CSC311 Homework 1

Guanyu Song

September 2021

1. (a) The graphs are shown below:

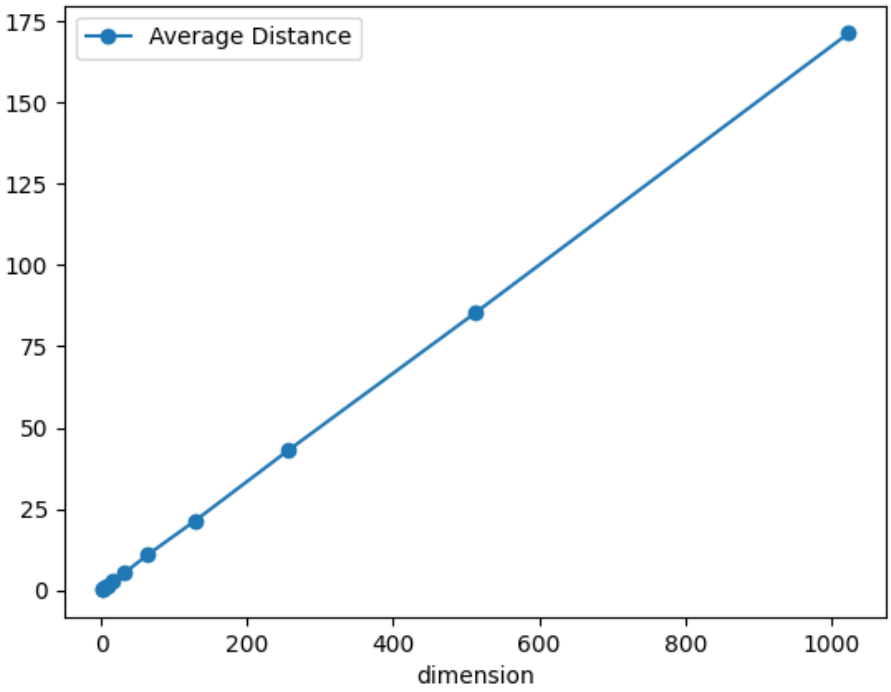

Figure 1: Average Distance vs Dimension. It's likely to be a linear relationship

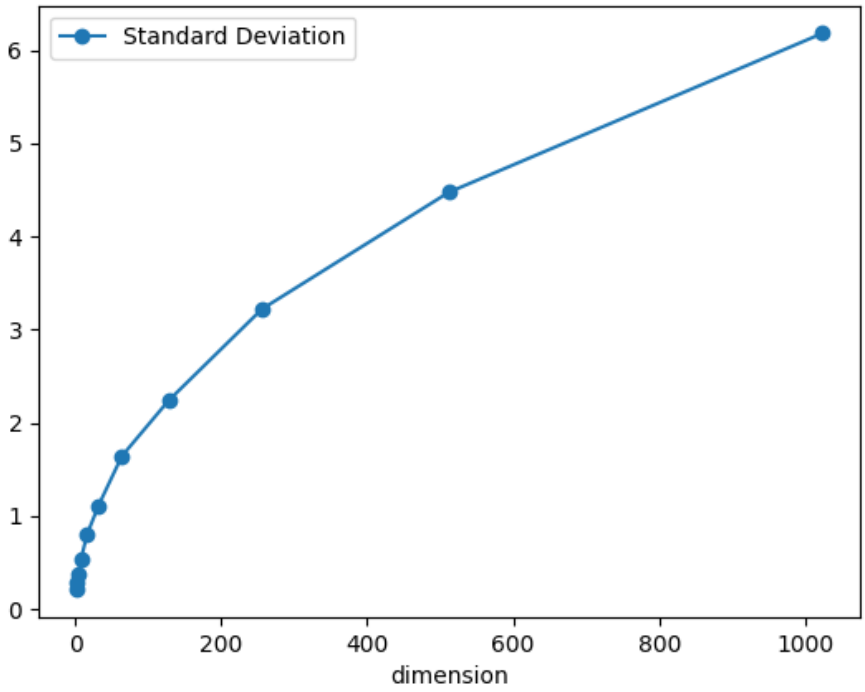

Figure 2: Standard Deviation. Likely to be a square root relationship

(b)

$$\begin{aligned} E(R) &= E(\sum_{i=1}^{d} Z_i) \\ &= \sum_{i=1}^{d} E(Z_i) \\ &= \sum_{i=1}^{d} \frac{1}{6} \\ &= \frac{d}{6} \end{aligned}$$

Since each dimension is independent of each others, the variance is linear.

$$\begin{aligned} Var(R) &= Var(\sum_{i=1}^{d} Z_i) \\ &= \sum_{i=1}^{d} Var(Z_i) \\ &= \sum_{i=1}^{d} \frac{7}{180} \\ &= \frac{7d}{180} \end{aligned}$$

Both E and Var are a linear function of d. It matches the graphs (in the graph, Std is a square root function of d, so $Var = Std^2$ = linear function of d).

2. (b)

```
=======================================
mode: gini
depth: 3, accuracy: 0.7061224489795919
depth: 5, accuracy: 0.7183673469387755
depth: 7, accuracy: 0.7224489795918367
depth: 9, accuracy: 0.7428571428571429
depth: 11, accuracy: 0.7448979591836735
depth: 13, accuracy: 0.7653061224489796
depth: 15, accuracy: 0.7693877551020408
=======================================
mode: entropy
depth: 3, accuracy: 0.6530612244897959
depth: 5, accuracy: 0.7081632653061225
depth: 7, accuracy: 0.7326530612244898
depth: 9, accuracy: 0.736734693877551
depth: 11, accuracy: 0.7489795918367347
depth: 13, accuracy: 0.7510204081632653
depth: 15, accuracy: 0.7448979591836735
```

Figure 3: The results

(c) The hyperparameters which achieved the highest validation accuracy are mode=gini, depth=15. Decision tree:

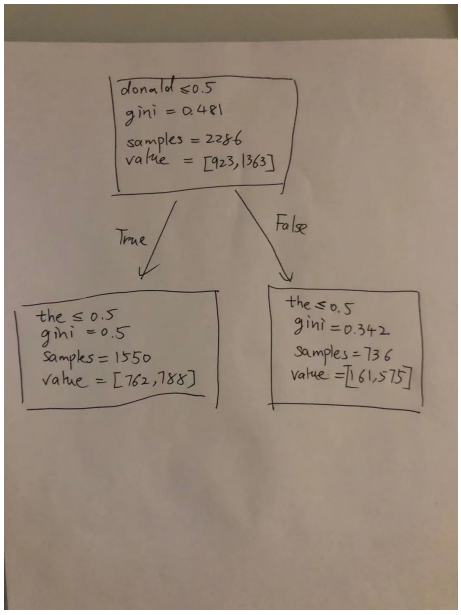

Figure 4: The first two layers of the decision tree

(d)

```
The information gain of a split at 'donald' is 0.051199764583789914
The information gain of a split at 'the' is 0.04755531139968236
The information gain of a split at 'hillary' is 0.03532590215455961
The information gain of a split at 'turnbull' is 0.011871533685102631
The information gain of a split at 'and' is 0.015417149865806085
The information gain of a split at 'trump' is 0.033572119865716954
```

Figure 5: The information gain of some keywords

3. (a)

$$\begin{aligned}\mathcal{J}_{reg}^{\beta}(w) &= \mathcal{J} + \mathcal{R} \\ &= \frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^(i))^2 + \frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2\end{aligned}$$

Calculate the partial derivatives of $\mathcal{J}$ and $\mathcal{R}$ separately:

$$\frac{\partial \mathcal{J}}{\partial w_j} = \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}(y^{(i)} - t^{(i)})$$

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

$$\frac{\partial \mathcal{R}}{\partial w_j} = \beta_j w_j$$

$$\frac{\partial \mathcal{R}}{\partial b} = 0$$

Because of the linearity of partial derivatives, we have:

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

We call it weight decay because the approaches the optimal value in the same way as an exponential decay function approaches its horizontal asymptote. The rate of change for w is in the form of $-aw$, while an exponential function $f(x) = e^{-a}$ also has $f'(x) = -af(x)$.

(b)

$$\begin{aligned}
\frac{\partial \mathcal{J}_{reg}^{\beta}}{\partial w_j} &= \frac{\partial \mathcal{J}}{\partial w_j} + \frac{\partial \mathcal{R}}{\partial w_j} \\
&= \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}(y^{(i)} - t^{(i)}) + \beta_j w_j \\
&= \frac{1}{N}(\sum_{i=1}^{N} x_j^{(i)}(\sum_{j'=1}^{D} w_{j'} x_{j'}^{(i)} - t^{(i)})) + \beta_j w_j \\
&= \frac{1}{N}(\sum_{j'=1}^{D}(\sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)}) + \beta_j w_j - \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} t^{(i)}
\end{aligned}$$

So $A_{jj'} = \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)}$ when $j' \neq j$ and $A_{jj'} = \frac{1}{N}(\sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)}) + \beta_j$ when $j' = j$.
$c_j = \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} t^{(i)}$.

(c) $\mathrm{A} = 1/\mathrm{N}\begin{bmatrix} \sum_{i=1}^{N} x_1^{(i)} x_1^{(i)} & \dots & \sum_{i=1}^{N} x_1^{(i)} x_D^{(i)} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} x_D^{(i)} x_1^{(i)} & \dots & \sum_{i=1}^{N} x_D^{(i)} x_D^{(i)} \end{bmatrix} + \begin{bmatrix} \beta_1 & 0 & \dots & 0 \\ 0 & \beta_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \beta_D \end{bmatrix}$, $\mathrm{c} = 1/\mathrm{N}\begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_1^{(i)} t^{(i)} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} x_D^{(i)} t^{(i)} \end{bmatrix}$,
$\mathrm{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_D \end{bmatrix}$.

Therefore, we can find w by solving $\mathrm{Aw} = \mathrm{c}$. $w = A^{-1}c$ if A is invertible.