CSC2515 HW1

Bingzhang Zhu

October 06, 2021

1 Nearest Neighbours and the Curse of Dimensionality

1.1 (a)

As we know that in a Uniform Distribution, where a continuous random variable (RV) that has equally likely outcomes over the domain, $a < x < b$, the mean is $\mu = \frac{a+b}{2}$ and the standard deviation is $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. The probability density function (PDF) is $f(x) = \frac{1}{b-a}$ for $a < x < b$. And thus, in this question, we get $E(X) = E(Y) = \frac{1}{2}$, $Var(X) = Var(Y) = \frac{1}{12}$, and the PDF for X, $f(x) = 1$ for $0 < x < 1$, the PDF for Y, $f(y) = 1$ for $1 < x < 1$.
The expectation of Z is:

$$\begin{aligned} E(Z) &= E(|X - Y|^2) \\ &= E(X^2 - 2XY + Y^2) \\ &= E(X^2) - 2E(XY) + E(Y^2) \end{aligned} \tag{1}$$

Since X and Y are independent random variables, $E(XY) = E(X)E(Y)$, we could get from above equation that:

$$E(Z) = E(X^2) - 2E(X)E(Y) + E(Y^2) \tag{2}$$

And from $Var(X) = E(X^2) - E(X)^2$, we get that:

$$\begin{aligned} E(X^2) &= Var(X) + E(X)^2 \\ &= \frac{1}{12} + (\frac{1}{2})^2 \\ &= \frac{1}{3} \end{aligned} \tag{3}$$

And we get $E(Y^2) = \frac{1}{3}$ in the same way.
Therefore, based on what we have now, we conclude that:

$$E(Z) = \frac{1}{3} - 2 \times \frac{1}{2} \times \frac{1}{2} + \frac{1}{3} = \frac{1}{6} \tag{4}$$

For the Variance of Z, based on $Var(X) = E(X^2) - E(X)^2$, we get:

$$\begin{aligned} Var(Z) = E(Z^2) - E(Z)^2 &= E(|X - Y|^4) - \frac{1}{36} \\ &= E(X^4 - 4X^3Y + 6X^2Y^2 - 4XY^3 + Y^4) - \frac{1}{36} \\ &= E(X^4) - 4E(X^3Y) + 6E(X^2Y^2) - 4E(XY^3) + E(Y^4) - \frac{1}{36} \end{aligned} \tag{5}$$

Since X and Y are independent random variables, we could know that X^3 and Y, X^2 and Y^2, X and Y^3 are also independent from each other. Therefore, based on equation (5) and all the value we got above, we could conclude that:

$$\begin{aligned}
Var(Z) &= E(X^4) - 4E(X^3)E(Y) + 6E(X^2)E(Y^2) - 4E(X)E(Y^3) + E(Y^4) - \frac{1}{36} \\
&= \int_0^1 x^4 f(x)\ dx - 2\int_0^1 x^3 f(x)\ dx + 6 \times \frac{1}{3} \times \frac{1}{3} - 2 \times \int_0^1 y^3 f(y)\ dy + \int_0^1 y^4 f(y)\ dy - \frac{1}{36} \\
&= \int_0^1 x^4\ dx - 2\int_0^1 x^3\ dx + \frac{2}{3} - 2\int_0^1 y^3\ dy + \int_0^1 y^4\ dy - \frac{1}{36} \\
&= \frac{1}{5}x^5\Big|_0^1 - 2 \times \frac{1}{4}x^4\Big|_0^1 + \frac{2}{3} - 2 \times \frac{1}{4}y^4\Big|_0^1 + \frac{1}{5} \times y^5\big|_0^1 - \frac{1}{36} \\
&= \frac{1}{5} - \frac{1}{2} + \frac{2}{3} - \frac{1}{2} + \frac{1}{5} - \frac{1}{36} \\
&= \frac{7}{180}
\end{aligned} \tag{6}$$

1.2 (b)

The expectation is :

$$\begin{aligned}
E(||X - Y||_2^2) &= E(R) \\
&= E(Z_1 + Z_2 + ... + Z_d) \\
&= E(Z_1) + E(Z_2) + ... + E(Z_d) \\
&= d \times E(Z) = d \times \frac{1}{6} = \frac{d}{6}
\end{aligned} \tag{7}$$

And since $Z_1, Z_2, ..., Z_d$ are in different dimensionality, we regard them as in-dependant variables, so the Variance is:

$$\begin{aligned}
Var(||X - Y||_2^2) &= Var(R) \\
&= Var(Z_1 + Z_2 + ... + Z_d) \\
&= Var(Z_1) + Var(Z_2) + ... + Var(Z_d) \\
&= d \times Var(Z) = d \times \frac{7}{180} = \frac{7d}{180}
\end{aligned} \tag{8}$$

1.3 (c)

Supposing that the edge length of the d-dimensional unit cube is l, and two d-dimensional points X and Y from a d-dimensional unit cube with a uniform distribution, i.e., $X, Y \in [0, 1]^d$. The maximum possible squared Euclidean distance between two points within the d-dimensional unit cube should be the distance between opposite corners of the max cube within $[0, 1]^d$ and it should be $\sum_1^d l^2$.

Since the edge length of the max cube should be $1(l = 1)$, we could get that $\sum_1^d l^2 = d$.

The mean $\mu = E(||X - Y||_2^2) = \frac{d}{6}$ grows accordingly if d grows, therefore, in high dimensions, like when d = 1200, the mean distance would be 200, which is pretty far, and can hardly be regarded as the near neighborhood sharing similarities.

At the same time, the standard deviation $\sigma = \sqrt{Var(||X - Y||_2^2)} = \sqrt{\frac{7d}{180}}$ would be $\sqrt{\frac{7}{180d}} =$

$\frac{1}{60}\sqrt{\frac{7}{60}} \approx 0.006$ fragment of the maximum distance, which is pretty small. Under this circumstances, we could regard that the distances between two points within the unit cube as approximately the same distance, which makes the "near neighborhood" meaningless.
Besides, if we consider from the relationship between the maximum distance and the density of the points in the cube, we would find that the distribution of the density grows to be like a spike with the growing of the dimensions. Most of the points would centered in the central of the graph, which shows that they are growing to be the same distance with each other and growing far away from each other. This supports the statement "most points are far away, and approximately the same distance" pretty much.

2 Information Theory

2.1 (a)

We could do some transformation that: $H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)} = -\sum_x p(x) \log_2 p(x)$. Since the value of $p(x)$ is between 0 and 1, $\log_2 p(x)$ must be negative. And thus, the $\sum_x p(x) \log_2 p(x)$ part must be negative. Therefore, $H(X) = -\sum_x p(x) \log_2 p(x)$ must be non-negative.

2.2 (b)

From $H(X) = -\sum_x p(x) \log_2 p(x)$, we know that:

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in\chi}\sum_{y\in Y} p(x,y) \log_2 p(x,y) \\
&= -\sum_{x\in\chi}\sum_{y\in Y} p(x,y) \log_2 p(x)p(y|x) \\
&= -\sum_{x\in\chi}\sum_{y\in Y} p(x,y) \log_2 p(x) - \sum_{x\in\chi}\sum_{y\in Y} p(x,y) \log_2 p(y|x) \\
&= -\sum_{x\in\chi}\sum_{y\in Y} p(x,y) \log_2 p(x) - \sum_{x\in\chi}\sum_{y\in Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)}) \\
&= -\sum_{x\in\chi} p(x) \log_2 p(x) - \sum_{y\in Y} p(y) \log_2 p(y) \\
&= H(X) + H(Y)
\end{aligned} \tag{9}$$

2.3 (c)

From $H(X) = -\sum_x p(x) \log_2 p(x)$, we know that:

$$\begin{aligned}
H(X,Y) &= -\sum_{x\in\chi}\sum_{y\in Y} p(x,y) \log_2 p(x,y) \\
&= -\sum_{x\in\chi}\sum_{y\in Y} p(x,y) \log_2 p(x)p(y|x) \\
&= -\sum_{x\in\chi}\sum_{y\in Y} p(x,y) \log_2 p(x) - \sum_{x\in\chi}\sum_{y\in Y} p(x,y) \log_2 p(y|x)
\end{aligned} \tag{10}$$

Since $p(y|x) = \frac{p(x,y)}{p(x)}$, we could know that $p(x,y) = p(y|x)p(x)$, and based on equation (11), we have:

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\chi}\sum_{y\in Y} p(x,y)\log_2 p(x) - \sum_{x\in\chi}\sum_{y\in Y} p(y|x)p(x)\log_2 p(y|x) \\ &= -\sum_{x\in\chi} p(x)\log_2 p(x) - \sum_{x\in\chi} p(y|x)\log_2 p(y|x) \\ &= H(x) + H(Y|X) \end{aligned} \tag{11}$$

2.4 (d)

$$\begin{aligned} KL(p||q) &= \sum_x p(x)\log_2 \frac{p(x)}{q(x)} \\ &= -(-\sum_x p(x)\log_2 \frac{p(x)}{q(x)}) \\ &= -(\sum_x p(x)\log_2 \frac{q(x)}{p(x)}) \\ &= -E(\log_2 \frac{q(x)}{p(x)}) \end{aligned} \tag{12}$$

From Jensen's inequality, we know that $E_{p(x)}(\log_2 \frac{q(x)}{p(x)}) \leq \log_2 E_{p(x)}(\frac{q(x)}{p(x)})$, since $\log_2 X$ is a convex function of x. So, can we get:

$$\begin{aligned} KL(p||q) &\geq -\log_2 E(\frac{q(x)}{p(x)}) \\ &= -\log_2 \int p(x)\frac{q(x)}{p(x)}dx \\ &= -\log_2 \int q(x)dx \\ &= -\log_2 1 \\ &= 0 \end{aligned} \tag{13}$$

Therefore, $KL(p||q)$ is non-negative.

2.5 (e)

$$\begin{aligned} I(Y,X) &= H(Y) - H(Y|X) \\ &= -\sum_y p(y)\log_2 p(y) - (-\sum_x\sum_y p(x,y)\log_2 p(y|x)) \\ &= -\sum_y p(y)\log_2 p(y) + \sum_x\sum_y p(x,y)\log_2 p(y|x) \\ &= -\sum_x\sum_y p(x,y)\log_2 p(y) + \sum_x\sum_y p(x,y)\log_2 \frac{p(x,y)}{p(x)} \end{aligned} \tag{14}$$

Since $\log_2 \frac{X}{Y} = \log_2 X - \log_2 Y$, we know that:

$$I(Y,X) = \sum_x\sum_y p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} = KL(p(x,y)||p(x)p(y)) \tag{15}$$

3 Decision Trees and K-Nearest Neighbour

3.1 For this part of homework, my solutions are based on the Jupyter Notebook file. Besides, I also collaborated with Zhimao Ling and wrote the python files together.

3.2 (b)

The accuracy of the decision tree (*max_depth* = 3, *criterion* = gini) is: 0.695918
The accuracy of the decision tree (*max_depth* = 3, *criterion* = entropy) is: 0.632653
The accuracy of the decision tree (*max_depth* = 4, *criterion* = gini) is: 0.693878
The accuracy of the decision tree (*max_depth* = 4, *criterion* = entropy) is: 0.679592
The accuracy of the decision tree (*max_depth* = 5, *criterion* = gini) is: 0.695918
The accuracy of the decision tree (*max_depth* = 5, *criterion* = entropy) is: 0.695918
The accuracy of the decision tree (*max_depth* = 6, *criterion* = gini) is: 0.712245
The accuracy of the decision tree (*max_depth* = 6, *criterion* = entropy) is: 0.695918
The accuracy of the decision tree (*max_depth* = 7, *criterion* = gini) is: 0.710204
The accuracy of the decision tree (*max_depth* = 7, *criterion* = entropy) is: 0.704082

3.3 (c)

The hyper-parameters which achieved the highest validation accuracy is *max_depth* = 6 and *criterion* = gini. With this model, we could get its accuracy on the test dataset as 0.69795. The visualization of the first two layer of the tree is shown as Figure 1.

3.4 (d)

From previous question, we know that the topmost split is whether "the" appears in the title. The information gain of this split, as we get from the function, is 0.052595.
Besides, the information gain of split on whether 'donald' appears in the title is: 0.056744; the information gain of split on whether 'trumps' appears in the title is: 0.041687; the information gain of split on whether 'era' appears in the title is: 0.000512; the information gain of split on whether 'hillary' appears in the title is: 0.032244.

3.5 (e)

The generated graph is shown as Figure 2. From Figure 2, we know that the model with best validation accuracy is the one taking k = 16. Its accuracy on the test data is 0.6694.

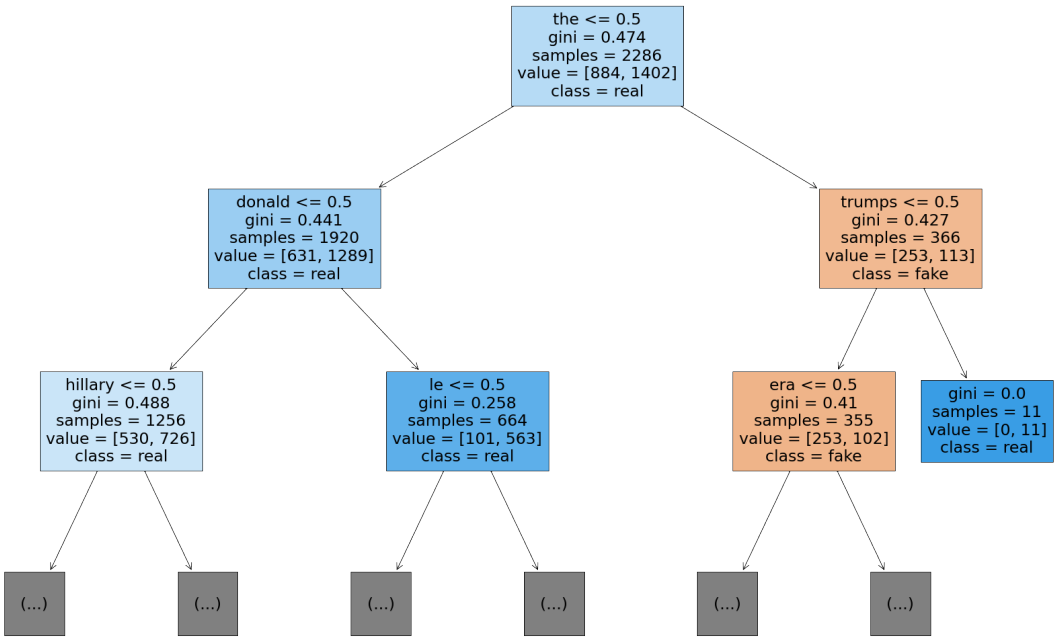

Figure 1: The first two layer of the tree

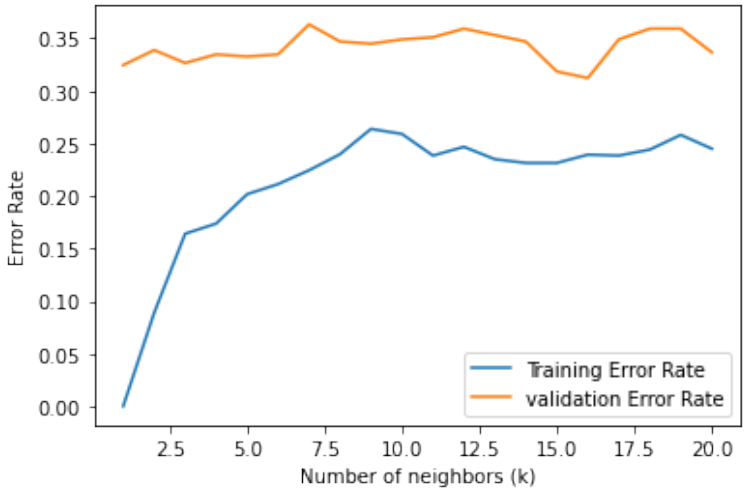

Figure 2: The first two layer of the tree