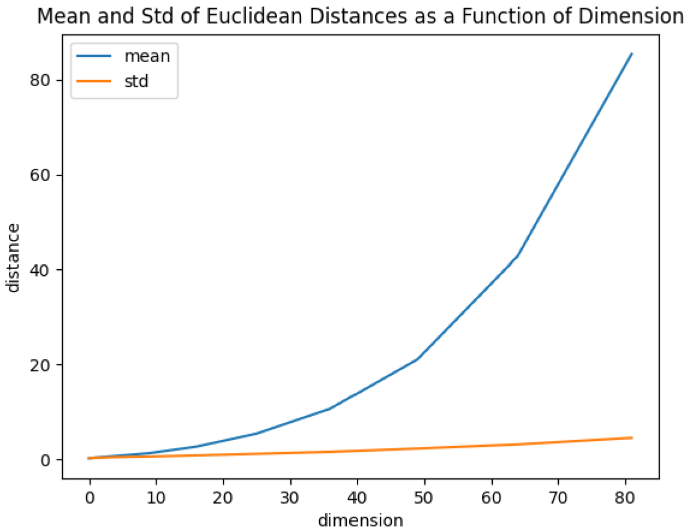
1 Nearest neighbours and the Curse of Dimensionality

a)



b) Deriving the analytical form of averaged distance and variance of distance:

$$\begin{aligned}
\mathbb{E}[R] &= \mathbb{E}[Z_1 + Z_2 + ... + Z_d] \\
&= \mathbb{E}\left[\sum_{i=0}^{d} Z_i\right] \\
&= \sum_{i=0}^{d} \mathbb{E}[Z_i] && \textbf{(by basic rule of expectation)} \\
&= \sum_{i=0}^{d} \frac{1}{6} = \frac{d}{6}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[R] &= \mathbb{E}[Z_1 + Z_2 + ... + Z_d] \\
&= \mathrm{Var}\left[\sum_{i=0}^{d} Z_i\right] \\
&= \sum_{i=0}^{d} \mathrm{Var}[Z_i] && \textbf{(by basic rule of variance } * \textbf{)} \\
&= \sum_{i=0}^{d} \frac{7}{180} = \frac{7d}{180}
\end{aligned}$$

2 Decision Trees

b) Top hyperparameters output: `[’entropy’, 7, 0.7287581699346405]`

c) Visualization of decision tree with highest scoring hyperparameters:

```
|--- donald <= 0.50
|   |--- hillary <= 0.50
|   |   |--- trumps <= 0.50
|   |   |   |--- truncated branch of depth 5
|   |   |--- trumps >  0.50
|   |   |   |--- truncated branch of depth 3
|   |--- hillary >  0.50
|   |   |--- saying <= 0.50
|   |   |   |--- class: 0
|   |   |--- saying >  0.50
|   |   |   |--- class: 1
|--- donald >  0.50
|   |--- the <= 0.50
|   |   |--- trump <= 0.50
|   |   |   |--- class: 1
|   |   |--- trump >  0.50
|   |   |   |--- truncated branch of depth 5
|   |--- the >  0.50
|   |   |--- trump <= 0.50
|   |   |   |--- class: 1
|   |   |--- trump >  0.50
|   |   |   |--- truncated branch of depth 5
```

3 Regularized Linear Regression

a) The regularized cost function $\mathcal{J}_{\text{reg}}^{\beta}$ can be formulated as so:

$$\mathcal{J}_{\text{reg}}^{\beta}(\mathbf{w}) = \frac{1}{2N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)^2 + \frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2$$

Furthermore, we can express the gradient descent update rules for $\mathcal{J}_{\text{reg}}^{\beta}$ with $\alpha > 0$ as:

$$\begin{aligned} w_j &\leftarrow w_j - \alpha\frac{\partial \mathcal{J}_{\text{reg}}^{\beta}}{\partial w_j} \\ b &\leftarrow b - \alpha\frac{\partial \mathcal{J}_{\text{reg}}^{\beta}}{\partial b} \end{aligned}$$

We will first derive $\frac{\partial \mathcal{J}_{\text{reg}}^{\beta}}{\partial w_j}$:

$$\begin{aligned} \frac{\partial \mathcal{J}_{\text{reg}}^{\beta}}{\partial w_j} &= \frac{\partial}{\partial w_j}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)^2 + \frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2\right) \\ &= \frac{\partial}{\partial w_j}\frac{1}{2N}\sum_{i=1}^{N}\left(\sum_{j'} w_{j'}x_{j'}^{(i)} + b - t^{(i)}\right)^2 + \frac{\partial}{\partial w_j}\frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}\left(\sum_{j'} w_{j'}x_{j'}^{(i)} + b - t^{(i)}\right) + \sum_{j=1}^{D}\beta_j w_j \\ &= \frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}(y^{(i)} - t^{(i)}) + \sum_{j=1}^{D}\beta_j w_j \end{aligned}$$

Next we will derive $\frac{\partial \mathcal{J}_{\text{reg}}^{\beta}}{\partial b}$:

$$\begin{aligned} \frac{\partial \mathcal{J}_{\text{reg}}^{\beta}}{\partial b} &= \frac{\partial}{\partial b}\left(\frac{1}{2N}\sum_{i=1}^{N}\left(y^{(i)} - t^{(i)}\right)^2 + \frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2\right) \\ &= \frac{\partial}{\partial b}\frac{1}{2N}\sum_{i=1}^{N}\left(\sum_{j'} w_{j'}x_{j'}^{(i)} + b - t^{(i)}\right)^2 + \frac{\partial}{\partial b}\frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{j'} w_{j'}x_{j'}^{(i)} + b - t^{(i)}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)}) \end{aligned}$$

We add regularization penalties to encourage the weights to stay small and prevent the gradient descent from exploding. With each iteration we minimize the weights towards the direction of steepest ascent of the gradient with respect to the loss function and over time they will decay to 0 as the weights decrease by very small factors.

b) We will derive a system of linear equations of the following form for $\mathcal{J}_{\text{reg}}^{\beta}$:

$$\begin{aligned}
\frac{\partial \mathcal{J}_{\text{reg}}^{\beta}}{\partial w_j} &= \sum_{j'=1}^{D} A_{jj'} w_{j'} - c_j = 0 \\
\implies \frac{\partial \mathcal{J}_{\text{reg}}^{\beta}}{\partial w_j} &= \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)} \left(\sum_{j'} w_{j'} x_{j'}^{(i)} - t^{(i)} \right) + \sum_{j=1}^{D} \beta_j w_j \\
&= \frac{1}{N} \sum_{j'=1}^{D} \left(\sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)} \right) w_{j'} - \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)} t^{(i)} + \sum_{j=1}^{D} \beta_j w_j \\
&= \frac{1}{N} \sum_{j'=1}^{D} \left(\left(\sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)} \right) w_{j'} + \beta_{j'} w_{j'} \right) - \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)} t^{(i)} \\
&= \sum_{j'=1}^{D} \frac{1}{N} \left(\left(\sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)} \right) + \beta_{j'} \right) w_{j'} - \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)} t^{(i)}
\end{aligned}$$

We have obtained the following formulas for $A_{jj'}$ and c_j:

$$\begin{aligned}
A_{jj'} &= \frac{1}{N} \left(\left(\sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)} \right) + \beta_{j'} \right) \\
c_j &= \frac{1}{N} \sum_{i=1}^{N} x_j^{(i)} t^{(i)}
\end{aligned}$$

c) We will derive a closed-form solution to the parameter $\mathbf{w}$. We will represent our inputs as an $N \times D$ matrix $\mathbf{X}$ where N is the number of training examples and D is the dimension. We can express $\mathbf{A}$ and $\mathbf{c}$:

$$\begin{aligned}
\mathbf{A} &= \frac{1}{N}(\mathbf{X}^{\top}\mathbf{X} + \beta) \\
\mathbf{c} &= \frac{1}{N}\mathbf{X}^{\top}\mathbf{t}
\end{aligned}$$

The solution to the linear system $\mathbf{Aw} = \mathbf{c}$ is given by $\mathbf{w} = \mathbf{A}^{-1}\mathbf{c}$ (assuming $\mathbf{A}$ is invertible), thus we have a formula for $\mathbf{w}$:

$$\mathbf{w} = (\mathbf{X}^{\top}\mathbf{X} + \beta)^{-1}\mathbf{X}^{\top}\mathbf{t}$$