CSC311 HOMEWORK 1

Q1.

(a)

$$\mathrm{E(Z)} = \mathrm{E}(|\mathrm{X}-\mathrm{Y}|^2) = E(X^2+Y^2-2XY)$$
$$= \mathrm{E}(\mathrm{X}^2) + E(Y^2) - 2E(XY)\ by\ linearity\ of\ expectation$$
$$= \mathrm{Var(X)} + E(X)^2 + Var(Y) + E(Y)^2 - 2E(XY)\ since\ Var(X) = E(X^2) - E(X)^2$$
$$= \mathrm{Var(X)} + E(X)^2 + Var(Y) + E(Y)^2 - 2E(X)E(Y)\ since\ X, Y\ are\ independent$$

Since X and Y are two independent univariate random variables sampled uniformly from the unit interval [0, 1], the expectation of X and Y is $\mathrm{E(X)} = E(Y) = \frac{1}{2}, \mathrm{Var(X)} = Var(Y) = \frac{1}{12}$

$$\mathrm{E}(\mathrm{X}^2) = \mathrm{E}(\mathrm{Y}^2) = \int_0^1 x^2 dx = \frac{1}{3}$$
$$\mathrm{E}(\mathrm{X}^3) = \mathrm{E}(\mathrm{Y}^3) = \int_0^1 x^3 dx = \frac{1}{4}$$
$$\mathrm{E}(\mathrm{X}^4) = \mathrm{E}(\mathrm{Y}^4) = \int_0^1 x^4 dx = \frac{1}{5}$$

$$\mathrm{E(Z)} = \mathrm{Var(X)} + E(X)^2 + Var(Y) + E(Y)^2 - 2E(X)E(Y)$$
$$= \frac{1}{12} + \frac{1}{4} + \frac{1}{12} + \frac{1}{4} - 2\frac{1}{4}$$
$$= \frac{1}{6}$$

$$\mathrm{Var(Z)} = E(Z^2) - E(Z)^2$$
$$= \mathrm{E}((\mathrm{X}-\mathrm{Y})^4) - \frac{1}{6^2}$$
$$= \mathrm{E}(\mathrm{X}^4 - 4\mathrm{X}^3\mathrm{Y} + 6\mathrm{X}^2\mathrm{Y}^2 - 4\mathrm{X}\mathrm{Y}^3 + \mathrm{Y}^4) - \frac{1}{36}$$

by linearity of expectation and the fact that X and Y are independent

$$= \mathrm{E}(\mathrm{X}^4) - 4\mathrm{E}(\mathrm{X}^3)\mathrm{E}(\mathrm{Y}) + 6\mathrm{E}(\mathrm{X}^2)\mathrm{E}(\mathrm{Y}^2) - 4\mathrm{E}(\mathrm{X})\mathrm{E}(\mathrm{Y}^3) + \mathrm{E}(\mathrm{Y}^4) - \frac{1}{36}$$
$$= \frac{1}{5} - 4 * \frac{1}{4} * \frac{1}{2} + 6 * \frac{1}{3} * \frac{1}{3} - 4 * \frac{1}{2} * \frac{1}{4} + \frac{1}{5} - \frac{1}{36}$$
$$= \frac{7}{180}$$

Therefore, the expectation of Z is $\frac{1}{6}$ and the variance of Z is $\frac{7}{180}$.

(b)

$$\mathrm{E(R)} = \mathrm{E}(\mathrm{Z}_1 + Z_2 + \cdots + Z_d)$$
$$= \mathrm{E}(\mathrm{Z}_1) + E(Z_2) + \cdots + E(Z_d)$$
$$= \mathrm{E}(|\mathrm{X}_1 - \mathrm{Y}_1|^2) + \mathrm{E}(|\mathrm{X}_2 - \mathrm{Y}_2|^2) + \cdots + \mathrm{E}(|\mathrm{X}_\mathrm{d} - \mathrm{Y}_\mathrm{d}|^2)$$

$$\begin{aligned}\mathrm{Var(R)} &= Var(\mathrm{Z_1} + Z_2 + \cdots + Z_d)\\ &= \mathrm{Var(Z_1)} + \cdots + Var(Z_d)\\ &= \mathrm{Var}(|\mathrm{X_1} - \mathrm{Y_1}|^2) + \mathrm{Var}(|\mathrm{X_2} - \mathrm{Y_2}|^2) + \cdots + \mathrm{Var}(|\mathrm{X_d} - \mathrm{Y_d}|^2)\end{aligned}$$

Since all random variables $\mathrm{X_1}, \dots, X_d$ and $\mathrm{Y_1}, \dots, Y_d$ are independently and uniformly form [0, 1], using the answer from part (a), we get

$$\begin{aligned}\mathrm{E(R)} &= \mathrm{E}(|\mathrm{X_1} - \mathrm{Y_1}|^2) + \mathrm{E}(|\mathrm{X_2} - \mathrm{Y_2}|^2) + \cdots + \mathrm{E}(|\mathrm{X_d} - \mathrm{Y_d}|^2)\\ &= \mathrm{dE(Z)}\\ &= \frac{\mathrm{d}}{6}\end{aligned}$$

$$\begin{aligned}\mathrm{Var(R)} &= \mathrm{Var}(|\mathrm{X_1} - \mathrm{Y_1}|^2) + \mathrm{Var}(|\mathrm{X_2} - \mathrm{Y_2}|^2) + \cdots + \mathrm{Var}(|\mathrm{X_d} - \mathrm{Y_d}|^2)\\ &= \mathrm{dVar(Z)}\\ &= \frac{7\mathrm{d}}{180}\end{aligned}$$

(c)

Let MED be the maximum possible squared Euclidean distance between two points within the d-dimensional unit cube (i.e. the squared Euclidean distance between opposite corners of the cube (0, 0, …, 0) and (1, 1, …, 1)).

$$\mathrm{MED} = (1-0)^2 + (1-0)^2 + (1-0)^2 + \cdots + (1-0)^2 = \mathrm{d}$$

Compare $\mathrm{E(R)}$, $\mathrm{SD(R)}$ to MED, as dimension goes to infinity:

$$\lim_{d\to\infty} E(R) = \lim_{d\to\infty} \frac{\mathrm{d}}{6} \to \infty$$

And the standard deviation relative to MED when dimension goes to infinity is

$$\lim_{d\to\infty} \frac{\sqrt{Var(R)}}{d} = \lim_{d\to\infty} \sqrt{\frac{7}{180}}\frac{\sqrt{d}}{d} \to 0$$

Therefore, in high dimensions, most points are far away, and approximately the same distance.

Q2.

(a)

Since p(x) is a probability mass function, $0 \le \mathrm{p(x)} \le 1$.

Then $\frac{1}{p(x)} \ge 1$ and by the property of log function, $\log_2\left(\frac{1}{p(x)}\right) \ge 0$.

So $\mathrm{p(x)}\log_2\left(\frac{1}{p(x)}\right) \ge 0$ for all possible x.

Since $\mathrm{H}(X) = \sum_x \mathrm{p(x)}\log_2\left(\frac{1}{p(x)}\right)$ where x is all possible values, $H(X) \ge 0$.

Hence H(X) is non-negative.

(b)

Since X, Y are independent random variables, p(x, y) = p(x)p(y)

Then

$$H(X,Y) = \sum_x \sum_y p(x,y)\log_2\left(\frac{1}{p(x,y)}\right)$$

$= \sum_x \sum_y p(x)p(y) \log_2 \left(\frac{1}{p(x)p(y)}\right)$

$= \sum_x \sum_y p(x)p(y) \left(\log_2 \left(\frac{1}{p(x)}\right) + \log_2 \left(\frac{1}{p(y)}\right)\right)$

$= \sum_x \sum_y p(x)p(y) \log_2 \left(\frac{1}{p(x)}\right) + \sum_x \sum_y p(x)p(y) \log_2 \left(\frac{1}{p(y)}\right)$

$= \sum_y p(y) \sum_x p(x) \log_2 \left(\frac{1}{p(x)}\right) + \sum_x p(x) \sum_y p(y) \log_2 \left(\frac{1}{p(y)}\right)$

Since $\sum_x p(x) = p(\chi) = 1 = \sum_y p(y)$

$= \sum_x p(x) \log_2 \left(\frac{1}{p(x)}\right) + \sum_y p(y) \log_2 \left(\frac{1}{p(y)}\right)$

$= H(X) + H(Y)$

(c)

By the definition of conditional entropy,

$$H(Y|X) = -\sum_x \sum_y p(x,y) \log_2 p(y|x) = -\sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)}$$

$$= \sum_x \sum_y p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)}\right)^{-1} = \sum_x \sum_y p(x,y) \log_2 \left(\frac{p(x)}{p(x,y)}\right)$$

$$= \sum_x \sum_y p(x,y) \log_2(p(x)) - \sum_x \sum_y p(x,y) \log_2(p(x,y))$$

$$= \sum_x p(x) \log_2(p(x)) + \sum_x \sum_y p(x,y) \log_2 \left(\frac{1}{p(x,y)}\right) \textit{ since } \sum_y p(x,y) = p(x)$$

$$= \sum_x \sum_y p(x,y) \log_2 \left(\frac{1}{p(x,y)}\right) - \sum_x p(x) \log_2 \left(\frac{1}{p(x)}\right)$$

$$= H(X,Y) - H(X)$$

Hence, $H(X,Y) = H(X) + H(Y|X)$

(d)

From the question, we know $\mathrm{KL}(\mathrm{p}||\mathrm{q}) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)} = -\sum_x p(x) \log_2 \frac{q(x)}{p(x)}$

To prove $\mathrm{KL}(\mathrm{p}||\mathrm{q}) \geq 0$,

we want to prove $\sum_x p(x) \log_2 \frac{q(x)}{p(x)} \leq 0$ so that $-\sum_x p(x) \log_2 \frac{q(x)}{p(x)} \geq 0$.

Since $-\log_2$ is a concave function, using Jensen's inequality $\emptyset(\mathrm{E}[\mathrm{X}]) \leq E[\emptyset(\mathrm{X})]$, we get

$$-\sum_x p(x) \log_2 \frac{q(x)}{p(x)} = \sum_x p(x) \left(-\log_2 \frac{q(x)}{p(x)}\right)$$

$$= E_{X\sim p}\left[-\log_2 \frac{q(x)}{p(x)}\right] \geq -\log_2 E_{X\sim p}\left[\frac{q(x)}{p(x)}\right]$$

$$= -\log_2 \sum_x p(x)\frac{q(x)}{p(x)} = -\log_2 \sum_x q(x)$$

Since q is a probability distribution, $\sum_x q(x) = 1 \rightarrow -\log_2 \sum_x q(x) = 0$

So $-\sum_x p(x) \log_2 \frac{q(x)}{p(x)} \geq -\log_2 \sum_x q(x) = 0$

Therefore $\mathrm{KL}(\mathrm{p}||\mathrm{q}) = -\sum_x p(x) \log_2 \frac{q(x)}{p(x)} \geq 0$ i.e. $\mathrm{KL}(\mathrm{p}||\mathrm{q})$ is non-negative.

(e)

By the definition of KL divergence,

$$\mathrm{KL}(\mathrm{p(x,y)}||\mathrm{p(x)p(y)}) = \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{\mathrm{p(x)p(y)}}$$

By the definition of information gain and entropy

$$\mathrm{I(Y;X)} = H(Y) - H(Y|X) = \sum_y \mathrm{p(y)} \log_2 \left(\frac{1}{p(y)}\right) + \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)}$$

$$= \sum_x \sum_y \mathrm{p(x,y)} \log_2 \left(\frac{1}{p(y)}\right) + \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)} \; since \; \sum_x p(x,y) = p(y)$$

$$= \sum_x \sum_y \mathrm{p(x,y)} \log_2 \left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Hence,

$$\mathrm{I(Y;X)} = \mathrm{KL}(\mathrm{p(x,y)}||\mathrm{p(x)p(y)})$$

3.

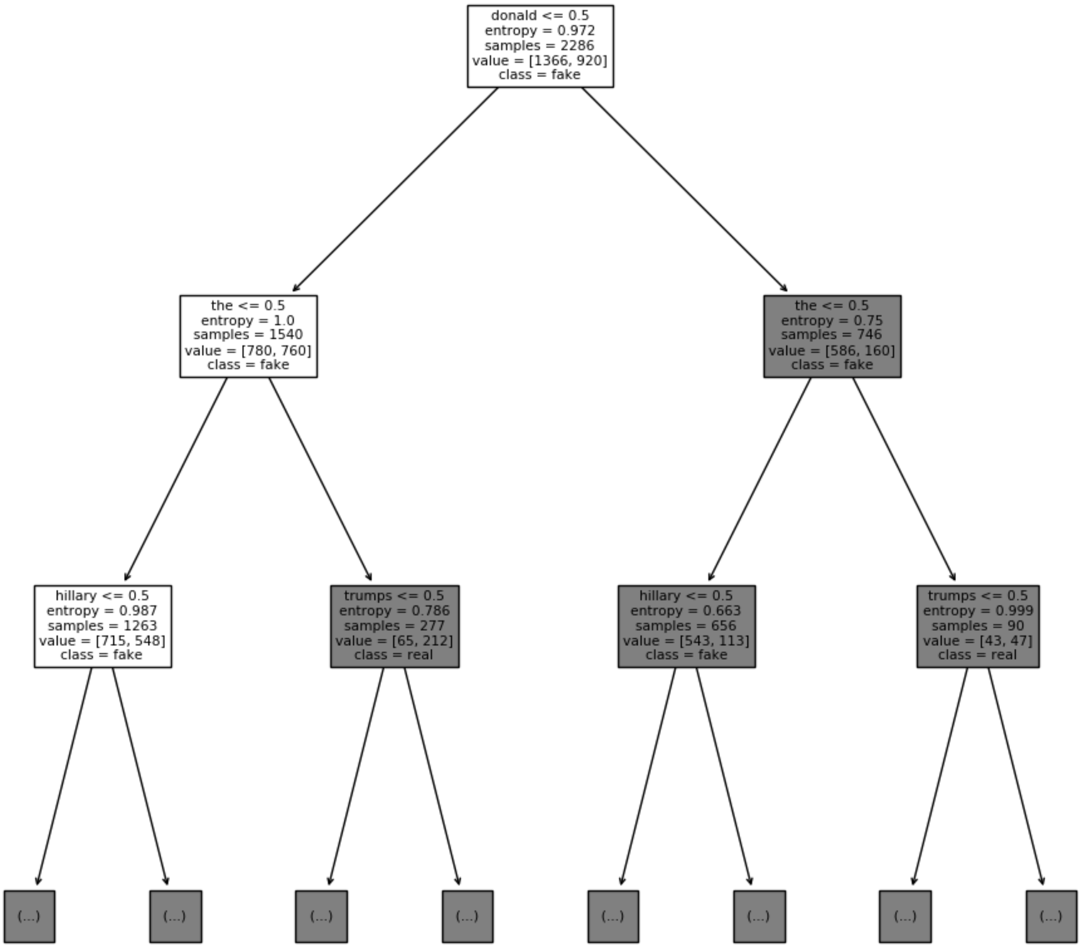
(b)

Accuracy for depth 16 and criteria entropy : 0.7346938775510204

Accuracy for depth 32 and criteria entropy : 0.753061224489796

Accuracy for depth 64 and criteria entropy : 0.7510204081632653
Accuracy for depth 128 and criteria entropy : 0.7346938775510204
Accuracy for depth 256 and criteria entropy : 0.746938775510204
Accuracy for depth 16 and criteria gini : 0.7306122448979592
Accuracy for depth 32 and criteria gini : 0.736734693877551
Accuracy for depth 64 and criteria gini : 0.753061224489796
Accuracy for depth 128 and criteria gini : 0.7489795918367347
Accuracy for depth 256 and criteria gini : 0.7448979591836735

(c)
Accuracy of the best hyperparameter on the test dataset: 0.746938775510204
Plot of the tree:

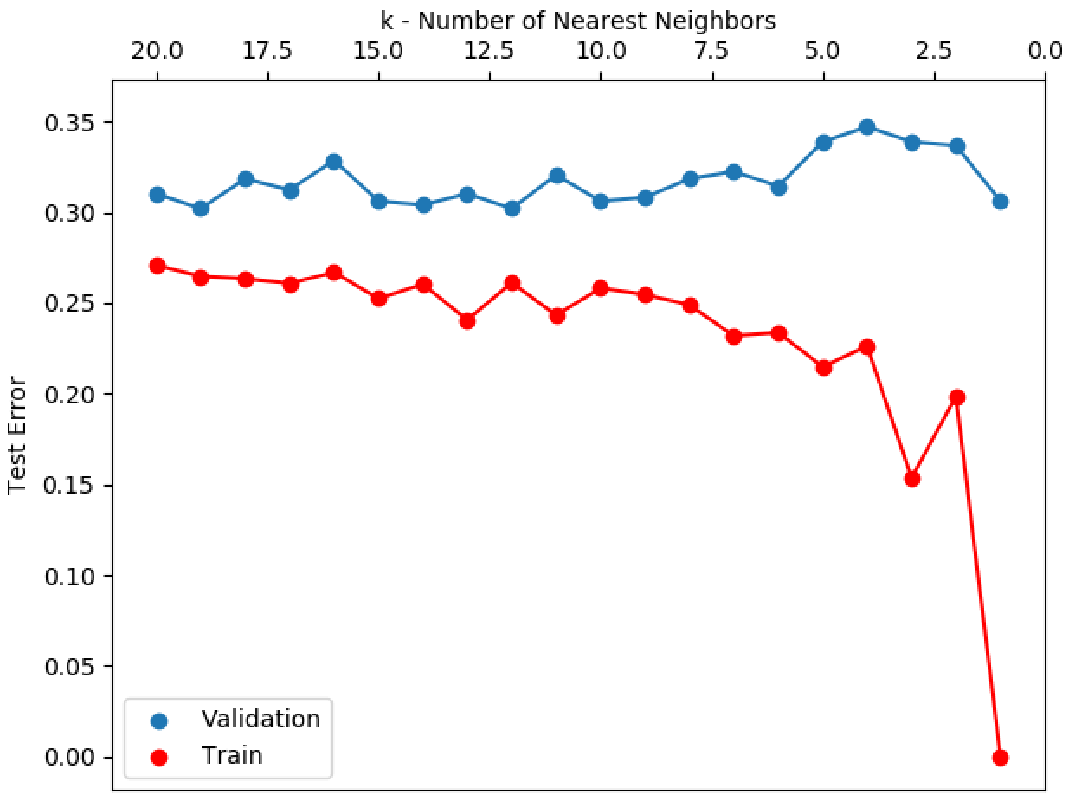


(d)
The information gain of 'donald': 0.05404605736121895
The information gain of 'trumps': 0.04240472003947697
The information gain of 'china': 0.0031134148983391124

The information gain of 'good': 0.00049429451797043554

(e)

The relationship between number of nearest neighbours and test error



Accuracy of the best KNN model on the test dataset: 0.6551020408163265