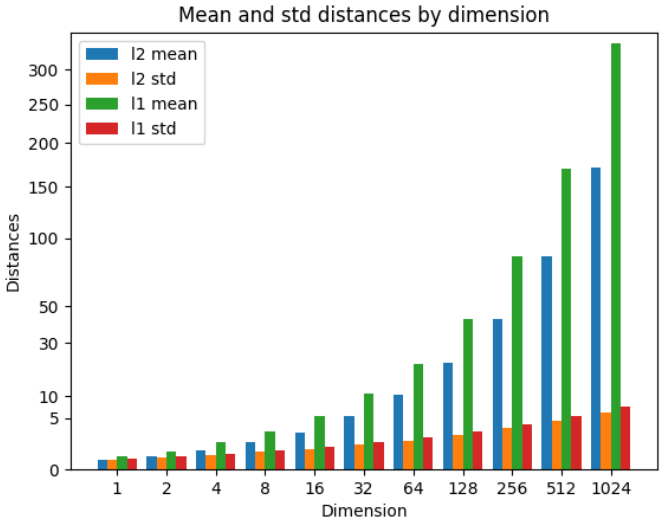
CSC311 A1 Writeup

Richard Yan

September 2022

Q1: Nearest Neighbour and the Curse of Dimensionality

(a) The plot shows that as dimension goes higher, the mean distance (either l1 or l2) grows, meaning that points eventually are getting farther and farther away from each other, while the standard deviation (either l1 or l2) doesn't grow as much (as I'm using an exponential scaling on the y-axis, the l1 standard deviation for 1024 dimension points is approximately 6), meaning that points are mostly "equal" distance from each other while being farther apart.



(b) Since $\mathbb{E}(Z_i) = \frac{1}{6}$ for every $1 \leq i \leq d$, then we have $\mathbb{E}(R) = \mathbb{E}(Z_1) + \cdots + \mathbb{E}(Z_d) = \frac{d}{6}$. Similarly, since the two points X and Y are chosen independently, so are their coordinates, which means Z_i and Z_j are all independent for all $1 \leq i, j \leq d$, $i \neq j$. Hence, $\text{Var}(R) = \text{Var}(Z_1 + \cdots + Z_d) = \text{Var}(Z_1) + \cdots + \text{Var}(Z_d) = \frac{7d}{180}$.

(c) (i) The event E is $|R - \mathbb{E}(R)| \geq d$, where R is defined in part b.

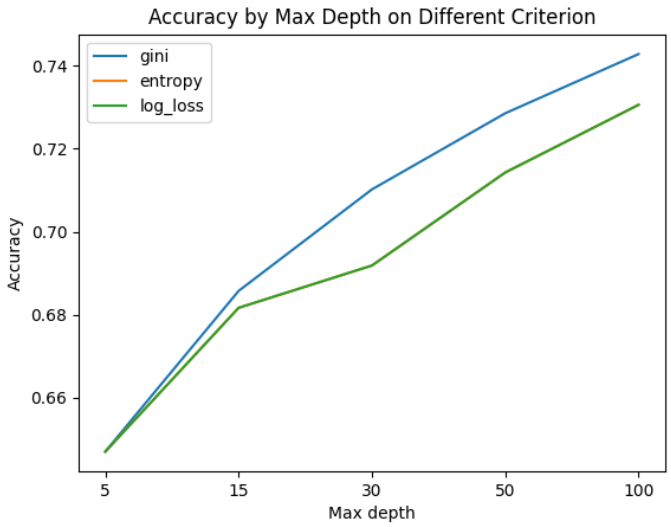
(ii)

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

(iii) As d (dimension) goes to ∞, we can see that $\mathbb{P}(E) \leq 0$, which means $\mathbb{P}(E) = 0$, thus equivalently saying that R, the distance between two random points, are never going to be far away from its expected value (average distance between random pairs of points), or in simpler words, pairwise distance between points are getting closer as dimension gets higher.

Q2: Decision Trees

(b)



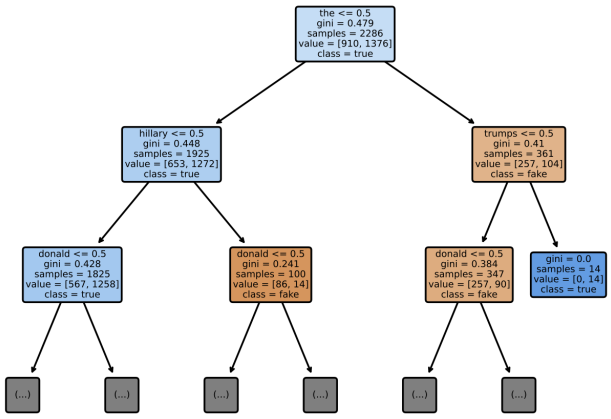
With gini criteria and max depth of 5: 0.6469387755102041
With gini criteria and max depth of 15: 0.6857142857142857
With gini criteria and max depth of 30: 0.710204081632653
With gini criteria and max depth of 50: 0.7285714285714285
With gini criteria and max depth of 100: 0.7428571428571429
With entropy criteria and max depth of 5: 0.6469387755102041
With entropy criteria and max depth of 15: 0.6816326530612244
With entropy criteria and max depth of 30: 0.6918367346938775
With entropy criteria and max depth of 50: 0.7142857142857143
With entropy criteria and max depth of 100: 0.7306122448979592

With log_loss criteria and max depth of 5: 0.6469387755102041
With log_loss criteria and max depth of 15: 0.6816326530612244
With log_loss criteria and max depth of 30: 0.6918367346938775
With log_loss criteria and max depth of 50: 0.7142857142857143
With log_loss criteria and max depth of 100: 0.7306122448979592

Notice that the entropy criterion and log_loss criterion are having the same accuracies because they are having the same underlining calculation, and I have fixed the randomness by introducing a random seed to control how the data is split (among various run) and how the split is chosen by the classifier (it tends to be different even with same training and validation data).

(c)

Decision Tree with Gini Criterion and Max Depth of 100



(d)

The information gain given the split on 'the': 0.05487510039397081 (topmost split)
The information gain given the split on 'hillary': 0.03926602747321262 (second topmost split)
The information gain given the split on 'trump': 0.03806974891343706 (second topmost split)
The information gain given the split on 'daily': 0.005635949130147977 (random split)
The information gain given the split on 'chomsky': 0.000581524206961781 (random split)

We can see that with the topmost split, we get the highest information gain (which tells why the classifier choose it as the topmost split), where both second topmost split having reasonable high information gain, while the random splits produce much less information gain (which tells why they are not on the top layers).

Q3: Regularized Linear Regression

(a) We can first rewrite our function $\mathcal{J}_{\text{reg}}^{\alpha\beta}(\mathbf{w})$ as

$$\mathcal{J}_{\text{reg}}^{\alpha\beta}(\mathbf{w}) = \frac{1}{2N}\sum_{i=1}^{N}(\mathbf{w}\mathbf{x}^{(i)} + b - t^{(i)})^2 + \sum_{j=1}^{D}\alpha_j|w_j| + \frac{1}{2}\sum_{j=1}^{D}\beta_j w_j^2$$

So let's look at this term by term.

The first term when taken derivative with respective to w_j will be:

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

The second term (since regularly not differentiable at 0) can be written out as piecewise function (and thus we can define the derivatives with respect to w_j piecewisely) as such: $\sum_{j=1}^{D}\alpha_j|w_j| = \begin{cases}-\sum_{j=1}^{D}\alpha_j w_j & w_j < 0\\ 0 & w_j = 0\\ \sum_{j=1}^{D}\alpha_j w_j & w_j > 0\end{cases}$

So derivative with respect to w_j is:

$\begin{cases}-\alpha_j & w_j < 0\\ 0 & w_j = 0\\ \alpha_j & w_j > 0\end{cases}$

Finally, the derivative of the last term with respect to w_j is:

$$\frac{1}{2}\cdot 2\beta_j w_j = \beta_j w_j$$

So put it together, for all j, we have:

$$\frac{\partial\mathcal{J}_{\text{reg}}^{\alpha\beta}}{\partial w_j} = \begin{cases}\frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}(y^{(i)} - t^{(i)}) - \alpha_j + \beta_j w_j & w_j < 0\\ \frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}(y^{(i)} - t^{(i)}) + \beta_j w_j & w_j = 0\\ \frac{1}{N}\sum_{i=1}^{N}x_j^{(i)}(y^{(i)} - t^{(i)}) + \alpha_j + \beta_j w_j & w_j > 0\end{cases}$$

Also for the constant bias term b, we can see that it only appears in the first term (inside $y^{(i)}$), so we have the derivative as:

$$\begin{aligned}\frac{\partial \mathcal{J}_{\text{reg}}^{\alpha\beta}}{\partial b} &= \frac{1}{2N}\sum_{i=1}^{N} 2(\mathbf{w}\mathbf{x}^{(i)} + b - t^{(i)}) \\ &= \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)})\end{aligned}$$

So we have that:
If $w_j > 0$:

$$\begin{aligned}w_j &\leftarrow w_j - (\frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}(y^{(i)} - t^{(i)}) + \alpha_j + \beta_j w_j) \\ b &\leftarrow b - (\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)}))\end{aligned}$$

If $w_j = 0$:

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

If $w_j < 0$:

$$\begin{aligned}w_j &\leftarrow w_j - (\frac{1}{N}\sum_{i=1}^{N} x_j^{(i)}(y^{(i)} - t^{(i)}) - \alpha_j + \beta_j w_j) \\ b &\leftarrow b - (\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - t^{(i)}))\end{aligned}$$

(b) From part (a) we can get that

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

Since we care about making it into linear equations in terms of each weight

variable, then let's take out the $w_{j'}$:

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

Notice that we have a term $\beta_j w_j$ in the end, which will affect the $w_{j'}$ only when $j^{'} = j$, hence we can have that:

$$\frac{\partial \mathcal{J}_{\text{reg}}^{\beta}}{\partial w_j} = \sum_{j'=1}^{D} A_{jj'} w_{j'} - c_j = 0$$

for

$$A_{jj'} = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)} & j^{'} \neq j \\ \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} x_{j'}^{(i)} + \beta_j & j^{'} = j \end{cases}$$
$$c_j = \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} t^{(i)}$$

(c) We can observe that when $j^{'} \neq j$, the matrix entries (i.e. those $A_{jj'}$) will give a matrix $\frac{1}{N}\mathbf{X}^T\mathbf{X}$. But we also need to consider the case that $j^{'} = j$, in which it's only affecting the diagonal of the matrix with adding a term. Hence we could have

$$\mathbf{A} = \frac{1}{N}\mathbf{X}^T\mathbf{X} + \text{diag}(\boldsymbol{\beta})$$

where $\text{diag}(\boldsymbol{\beta})$ means the diagonal matrix with entries (j, j) as β_j. And

$$\mathbf{c} = \frac{1}{N}\mathbf{X}^T\mathbf{t}$$

So the solution of the linear system $\mathbf{A}\mathbf{w} - \mathbf{c} = 0$ is

$$\begin{aligned} \mathbf{w} &= \mathbf{A}^{-1}\mathbf{c} \\ &= (\frac{1}{N}\mathbf{X}^T\mathbf{X} + \text{diag}(\boldsymbol{\beta}))^{-1}\frac{1}{N}\mathbf{X}^T\mathbf{t} \\ &= N(\mathbf{X}^T\mathbf{X} + \text{diag}(\boldsymbol{\beta}))^{-1}\frac{1}{N}\mathbf{X}^T\mathbf{t} \\ &= (\mathbf{X}^T\mathbf{X} + \text{diag}(\boldsymbol{\beta}))^{-1}\mathbf{X}^T\mathbf{t} \end{aligned}$$