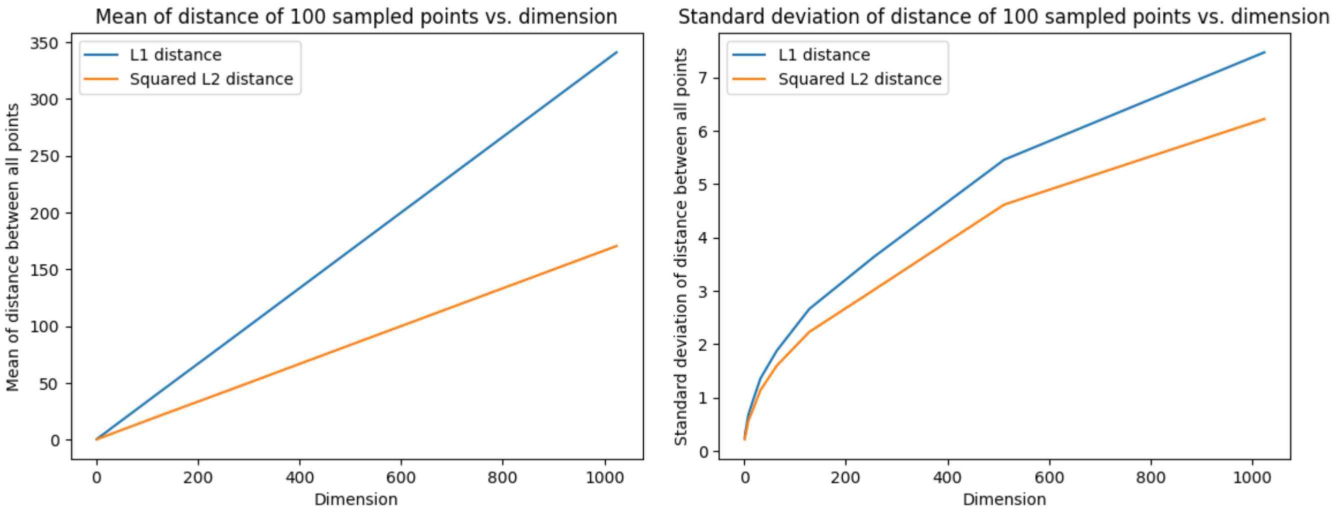
CSC311 Assignment 1

June 5, 2023

1. (a) To guarantee that any new point will be ≤ 0.01 of an existing one, the existing points can be 0.02 apart since the new point can lie in the middle of every pair ($= 0.01$ to each point) or closer to one point (< 0.01 to one point). In this scenario, $\dfrac{1}{0.02} = 50$ points are needed.

   (b) For 10 features, the volume of $[0,1]^{10}$ is still 1, while the volume of the neighborhood of a point is $0.02^{10} = 1.024 \times 10^{-17}$. Since $\dfrac{1}{1.024 \times 10^{-17}} = 9.766 \times 10^{16}$ points are now needed, which is much more than the 50 points for one feature, the guarantee is harder to maintain.

   (c) 

   (d) $\mathbb{E}[Z] = \mathbb{E}[\sum_{i=1}^{d} Z_i] = \sum_{i=1}^{d} \mathbb{E}[Z_i] = \dfrac{d}{6}$ by linearity of expectation.

   $\text{Var}[Z] = \text{Var}[\sum_{i=1}^{d} Z_i] = \sum_{i=1}^{d} \text{Var}[Z_i] = \dfrac{7d}{180}$ since all X_i, Y_i are independent and so all $Z_i = (X_i - Y_i)^2$ are independent as well.

   (e) i. E can be written as "$|R - \mathbb{E}[R]| \geq k$".

   ii. $\mathbb{P}[E] = \mathbb{P}[|R - \mathbb{E}(R)| \geq k] \leq \dfrac{\text{Var}[R]}{k^2}$.

   iii. Taking the limit yields

$$\lim_{d\to\infty} \mathbb{P}[E] \leq \lim_{d\to\infty} \frac{\text{Var}[R]}{k^2} = \lim_{d\to\infty} \frac{7d}{180k^2} = \lim_{d\to\infty} \frac{7}{180c^2 d} = 0$$

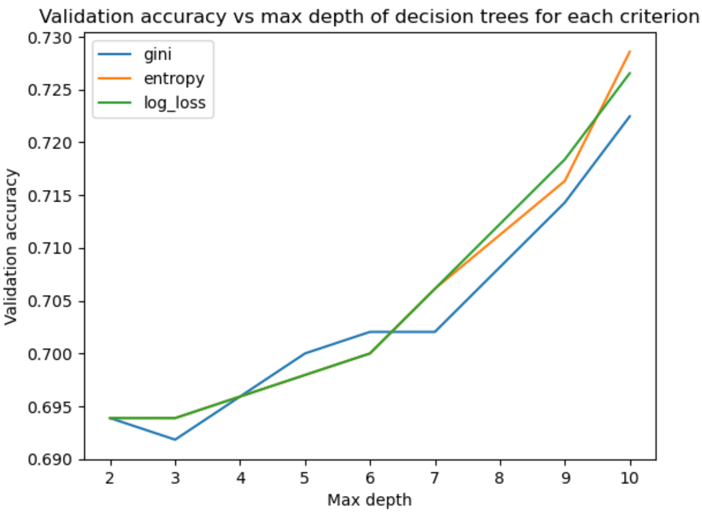
   since $\text{Var}[R] = \dfrac{7d}{180}$ by part (d) and $k = cd$ for $c > 0$. Since $\mathbb{P}[E] \geq 0$ by definition, we apply the squeeze theorem to get

$$\lim_{d\to\infty} \mathbb{P}[E] = 0$$

   as needed.

2. (a) See the Python code.

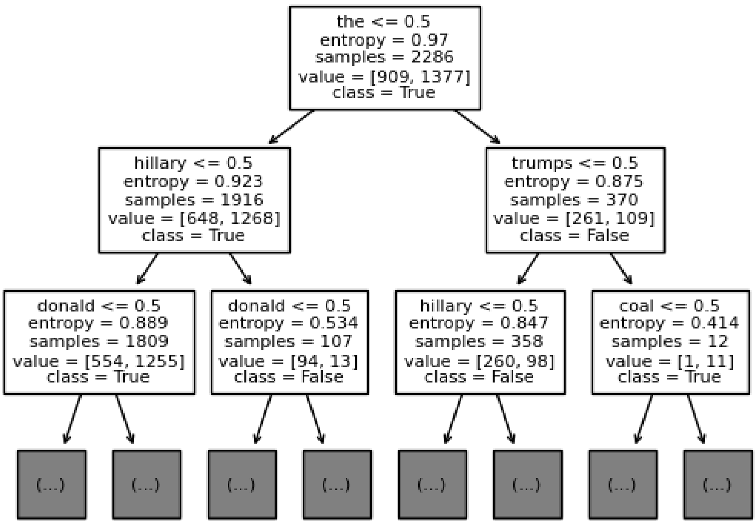
(b) The plot and function output are shown below. The test accuracy of the hyperparameters with the highest validation accuracy is 0.697959.




```
Criterion: gini, max_depth: 2, accuracy: 0.6938775510204082
Criterion: gini, max_depth: 3, accuracy: 0.6918367346938775
Criterion: gini, max_depth: 5, accuracy: 0.7
Criterion: gini, max_depth: 6, accuracy: 0.7020408163265306
Criterion: gini, max_depth: 7, accuracy: 0.7061224489795919
Criterion: gini, max_depth: 9, accuracy: 0.7142857142857143
Criterion: gini, max_depth: 10, accuracy: 0.7244897959183674
Criterion: entropy, max_depth: 2, accuracy: 0.6938775510204082
Criterion: entropy, max_depth: 3, accuracy: 0.6938775510204082
Criterion: entropy, max_depth: 5, accuracy: 0.6979591836734694
Criterion: entropy, max_depth: 6, accuracy: 0.7020408163265306
Criterion: entropy, max_depth: 7, accuracy: 0.7061224489795919
Criterion: entropy, max_depth: 9, accuracy: 0.7183673469387755
Criterion: entropy, max_depth: 10, accuracy: 0.726530612244898
Criterion: log_loss, max_depth: 2, accuracy: 0.6938775510204082
Criterion: log_loss, max_depth: 3, accuracy: 0.6938775510204082
Criterion: log_loss, max_depth: 5, accuracy: 0.6979591836734694
Criterion: log_loss, max_depth: 6, accuracy: 0.7020408163265306
Criterion: log_loss, max_depth: 7, accuracy: 0.7061224489795919
Criterion: log_loss, max_depth: 9, accuracy: 0.7183673469387755
Criterion: log_loss, max_depth: 10, accuracy: 0.726530612244898
Test accuracy of tree with the highest validation accuracy: 0.697959
```

(1)

(c) The visualization is shown below.



(2)

(d) The topmost split is 'the' ≤ 0.5, which has an information gain of 0.054313. Other information gains include 0.023501 for 'us', 0.042801 for 'trumps', 0.047085 for 'hillary', and 0.046617 for 'donald'. The function output is shown below.

```
Information gain for the: 0.05431294336040304
Information gain for us: 0.023500563840293753
Information gain for trumps: 0.04280069136897102
Information gain for hillary: 0.04708470956770994
Information gain for donald: 0.046616794619072865
```

(3)

3. See next page.

3. a) For $\vec{w}=(w_1,\dots,w_D)$, define $\mathcal{J}(\vec{w})=\frac{1}{2N}\sum_{i=1}^{N}(y^{(i)}-t^{(i)})^2=\frac{1}{2N}\sum_{i=1}^{N}(\sum_{j=1}^{D}w_jx_j^{(i)}+b-t^{(i)})^2$, $f(\vec{w})=\sum_{j=1}^{D}\alpha_j|w_j|$, and $g(\vec{w})=\frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2$. Then,

$$\frac{\partial\mathcal{J}(\vec{w})}{\partial w_j}=\frac{\partial}{\partial w_j}\frac{1}{2N}\sum_{i=1}^{N}(\sum_{j'=1}^{D}w_{j'}x_{j'}^{(i)}+b-t^{(i)})^2=\frac{1}{2N}\sum_{i=1}^{N}\left[2(\sum_{j'=1}^{D}w_{j'}x_{j'}^{(i)}+b-t^{(i)})\frac{\partial}{\partial w_j}(\sum_{j'=1}^{D}w_{j'}x_{j'}^{(i)}+b-t^{(i)})\right]$$

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

$$\frac{\partial\mathcal{J}(\vec{w})}{\partial b}=\frac{\partial}{\partial b}\frac{1}{2N}\sum_{i=1}^{N}(\sum_{j'=1}^{D}w_{j'}x_{j'}^{(i)}+b-t^{(i)})^2=\frac{1}{2N}\sum_{i=1}^{N}\left[2(\sum_{j'=1}^{D}w_{j'}x_{j'}^{(i)}+b-t^{(i)})\frac{\partial}{\partial b}(\sum_{j'=1}^{D}w_{j'}x_{j'}^{(i)}+b-t^{(i)})\right]$$

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

$$\frac{\partial f(\vec{w})}{\partial w_j}=\begin{cases}\frac{\partial}{\partial w_j}\sum_{j'=1}^{D}\alpha_{j'}w_{j'}=\frac{\partial}{\partial w_j}(\alpha_1w_1+\dots+\alpha_jw_j+\dots+\alpha_Dw_D)=\alpha_j & \text{when } w_j>0\\ \frac{\partial}{\partial w_j}\sum_{j'=1}^{D}\alpha_{j'}w_{j'}=\frac{\partial}{\partial w_j}(\alpha_1w_1+\dots+\alpha_j(0)+\dots+\alpha_Dw_D)=0 & \text{when } w_j=0\\ \frac{\partial}{\partial w_j}\sum_{j'=1}^{D}\alpha_{j'}w_{j'}=\frac{\partial}{\partial w_j}(\alpha_1w_1+\dots+\alpha_j(-w_j)+\dots+\alpha_Dw_D)=-\alpha_j & \text{when } w_j<0\end{cases}$$

$$\frac{\partial f(\vec{w})}{\partial b}=0$$

$$\frac{\partial g(\vec{w})}{\partial w_j}=\frac{\partial}{\partial w_j}\frac{1}{2}\sum_{j'=1}^{D}\beta_{j'}w_{j'}^2=\frac{\partial}{\partial w_j}(\tfrac{1}{2}\beta_1w_1^2+\dots+\tfrac{1}{2}\beta_jw_j^2+\dots+\tfrac{1}{2}\beta_Dw_D^2)=\tfrac{1}{2}\cdot2\beta_jw_j=\beta_jw_j$$

$$\frac{\partial g(\vec{w})}{\partial b}=0$$

Since the gradient descent rule has the form $w_j\leftarrow w_j-\lambda\frac{\partial\mathcal{J}_{reg}^{\alpha\beta}(\vec{w})}{\partial w_j}$ and $b\leftarrow b-\lambda\frac{\partial\mathcal{J}_{reg}^{\alpha\beta}(\vec{w})}{\partial w_j}$ where λ is the learning rate, we have:

if $w_j>0$: $$\frac{\partial\mathcal{J}_{reg}^{\alpha\beta}(\vec{w})}{\partial w_j}=\frac{\partial}{\partial w_j}(\mathcal{J}(\vec{w})+f(\vec{w})+g(\vec{w}))=\frac{1}{N}\sum_{i=1}^{N}(y^{(i)}-t^{(i)})x_j^{(i)}+\alpha_j+\beta_jw_j$$

$$\frac{\partial\mathcal{J}_{reg}^{\alpha\beta}(\vec{w})}{\partial b}=\frac{\partial}{\partial b}(\mathcal{J}(\vec{w})+f(\vec{w})+g(\vec{w}))=\frac{1}{N}\sum_{i=1}^{N}(y^{(i)}-t^{(i)})$$

so the rules are $$w_j\leftarrow w_j-\lambda\left[\frac{1}{N}\sum_{i=1}^{N}(y^{(i)}-t^{(i)})x_j^{(i)}+\alpha_j+\beta_jw_j\right]$$

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

if $w_j = 0$: $$\frac{\partial \mathcal{J}_{reg}^{\alpha\beta}(\vec{w})}{\partial w_j} = \frac{\partial}{\partial w_j}\left(\mathcal{J}(\vec{w}) + f(\vec{w}) + g(\vec{w})\right) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)x_j^{(i)} + 0 + \beta_j(0)$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)x_j^{(i)}$$

$$\frac{\partial \mathcal{J}_{reg}^{\alpha\beta}(\vec{w})}{\partial b} = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right) \quad \text{as before}$$

so the rules are $w_j \leftarrow w_j - \frac{\lambda}{N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)x_j^{(i)}$

$$b \leftarrow b - \frac{\lambda}{N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)$$

if $w_j < 0$: $$\frac{\partial \mathcal{J}_{reg}^{\alpha\beta}(\vec{w})}{\partial w_j} = \frac{\partial}{\partial w_j}\left(\mathcal{J}(\vec{w}) + f(\vec{w}) + g(\vec{w})\right) = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)x_j^{(i)} - \alpha_j + \beta_j w_j$$

$$\frac{\partial \mathcal{J}_{reg}^{\alpha\beta}(\vec{w})}{\partial b} = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right) \quad \text{as before}$$

so the rules are $w_j \leftarrow w_j - \lambda\left[\frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)x_j^{(i)} - \alpha_j + \beta_j w_j\right]$

$$b \leftarrow b - \frac{\lambda}{N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)$$

It is called "weight decay" likely because the weight decays in proportion to its size due to the $-\lambda\beta_j w_j$ term.

b) Define $\mathcal{J}_{reg}^{\beta}(\vec{w}) = \frac{1}{2N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)^2 + \frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2$. Also, define $\mathbb{1}_{jj'} = \begin{cases} 1 & \text{if } j = j' \\ 0 & \text{if } j = j' \end{cases}$. Then,

$$\frac{\partial \mathcal{J}_{reg}^{\beta}(\vec{w})}{\partial w_j} = \frac{1}{N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)\frac{\partial}{\partial w_j}\left(y^{(i)} - t^{(i)}\right) + \frac{\partial}{\partial w_j}\frac{1}{2}\sum_{j'=1}^{D}\beta_{j'} w_{j'}^2$$

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

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

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

since $\mathbb{1}_{jj'}(N\beta_j w_j) = N\beta_j w_j$ iff $j = j'$

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

since $w_j = w_{j'}$ when $\mathbb{1}_{jj'} = 1$

$$= \sum_{j'=1}^{D}\left(A_{jj'} w_{j'}\right) - c_j = 0$$

(continued on next page)

Thus, $A_{jj'} = \frac{1}{N}\left[\sum_{i=1}^{N} x_{j'}^{(i)} x_j^{(i)} + \mathbb{1}_{jj'}(N\beta_j)\right]$ and $c_j = \frac{1}{N}\sum_{i=1}^{N} t^{(i)} x_j^{(i)}$ as needed.

c) Since $A_{jj'} = \frac{1}{N}\left[\sum_{i=1}^{N} x_{j'}^{(i)} x_j^{(i)} + \mathbb{1}_{jj'}(N\beta_j)\right]$, it follows that $A = \frac{1}{N}\left[X^T X + N\vec{\beta} I\right]$ where I is the identity matrix. This is because:

(1) X is an $N \times D$ matrix, so $X^T X$ is a $D \times D$ matrix. For $j \in \{1, \dots, D\}$ and $j' \in \{1, \dots, D\}$, $\sum_{i=1}^{N} x_{j'}^{(i)} x_j^{(i)} = \sum_{i=1}^{N} X_{j'i}^T X_{ij}$ by properties of matrix multiplication.

(2) $\mathbb{1}_{jj'}(N\beta_j) = NB_j$ only when $j = j'$, which is the diagonal of A, so this corresponds to $N\vec{\beta}$ multiplied by the identity matrix, where $\vec{\beta} = (\beta_1, \dots, \beta_D)$ is a D-dimensional vector.

Since $c_j = \frac{1}{N}\sum_{i=1}^{N} t^{(i)} x_j^{(i)}$, $\vec{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_D \end{bmatrix} = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} t^{(i)} x_1^{(i)} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} t^{(i)} x_D^{(i)} \end{bmatrix} = \frac{1}{N} X^T \vec{t}$ for $\vec{t} = (t_1, \dots, t_N)$.

Since $A\vec{w} - \vec{c} = 0$, $\vec{w} = A^{-1}\vec{c} = \left(\frac{1}{N}\left[X^T X + N\vec{\beta} I\right]\right)^{-1} \frac{1}{N} X^T \vec{t} = (X^T X + N\vec{\beta} I)^{-1} X^T \vec{t}$ as needed.