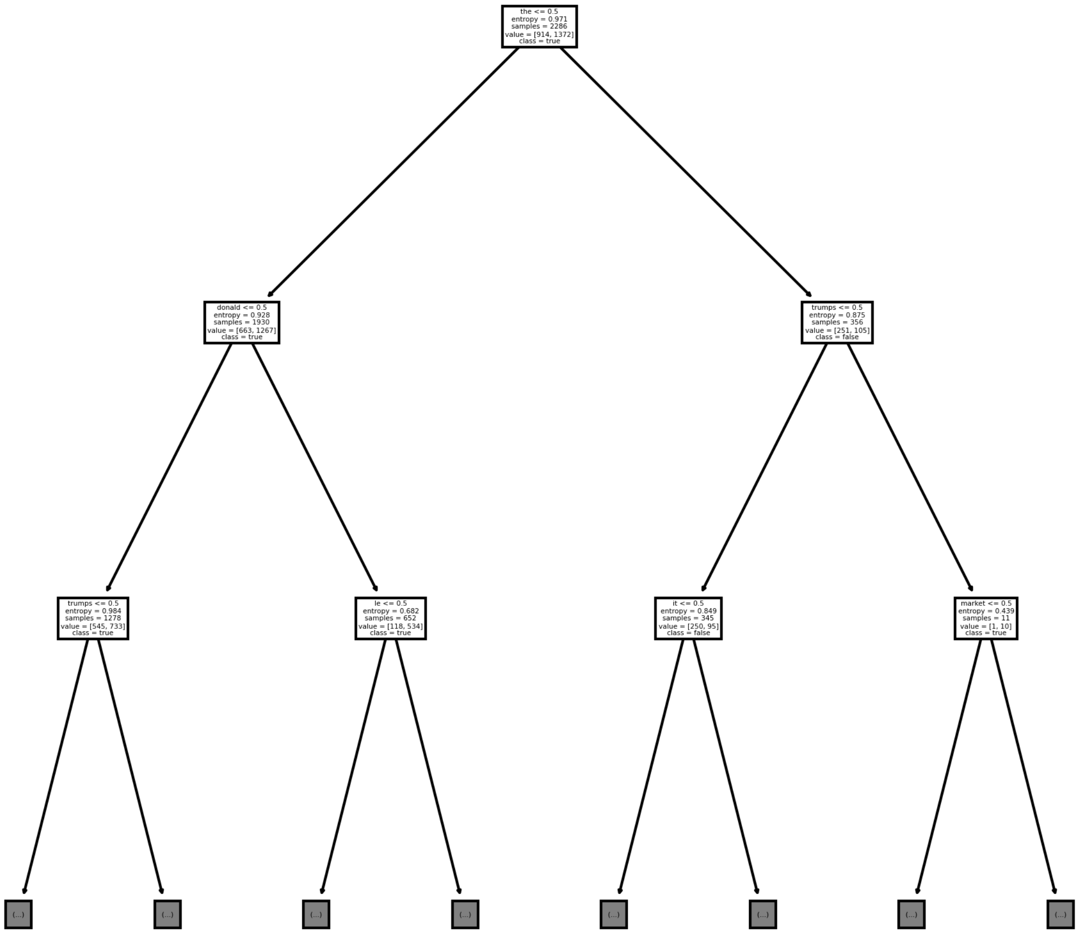
Output:

```
C:\Users\shive\anaconda3\envs\hw1_code\python.exe C:/Users/shive/PycharmProjects/hw1_code/hw1
The accuracy with Gini Coefficient and max_depth = 2 is: 0.6693877551020408
The accuracy with Gini Coefficient and max_depth = 3 is: 0.6693877551020408
The accuracy with Gini Coefficient and max_depth = 5 is: 0.6612244897959184
The accuracy with Gini Coefficient and max_depth = 7 is: 0.6836734693877551
The accuracy with Gini Coefficient and max_depth = 10 is: 0.7
The accuracy with Information Gain and max_depth = 2 is: 0.636734693877551
The accuracy with Information Gain and max_depth = 3 is: 0.6693877551020408
The accuracy with Information Gain and max_depth = 5 is: 0.6612244897959184
The accuracy with Information Gain and max_depth = 7 is: 0.6755102040816326
The accuracy with Information Gain and max_depth = 10 is: 0.7020408163265306

Process finished with exit code 0
```

Thus the hyper-parameters with the highest validation accuracy is with Information Gain and max_depth = 10

2.2 c)


the <= 0.5
entropy = 0.971
samples = 2286
value = [914, 1372]
class = true
donald <= 0.5
entropy = 0.928
samples = 1930
value = [663, 1267]
class = true
trumps <= 0.5
entropy = 0.875
samples = 356
value = [251, 105]
class = false
trumps <= 0.5
entropy = 0.984
samples = 1278
value = [545, 733]
class = true
le <= 0.5
entropy = 0.682
samples = 652
value = [118, 534]
class = true
it <= 0.5
entropy = 0.849
samples = 345
value = [250, 95]
class = false
market <= 0.5
entropy = 0.439
samples = 11
value = [1, 10]
class = true
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

2.3 d)

```
the Information Gain for the word the is: 0.0508217499222694
the Information Gain for the word donald is: 0.05107848514883384
the Information Gain for the word trumps is: 0.043014604836436376
the Information Gain for the word le is: 0.002486932652073781
the Information Gain for the word it is: 0.004666524574022279
the Information Gain for the word market is: 2.433624434172721e-05
```

3 Regularized Linear Regression

3.1 a)

We know that: $R(w) = \frac{\lambda}{2}\mathbf{w}^\top\mathbf{w} = \frac{\lambda}{2}\sum_{j=1}^{D} w_j^2$

And that: $\Im_{reg}^{\beta}(\mathbf{w}) = \Im(\mathbf{w}) + \lambda\Re(\mathbf{w}) = \frac{1}{2N}\sum_{i=1}^{N}(y^i - t^i)^2 + \frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2$

Thus we can say: $\frac{\partial\Im_{reg}^{\beta}}{\partial\mathbf{w}} = \frac{\partial\Im}{\partial\mathbf{w}} + \frac{\partial\Re}{\partial\mathbf{w}}$

We further know from lecture that: $\frac{\partial\Im}{\partial\Im} = \frac{1}{N}\sum_{i=1}^{N}(y^i - t^i)\mathbf{x}^i$

Solving for $\frac{\partial\Re}{\partial\mathbf{w}}$ we get: $\frac{\partial}{\partial\mathbf{w}}\frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2 = w_j\beta_j$

We can now find the gradient descent update rules for some $\alpha > 0$:

$w_j \leftarrow w_j - \alpha\frac{\partial\Im_{reg}^{\beta}}{\partial\mathbf{w}}$

$w_j \leftarrow w_j - \alpha(\frac{\partial\Im}{\partial\mathbf{w}} + \frac{\partial\Re}{\partial\mathbf{w}})$

$w_j \leftarrow w_j - (\frac{\alpha}{N}\sum_{i=1}^{N}(y^i - t^i)\mathbf{x}^i + \alpha w_j\beta_j)$

$b \leftarrow b - \alpha\frac{\partial\Im_{reg}^{\beta}}{\partial\mathbf{w}}$

Since there is no regularization penalty on the bias parameter, we can then rewrite the above equation as:

$b \leftarrow b - \alpha\frac{\partial\Im}{\partial\mathbf{w}}$

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

It is called the weight decay as the term causes the weight to exponentially decay down to zero. It further removes more of the weight for larger weights and less for smaller weights.

3.2 b)

We know that: $\Im_{reg}^{\beta}(\mathbf{w}) = \frac{1}{2N}\sum_{i=1}^{N}(y^i - t^i)^2 + \frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2$

And that: $\frac{\partial\Im_{reg}^{\beta}}{\partial\mathbf{w}} = \sum_{j'=1}^{D} A_{jj'}w_{j'} - c_j = \frac{\partial\Im}{\partial\mathbf{w}} + \frac{\partial\Re}{\partial\mathbf{w}} = (\frac{1}{N}\sum_{i=1}^{N}(y^i - t^i)\mathbf{x}^i + w_j\beta_j$

If we expand the term $(y^i + t^i)\mathbf{x}^i$ we get: $\frac{\partial\Im_{reg}^{\beta}}{\partial\mathbf{w}} = \frac{1}{N}\sum_{i=1}^{N}(y^i\mathbf{x}^i - t^i\mathbf{x}^i) + w_j\beta_j$

We can then distribute the summation to get: $\frac{\partial\Im_{reg}^{\beta}}{\partial\mathbf{w}} = \frac{1}{N}\sum_{i=1}^{N} y^i\mathbf{x}^i - \frac{1}{N}\sum_{i=1}^{N} t^i\mathbf{x}^i + w_j\beta_j$

Therefore we can get that:

$A_{jj'} = \frac{1}{N}\sum_{i=1}^{N} y^i\mathbf{x}^i$

And thus:

$c_j = \frac{1}{N}\sum_{i=1}^{N} t^i\mathbf{x}^i + w_j\beta_j$