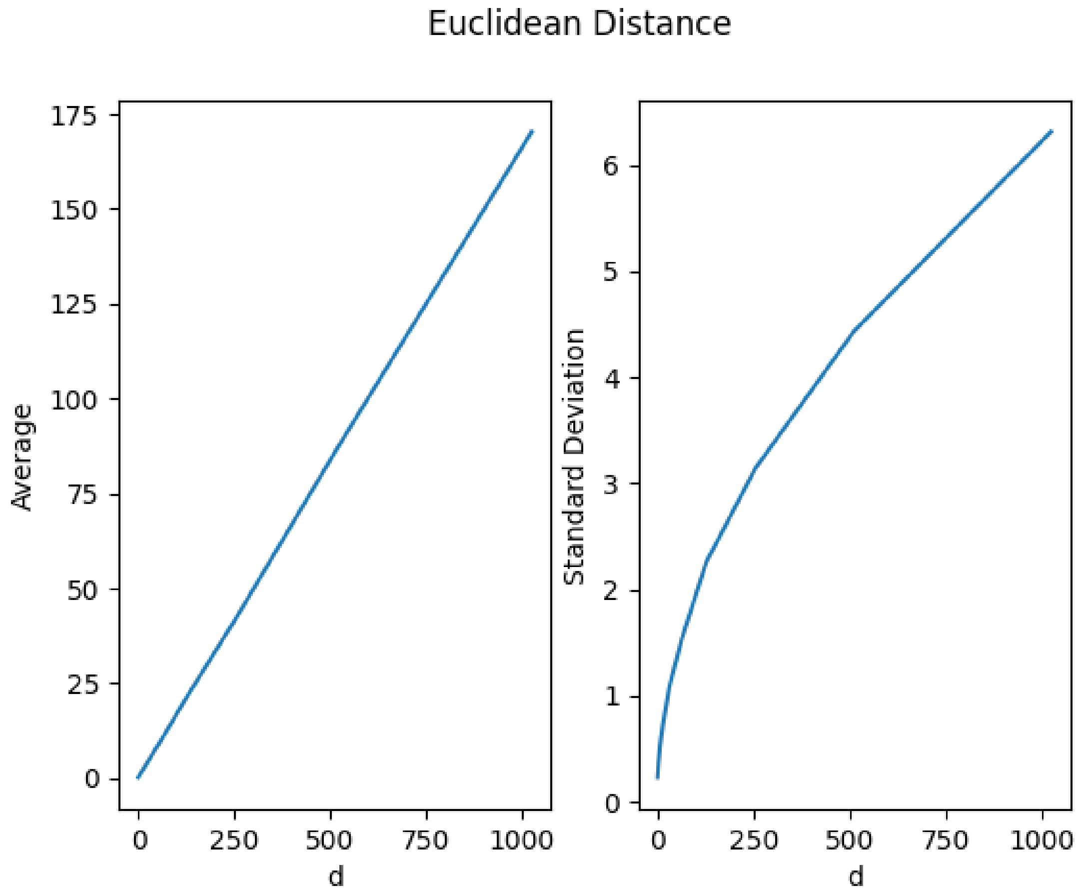
CSC311, Fall 2022, Homework 1

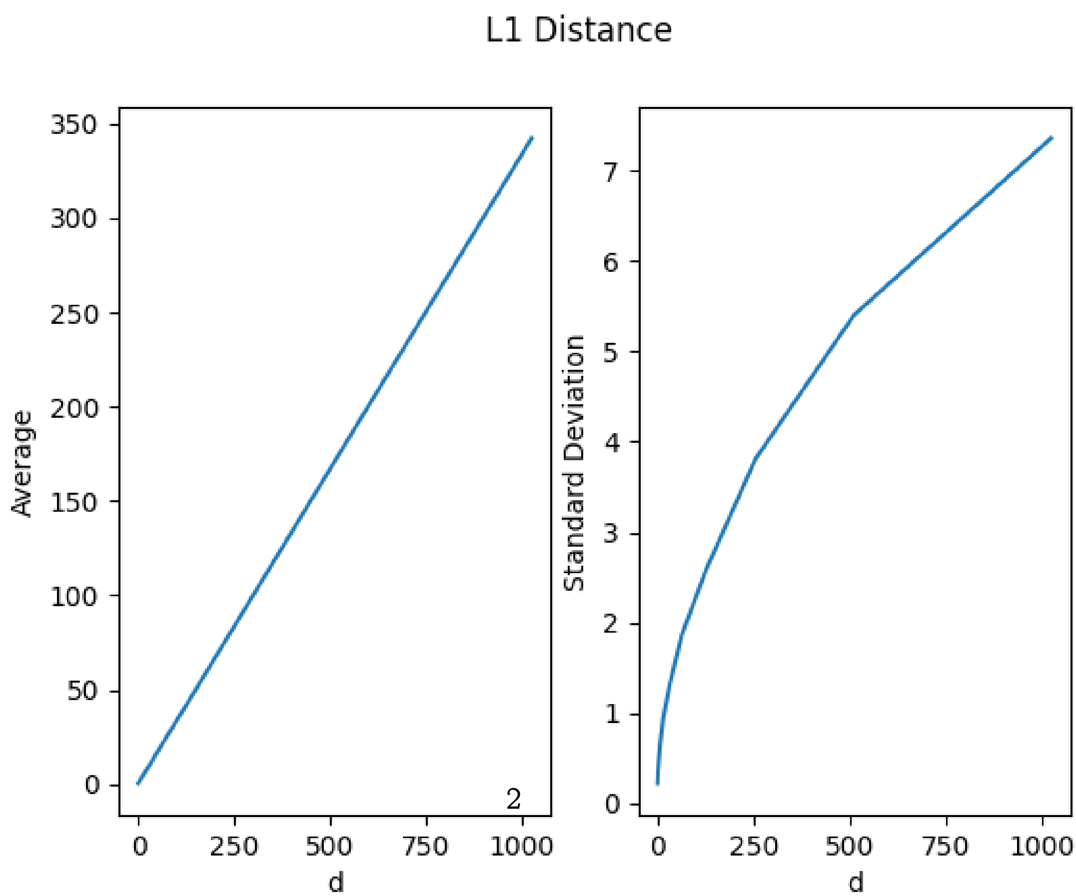
Huipeng Lei, 1005908154

Due October 3, 2022

Problem 1


Euclidean Distance
Average
Standard Deviation
d
d
0
25
50
75
100
125
150
175
0
1
2
3
4
5
6
0
250
500
750
1000

a)


L1 Distance
Average
Standard Deviation
d
d
0
50
100
150
200
250
300
350
0
1
2
3
4
5
6
7
0
250
500
750
1000

b)

$$\begin{aligned} E[R] &= E[Z_1 + Z_2 + \cdots + Z_d] \\ &= E[Z_1] + \cdots + E[Z_d] && \text{(by linearity of expectation)} \\ &= d \cdot \frac{1}{6} \end{aligned}$$

$$\begin{aligned} Var[R] &= Var[Z_1 + Z_2 + \cdots + Z_d] \\ &= Var[Z_1] + \cdots + Var[Z_d] && \text{(by linearity of variance w/ independence)} \\ &= d \cdot \frac{7}{180} \end{aligned}$$

c) Notice for any random variable Z,

$$-\mathbb{P}\Big(|Z - E[Z]| \geq d\Big) \geq -\frac{Var[Z]}{d^2}$$
$$\Rightarrow 1 - \mathbb{P}\Big(|Z - E[Z]| \geq d\Big) \geq 1 - \frac{Var[Z]}{d^2}$$

and

$$\mathbb{P}(E) = \mathbb{P}\Big(|R - E[R]| \leq d\Big) = 1 - \mathbb{P}\Big(|R - E[R]| \geq d\Big)$$

We have

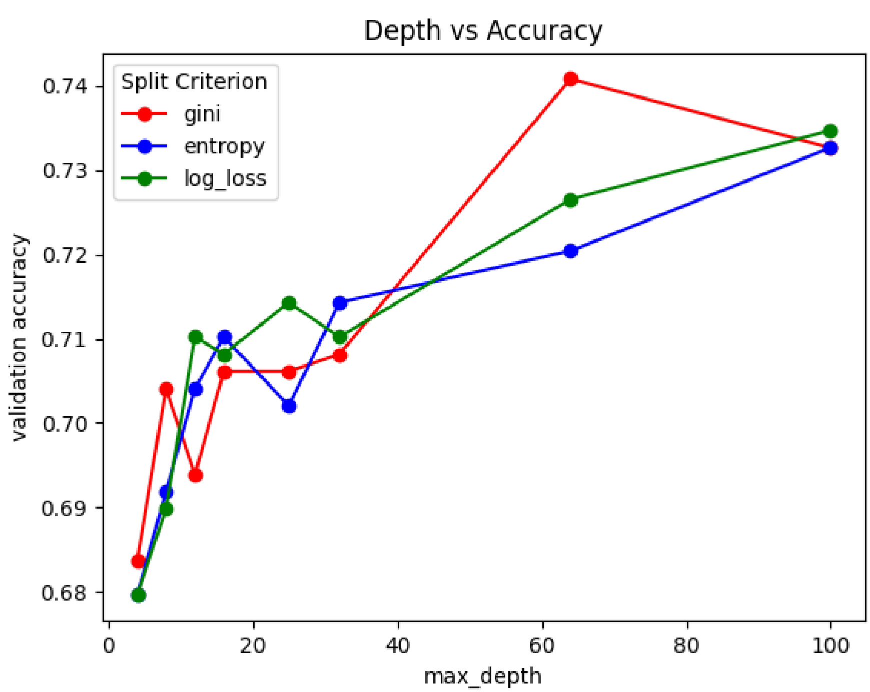
$$\begin{aligned} \mathbb{P}(E) &= 1 - \mathbb{P}\Big(|R - E[R]| \geq d\Big) \\ &\geq 1 - \frac{Var[R]}{d^2} \\ &= 1 - \frac{7 \cdot d}{180 \cdot d^2} \\ &= 1 - \frac{7}{180 \cdot d} \end{aligned}$$

$\mathbb{P}(E)$ approaches 1 as $d \to \infty$. Therefore, in high dimensions, the distance among 'most' points are almost the same even though each of them are far away from each other.

Problem 2

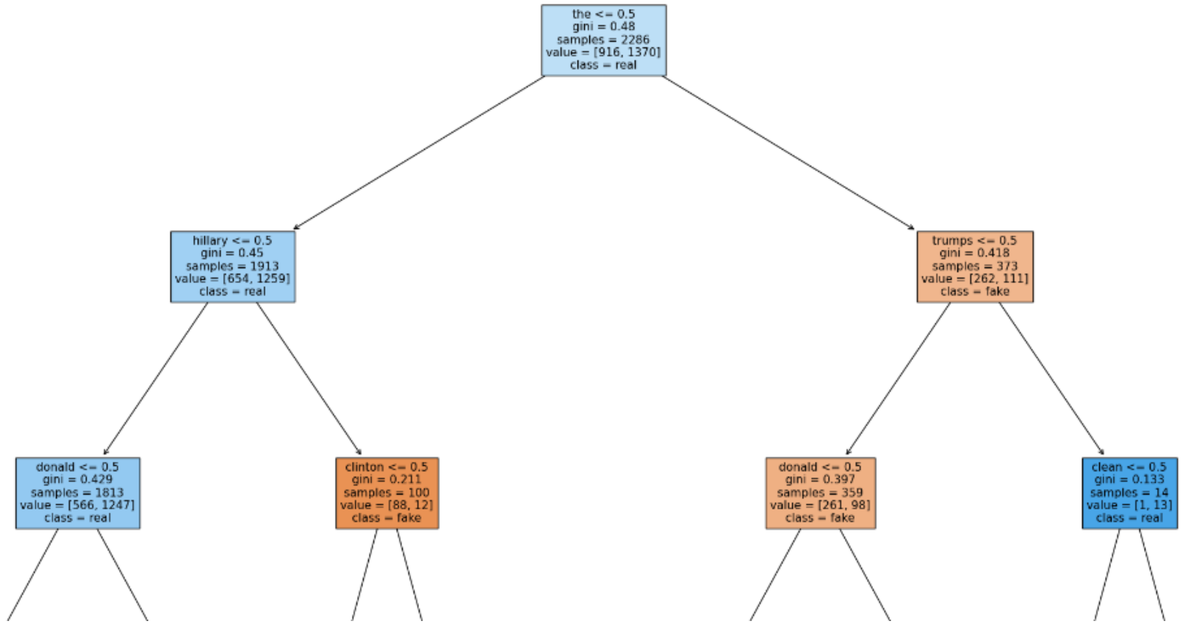
b)

```
Accuracy under `Information Gain` criterion with max_depth=4: 0.6795918367346939
Accuracy under `Log loss` criterion with max_depth=4: 0.6795918367346939
Accuracy under `Gini coefficient` criterion with max_depth=4: 0.6836734693877551
Accuracy under `Information Gain` criterion with max_depth=8: 0.6918367346938775
Accuracy under `Log loss` criterion with max_depth=8: 0.689795918367347
Accuracy under `Gini coefficient` criterion with max_depth=8: 0.7040816326530612
Accuracy under `Information Gain` criterion with max_depth=12: 0.7040816326530612
Accuracy under `Log loss` criterion with max_depth=12: 0.710204081632653
Accuracy under `Gini coefficient` criterion with max_depth=12: 0.6938775510204082
Accuracy under `Information Gain` criterion with max_depth=16: 0.710204081632653
Accuracy under `Log loss` criterion with max_depth=16: 0.7081632653061225
Accuracy under `Gini coefficient` criterion with max_depth=16: 0.7061224489795919
Accuracy under `Information Gain` criterion with max_depth=25: 0.7020408163265306
Accuracy under `Log loss` criterion with max_depth=25: 0.7142857142857143
Accuracy under `Gini coefficient` criterion with max_depth=25: 0.7061224489795919
Accuracy under `Information Gain` criterion with max_depth=32: 0.7142857142857143
Accuracy under `Log loss` criterion with max_depth=32: 0.710204081632653
Accuracy under `Gini coefficient` criterion with max_depth=32: 0.7081632653061225
Accuracy under `Information Gain` criterion with max_depth=64: 0.7204081632653061
Accuracy under `Log loss` criterion with max_depth=64: 0.726530612244898
Accuracy under `Gini coefficient` criterion with max_depth=64: 0.7408163265306122
Accuracy under `Information Gain` criterion with max_depth=100: 0.7326530612244898
Accuracy under `Log loss` criterion with max_depth=100: 0.7346938775510204
Accuracy under `Gini coefficient` criterion with max_depth=100: 0.7326530612244898
```



Notice that log loss and entropy is not identical because the DecisionTreeClassifier splits the data randomly.

c)



d)

```
The information gain for feature "the" is 0.05263747727044332
The information gain for feature "hillary" is 0.0443445873158429
The information gain for feature "donald" is 0.049398847926479306
The information gain for feature "trumps" is 0.04500636360104682
The information gain for feature "clinton" is 0.011983306127556492
```

The information gain computed may be inconsistent because the split of data is random.

Problem 3

a)

$$\frac{\partial \mathcal{J}}{\partial w_j} = \frac{\partial}{\partial w_j}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t_i\right)^2\right) = \frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})x_j^{(i)}}{N}$$

$$\frac{\partial}{\partial w_j}\sum_{j=1}^{D}\alpha_j|w_j| = \begin{cases} 0, & w_j = 0 \\ \alpha_j, & w_j > 0 \\ -\alpha_j, & w_j < 0 \end{cases} \tag{1}$$

Let $\gamma > 0$ be the learning rate.

If $\underline{w_j > 0}$:

$$\frac{\partial \mathcal{J}_{reg}^{\alpha\beta}}{\partial w_j} = \frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})x_j^{(i)} + Na_j + N\beta_j w_j}{N}$$

And

$$\frac{\partial \mathcal{J}_{reg}^{\alpha\beta}}{\partial b} = \frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})}{N}$$

We have

$$w_j \leftarrow w_j - \gamma\frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})x_j^{(i)} + Na_j + N\beta_j w_j}{N}$$

$$b \leftarrow b - \gamma\frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})}{N}$$

If $\underline{w_j = 0}$:

$$\frac{\partial \mathcal{J}_{reg}^{\alpha\beta}}{\partial w_j} = \frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})x_j^{(i)} + N\beta_j w_j}{N}$$

And

$$\frac{\partial \mathcal{J}_{reg}^{\alpha\beta}}{\partial b} = \frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})}{N}$$

We have

$$w_j \leftarrow w_j - \gamma\frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})x_j^{(i)} + N\beta_j w_j}{N}$$

$$b \leftarrow b - \gamma\frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})}{N}$$

If $\underline{w_j < 0}$:

$$\frac{\partial \mathcal{J}_{reg}^{\alpha\beta}}{\partial w_j} = \frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})x_j^{(i)} - Na_j + N\beta_j w_j}{N}$$

And

$$\frac{\partial \mathcal{J}_{reg}^{\alpha\beta}}{\partial b} = \frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})}{N}$$

We have

$$w_j \leftarrow w_j - \gamma \frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})x_j^{(i)} - Na_j + N\beta_j w_j}{N}$$
$$b \leftarrow b - \frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})}{N}$$

Equivalently:

$$w_j \leftarrow w_j \frac{(1-\gamma\beta_j)}{N} - \gamma \frac{\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} + b - t^{(i)})x_j^{(i)} - Na_j}{N}$$

The regularization makes the weight w_j smaller, since we rescale w_j by $(1-\gamma\beta_j)/N$, so the weight decays.

b) Let $B = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_D \end{bmatrix}$. NI_D is the diagonal DxD matrix. We have:

$$\mathcal{J}_{reg}^{\beta} = \frac{1}{2N}\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} - t^{(i)}) + \frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2 \tag{2}$$

Then

$$\begin{aligned}
\frac{\partial \mathcal{J}_{reg}^{\beta}}{\partial w_j} &= \frac{1}{N}\sum_{i=1}^{N}(\sum_{j=1}^{D} w_j x_j^{(i)} - t^{(i)})x_j^{(i)} + \beta_j w_j \\
&= \frac{1}{N}\sum_{j'=1}^{D} w_{j'} x_{j'}^T x_{j'} - \frac{1}{N}x_j^T t + \beta_j w_j \\
&= \frac{1}{N}\sum_{j'=1}^{D}(w_{j'} x_{j'}^T x_{j'} + N\beta_j w_j) - \frac{1}{N}x_j^T t \\
&= \sum_{j'=1}^{D}\frac{1}{N}(X^T X + NI_D B)_{jj'} w_j' - \frac{1}{N}x_j^T t \\
&= \sum_{j'=1}^{D} A_{jj'} w_{j'} - c_j
\end{aligned}$$

Where $A_{jj'} = \frac{1}{N}(X^T X + NI_D B)_{jj'}$ and $c_j = \frac{1}{N}x_j^T t$.

c) From b), we get $A = \frac{1}{N}(X^T X + NI_D B)$ and $c = \frac{1}{N} X^T t$

$$\begin{aligned}
\frac{\partial \mathcal{J}_{reg}^{\beta}}{\partial w} &= \sum_{j=1}^{D} (\sum_{j'=1}^{D} A_{jj'} w_{j'} - c_j) \\
&= \sum_{j=1}^{D} (\sum_{j'=1}^{D} (A_{jj'} w_{j'}) - c \\
&= Aw - c \\
&= \frac{1}{N}(X^T X + NI_D B)w - \frac{1}{N} X^T t
\end{aligned}$$

Then

$$\begin{aligned}
(X^T X + NI_D B)w &= \frac{1}{N} X^T t \\
w &= (X^T X + NI_D B)^{-1} X^T t
\end{aligned}$$

Therefore, the closed form solution for w is $w = (X^T X + NI_D B)^{-1} X^T t = A^{-1} c$