CSC311-hw1

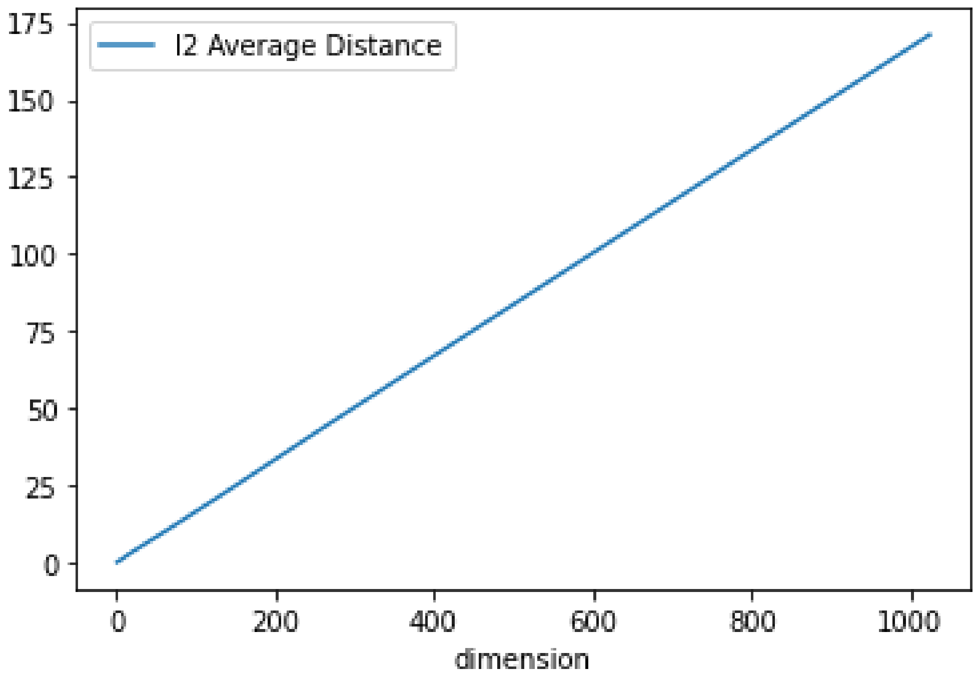
siwei tang

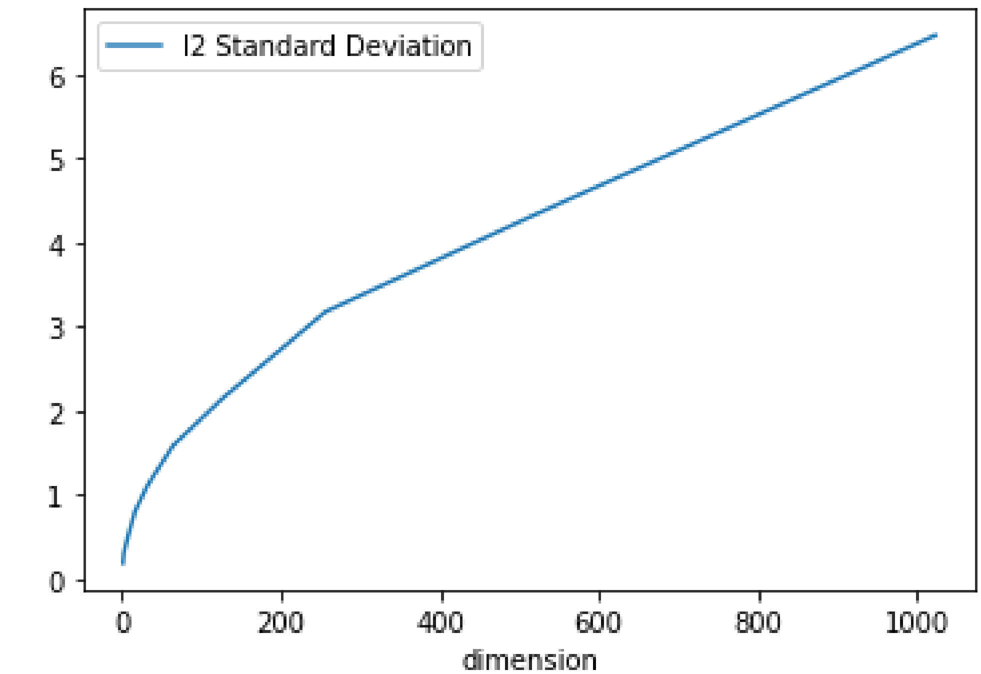
September 2022

1 Nearest Neighbours and the Curse of Dimensionality

(a)

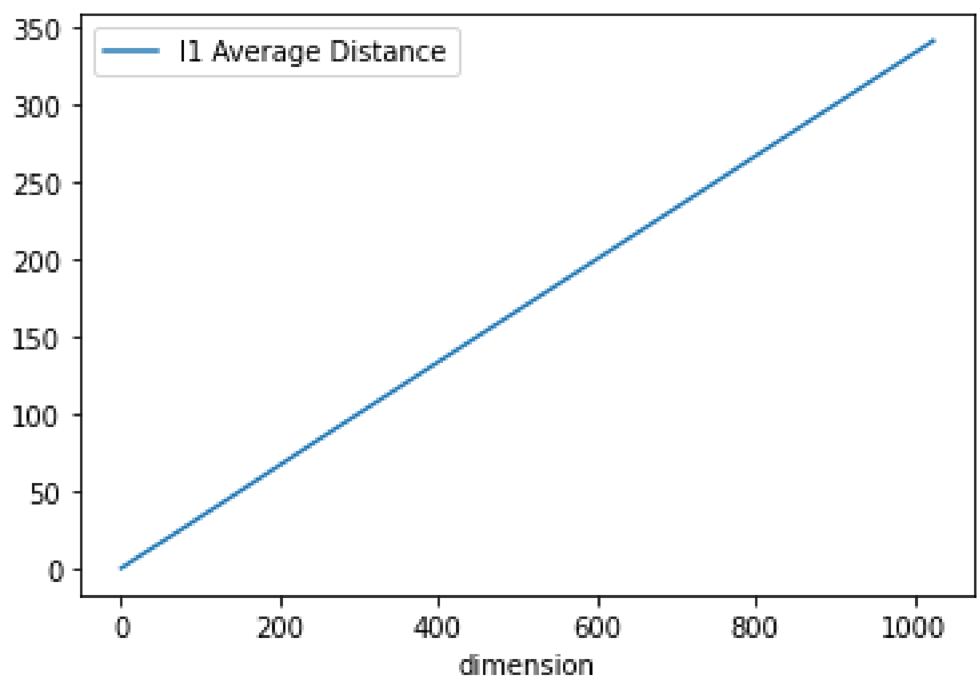
(i)



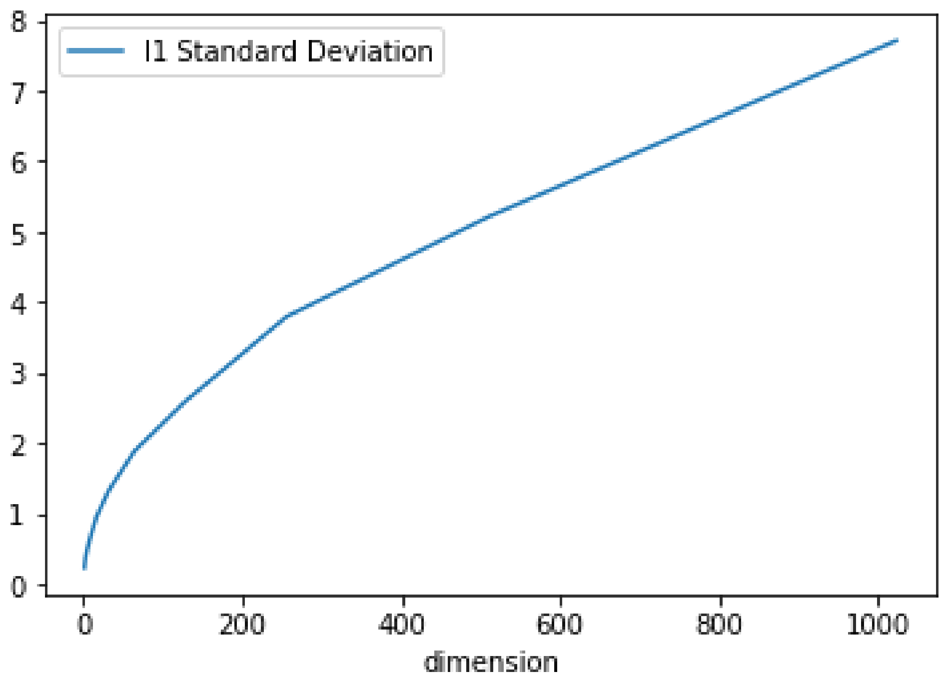


l2 Standard Deviation
0
1
2
3
4
5
6
0
200
400
600
800
1000
dimension

(ii)



l1 Average Distance
0
50
100
150
200
250
300
350
0
200
400
600
800
1000
dimension



(b)

$$
\begin{aligned}
\mathbb{E}[R] &= \mathbb{E}[Z_1 + ... + Z_d] \\
&= \mathbb{E}\left[\sum_{i=0}^{d} Z_i\right] \\
&= \sum_{i=1}^{d} \mathbb{E}[Z_i] \\
&= \sum_{i=1}^{d} \frac{1}{6} \\
&= \frac{d}{6}
\end{aligned}
$$

$$
\begin{aligned}
Var[R] &= Var[Z_1 + ..._+ Z_d] \\
&= Var\left[\sum_{i=0}^{d} Z_i\right] \\
&= \sum_{i=1}^{d} Var[Z_i] \\
&= \sum_{i=1}^{d} \frac{7}{180} \\
&= \frac{7d}{180}
\end{aligned}
$$

(c)

(i)

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

(ii)

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

(iii)

As d goes to $\infty, \mathbb{P}$(E) approaches 1
Markov's inequality suggests that probability of euclidean distance far away from expected value is unlikely, so that the distance between two point is likely to be similar.

2 Decision Tree

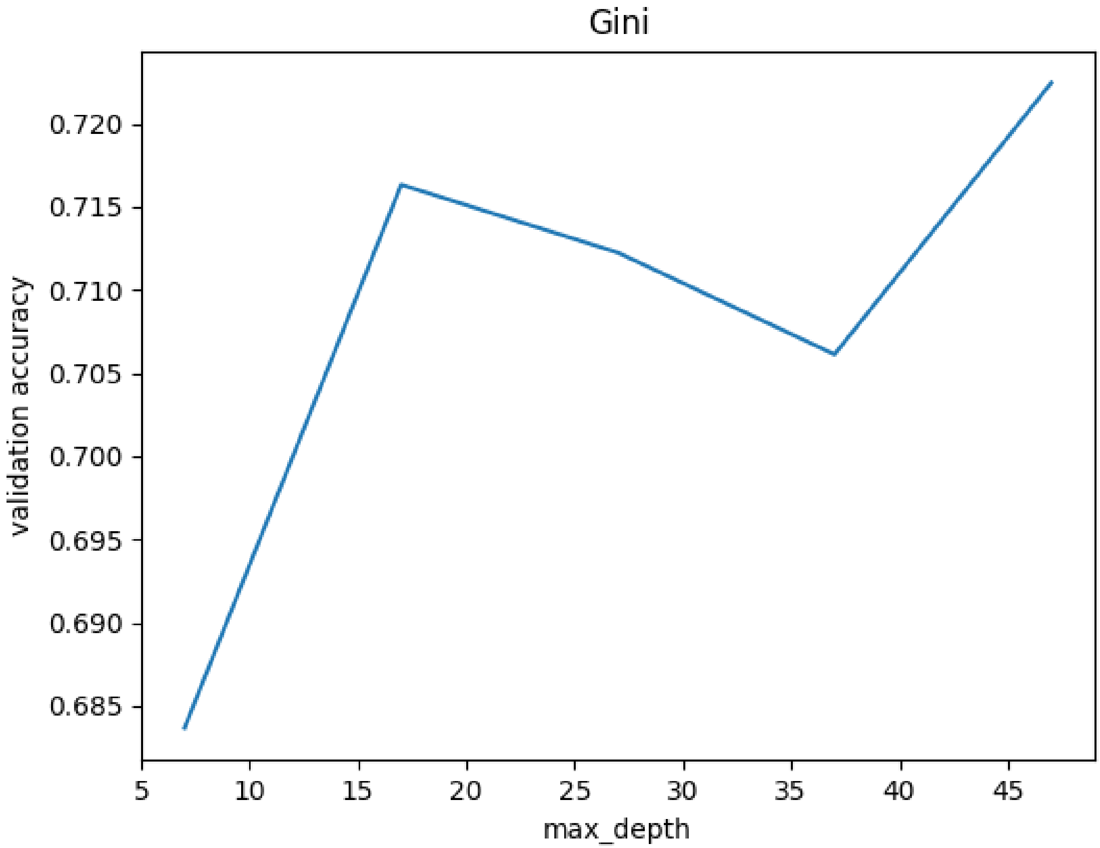
(a)

See python file

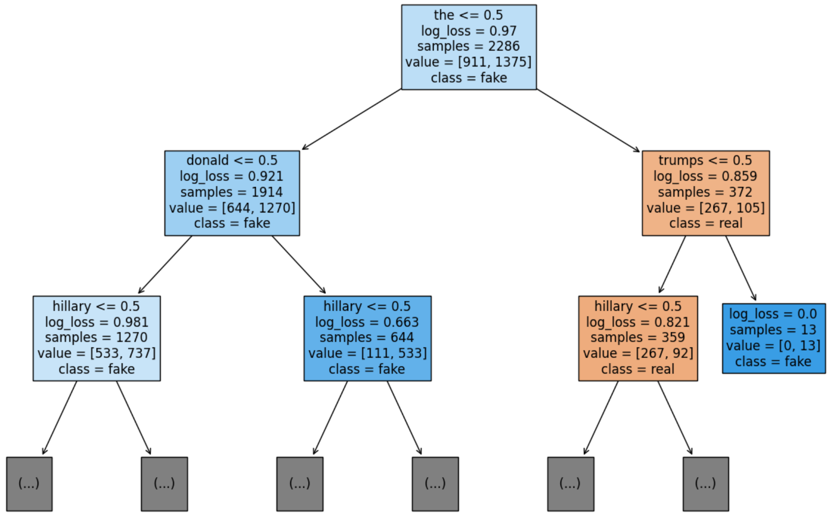
(b)

max depth	gini	log loss	entropy
7	0.6816	0.6857	0.6857
17	0.7041	0.7204	0.7265
27	0.7286	0.7347	0.7265
37	0.7102	0.7245	0.7163
47	0.7122	0.7245	0.7286

It is for Gini criteria.



(c)



(d)

The top most split word is 'the', the information gain is 0.058909111805979575. For other words, see the below picture.

```
The information gain of a split at 'donald' is 0.04773248640015862
The information gain of a split at 'the' is 0.058909111805979575
The information gain of a split at 'hillary' is 0.034264881132076846
The information gain of a split at 'trumps' is 0.045315207919737845
The information gain of a split at 'de' is 0.0015217665211054569
The information gain of a split at 'election' is 0.00028016755682347405
The information gain of a split at 'are' is 0.01367884585470175
```

3 Regularized Linear Tree

(a)

$$w_j > 0, J_{reg}^{\alpha\beta}(w) = \frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})^2 + \sum_{j=1}^{D}\alpha_j w_j + \frac{1}{2}\sum_{j=1}^{D}\beta_j {w_j}^2$$

$$w_j = 0, J_{reg}^{\alpha\beta}(w) = \frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})^2 + \frac{1}{2}\sum_{j=1}^{D}\beta_j {w_j}^2$$

$$w_j < 0, J_{reg}^{\alpha\beta}(w) = \frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})^2 - \sum_{j=1}^{D}\alpha_j w_j + \frac{1}{2}\sum_{j=1}^{D}\beta_j {w_j}^2$$

$$\begin{aligned}
J &= \frac{1}{2N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})^2 \\
&= \frac{1}{2N}\sum_{i=1}^{N}(w^T x^{(i)} + b - t^{(i)})^2 \\
\Rightarrow \frac{\partial J}{\partial w_j} &= \frac{\partial J}{\partial y^{(i)}}\frac{\partial y^{(i)}}{\partial w_j} = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)}) \cdot x_j^{(i)} \\
\frac{\partial J}{\partial b} &= \frac{\partial J}{\partial y^{(i)}}\frac{\partial y^{(i)}}{\partial b} = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})
\end{aligned}$$

If $w_j > 0$:

$$\begin{aligned}
R &= \sum_{j=1}^{D}\alpha_j w_j + \frac{1}{2}\sum_{j=1}^{D}\beta_j {w_j}^2 = \sum_{j=1}^{D}(\alpha_j w_j + \frac{1}{2}\beta_j {w_j}^2) \\
\Rightarrow \frac{\partial R}{\partial w_j} &= \alpha_j + \beta_j w_j,\ \frac{\partial R}{\partial b} = 0 \\
w_j &\leftarrow w_j - \theta\frac{\partial J_{reg}^{\alpha\beta}}{\partial w_j} \\
&= w_j - \theta \cdot (\frac{\partial J}{\partial w_j} + \frac{\partial R}{\partial w_j}) \\
&= w_j - \theta(\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)}) \cdot x_j^{(i)} + \alpha_j + \beta_j w_j) \\
&= (1 - \theta\beta_j)w_j - \frac{\alpha}{N}\sum_{i=1}^{N}(x_j^{(i)}(y^{(i)} - t^{(i)})) - \theta\alpha_j
\end{aligned}$$

$$
\begin{aligned}
b &\leftarrow b - \theta \frac{\partial J_{reg}^{\alpha\beta}}{\partial b} \\
&= b - \theta \cdot (\frac{\partial J}{\partial b} + \frac{\partial R}{\partial b}) \\
&= b - \theta \cdot (\frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)}) + 0) \\
&= b - \frac{\theta}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)})
\end{aligned}
$$

If $w_j = 0:$

$$
\begin{aligned}
R &= \frac{1}{2} \sum_{j=1}^{D} \beta_j {w_j}^2 \\
\Rightarrow \frac{\partial R}{\partial w_j} &= \beta_j w_j, \ \frac{\partial R}{\partial b} = 0 \\
w_j &\leftarrow w_j - \theta \frac{\partial J_{reg}^{\alpha\beta}}{\partial w_j} \\
&= w_j - \theta \cdot (\frac{\partial J}{\partial w_j} + \frac{\partial R}{\partial w_j}) \\
&= w_j - \theta(\frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)}) \cdot x_j^{(i)} + \beta_j w_j) \\
&= (1 - \theta\beta_j) w_j - \frac{\theta}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)}) \cdot x_j^{(i)}
\end{aligned}
$$

$$
\begin{aligned}
b &\leftarrow b - \theta \frac{\partial J_{reg}^{\alpha\beta}}{\partial b} \\
&= b - \theta \cdot (\frac{\partial J}{\partial b} + \frac{\partial R}{\partial b}) \\
&= b - \theta \cdot (\frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)}) + 0) \\
&= b - \frac{\theta}{N} \sum_{i=1}^{N} (y^{(i)} - t^{(i)})
\end{aligned}
$$

If $w_j < 0$:

$$
\begin{aligned}
R &= -\sum_{j=1}^{D}\alpha_j w_j + \frac{1}{2}\sum_{j=1}^{D}\beta_j {w_j}^2 = \sum_{j=1}^{D}(\frac{1}{2}\beta_j {w_j}^2 - \alpha_j w_j) \\
\Rightarrow \frac{\partial R}{\partial w_j} &= \beta_j w_j - \alpha_j,\ \frac{\partial R}{\partial b} = 0 \\
w_j &\leftarrow w_j - \theta\frac{\partial J_{reg}^{\alpha\beta}}{\partial w_j} \\
&= w_j - \theta\cdot(\frac{\partial J}{\partial w_j} + \frac{\partial R}{\partial w_j}) \\
&= w_j - \theta(\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})\cdot x_j^{(i)} + \beta_j w_j - \alpha_j) \\
&= (1-\theta\beta_j)w_j - \frac{\alpha}{N}\sum_{i=1}^{N}(x_j^{(i)}(y^{(i)} - t^{(i)})) + \theta\alpha_j
\end{aligned}
$$

$$
\begin{aligned}
b &\leftarrow b - \theta\frac{\partial J_{reg}^{\alpha\beta}}{\partial b} \\
&= b - \theta\cdot(\frac{\partial J}{\partial b} + \frac{\partial R}{\partial b}) \\
&= b - \theta\cdot(\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)}) + 0) \\
&= b - \frac{\theta}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})
\end{aligned}
$$

Since the parameter before w is smaller than 1 so that means weight becomes smaller, that's why we call it weight decay.

(b)

$$
\begin{aligned}
\frac{\partial J_{reg}^{\beta}}{\partial w_j} &= \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}(y^{(i)} - t^{(i)}) + \beta_j w_j \\
&= \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}(\sum_{j'=1}^{D} w_{j'} x_{j'}^{(i)} - t^{(i)}) + \beta_j w_j \\
&= \sum_{j'=1}^{D} \frac{1}{N}(\sum_{i=1}^{N}(x_j^{(i)} x_{j'}^{(i)} w_{j'}) + N\beta_j w_j) - \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} t^{(i)} \\
&= \sum_{j'=1}^{D} \frac{1}{N}\sum_{i=1}^{N}(x_j^{(i)} x_{j'}^{(i)} w_{j'} + \beta_j w_j) - \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} t^{(i)} \\
\Rightarrow A_{jj'} &= \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)} \text{ , if j' } \neq \text{ j} \\
A_{jj'} &= \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)} + \beta_j \text{ , if j' = j} \\
c_j &= \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} t^{(i)}
\end{aligned}
$$

(c)

$$
\begin{aligned}
A &= \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} x_1^{(i)} x_1^{(i)} & \dots & \frac{1}{N}\sum_{i=1}^{N} x_1^{(i)} x_D^{(i)} \\ \vdots & \ddots & \vdots \\ \frac{1}{N}\sum_{i=1}^{N} x_D^{(i)} x_1^{(i)} & \dots & \frac{1}{N}\sum_{i=1}^{N} x_D^{(i)} x_D^{(i)} \end{pmatrix} + \begin{pmatrix} \beta_1 & 0 & \dots & 0 \\ 0 & \beta_2 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \beta_D \end{pmatrix} \\
&= \frac{1}{N} \begin{pmatrix} \sum_{i=1}^{N} x_1^{(i)} x_1^{(i)} & \dots & \sum_{i=1}^{N} x_1^{(i)} x_D^{(i)} \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} x_D^{(i)} x_1^{(i)} & \dots & \sum_{i=1}^{N} x_D^{(i)} x_D^{(i)} \end{pmatrix} + \begin{pmatrix} \beta_1 & 0 & \dots & 0 \\ 0 & \beta_2 & \dots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \beta_D \end{pmatrix} \\
&= \frac{1}{N} X^T X + \beta I
\end{aligned}
$$

$$
c = \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N} x_1^{(i)} t^{(i)} \\ \vdots \\ \frac{1}{N}\sum_{i=1}^{N} x_D^{(i)} t^{(i)} \end{pmatrix} = \frac{1}{N} \begin{pmatrix} \sum_{i=1}^{N} x_1^{(i)} t^{(i)} \\ \vdots \\ \sum_{i=1}^{N} x_D^{(i)} t^{(i)} \end{pmatrix} = \frac{1}{N} X^T t
$$

$$
w = \begin{pmatrix} w_1 \\ \vdots \\ w_D \end{pmatrix}
$$

$$
\begin{aligned}
\Rightarrow Aw &= c \\
\Rightarrow w &= A^{-1} c \\
&= (X^T X + \beta I)^{-1} X^T t
\end{aligned}
$$