CSC311 A1

Wilson Wei-Sheng Hsu

February 2020

Question 1

Part A

Find $E(Z)$ and $V(Z)$:

$$\begin{aligned}
E(Z) &= E(|X-Y|^2) = E((X-Y)^2) \\
&= E(X^2 - 2XY + Y^2) \\
&= E(X^2) - 2E(XY) + E(Y^2) && \text{Property: } E(aX+bY) = aE(X) + bE(Y) \\
&= E(X^2) - 2E(X)E(Y) + E(Y^2) && \text{Property: If X and Y are independent, then } E(XY) = E(X) \times E(Y)
\end{aligned}$$

Since X, Y are uniformly distributed, we know that expectation of an uniform distribution $Uniform(a,b)$ is $E(X) = \frac{a+b}{2}$. We find that

$$E(X) = E(Y) = \frac{0+1}{2}$$

and

$$E(X^2) = E(Y^2) = \int_0^1 x^2 dx = \frac{x^3}{3}|_0^1 = \frac{1}{3}$$

Notice, $f(x)$ for $Uniform(a,b) = \frac{1}{b-a}$ and in this case it is equal to 1.
So

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

$$\begin{aligned}
V(Z) &= E(Z^2) - E(Z)^2 \\
&= E((X-Y)^4) - \frac{1}{36} && E(Z) = \tfrac{1}{6} \text{ from last part} \\
&= E(X^4 - 4X^3Y + 6X^2Y^2 - 4XY^3 + Y^4) - \frac{1}{36} \\
&= E(X^4) - 4E(X^3)E(Y) + 6E(X^2)E(Y^2) - 4E(X)E(Y^3) + E(Y^4) - \frac{1}{36}
\end{aligned}$$

Know $E(X^3) = E(Y^3) = \int_0^1 x^3 dx = \frac{x^4}{4}|_0^1 = \frac{1}{4}$, $E(X^4) = E(Y^4) = \int_0^1 x^4 dx = \frac{x^5}{5}|_0^1 = \frac{1}{5}$.
We get that

$$V(Z) = \frac{1}{5} - \frac{1}{2} + \frac{2}{3} - \frac{1}{2} + \frac{1}{5} - \frac{1}{36} = \frac{7}{180}$$

Part B

Determine $E[||X-Y||_2^2] = E(R)$ and $V[||X-Y||_2^2] = V(R)$

$$\begin{aligned} E[||X-Y||_2^2] &= E(R) \\ &= E(Z_1 + ... + Z_d) \\ &= E(Z_1) + ... + E(Z_d) && \text{Property: } E(aX+bY) = aE(X) + bE(Y) \\ &= \frac{d}{6} = dE(Z) && \text{From last part} \end{aligned}$$

$$\begin{aligned} V[||X-Y||_2^2] &= V(R) \\ &= V(Z_1 + ... + Z_d) \\ &= V(Z_1) + ... + V(Z_d) && \text{Property: If X and Y are independent } V(X+Y) = V(X) + V(Y) \\ &= \frac{7d}{180} = dV(Z) && \text{From last part} \end{aligned}$$

Part C

Squared Euclidean Distance of two points in dimension d is given by

$$Squared\_Euclidean\_distance(p,q) = (p_1 - q_1)^2 + ... + (p_d - q_d)^2$$

and maximum possible value for $p_i - q_i$ is 1, therefore maximum squared Euclidean distance would be d.
We know from Part B that $E(R) = \frac{d}{6}, V(R) = \frac{7d}{180}$ and so $\sigma(R) = \sqrt{\frac{7d}{180}}$ which is relatively small compare to mean and maximum distance. Since the standard deviation is small, we have a sense that most of the points are around the mean, therefore they are around the same distance away from each other.
If we compare our $E(R) = \frac{d}{6}$ with maximum squared Euclidean distance , we see that our average squared Euclidean distance between points is only $\frac{E(R)}{d} = \frac{1}{6}$ of the maximum squared Euclidean distance. And as in high dimension, when d becomes very large, we know that our $E(R)$ becomes very large as well, therefore most points in high dimension are far away from each other.

Question 2

Part A

Prove that $H(X) = \sum_x p(x) log_2(\frac{1}{p(x)})$ is non-negative:
By criteria of a PMF, $0 \leq p(x) \leq 1$, which is always non-negative.
Using the inequality we get:

$$1 \leq \frac{1}{p(x)} < \infty$$

in the $p(x) = 0$ case, if p(x) = 0, our H(X) = 0.

$$0 \leq log_2(\frac{1}{p(x)}) < \infty$$

and therefore since every term in $H(X) = \sum_x p(x) log_2(\frac{1}{p(x)})$ is non-negative, we can conclude that $H(X)$ is non-negative as well.

Part B

Show that If X, Y are independent, then $H(X,Y) = H(X) + H(Y)$
By definition:

$$\begin{aligned}
H(X,Y) &= \sum_x \sum_y p(x,y) log_2(\frac{1}{p(x,y)}) \\
&= \sum_x \sum_y p(x)p(y) log_2(\frac{1}{p(x)p(y)}) && \text{Property: If } X, Y \text{ are independent, then } p(x,y) = p(x)p(y) \\
&= \sum_x \sum_y p(x)p(y)(log_2(\frac{1}{p(x)}) + log_2(\frac{1}{p(y)})) && \text{Logarithmic property} \\
&= \sum_x \sum_y p(x)p(y) log_2(\frac{1}{p(x)}) + \sum_x \sum_y p(x)p(y) log_2(\frac{1}{p(y)}) \\
&= \sum_y p(y) \sum_x p(x) log_2(\frac{1}{p(x)}) + \sum_x p(x) \sum_y p(y) log_2(\frac{1}{p(y)}) \\
&= \sum_x p(x) log_2(\frac{1}{p(x)}) + \sum_y p(y) log_2(\frac{1}{p(y)}) && \text{PMF Criteria: } \textstyle\sum_x p(x) = \sum_y p(y) = 1 \\
&= H(X) + H(Y)
\end{aligned}$$

Concludes.

Part C

Show that $H(X,Y) = H(X) + H(Y|X)$
By definition:

$$\begin{aligned}
H(X,Y) &= \sum_x \sum_y p(x,y) log_2(\frac{1}{p(x,y)}) \\
&= \sum_x \sum_y p(x)p(y|x) log_2(\frac{1}{p(x)p(y|x)}) && \text{Definition: } p(x,y) = p(x)p(y|x) \\
&= \sum_x \sum_y p(x)p(y|x) log_2(\frac{1}{p(x)}) + \sum_x \sum_y p(x,y) log_2(\frac{1}{p(y|x)}) \\
& && \text{Logarithmic property} \\
&= \sum_x p(x)(\frac{1}{p(x)}) \sum_y p(y|x) + \sum_x \sum_y p(x,y) log_2(\frac{1}{p(y|x)}) \\
&= \sum_x p(x)(\frac{1}{p(x)}) + \sum_x \sum_y p(x,y) log_2(\frac{1}{p(y|x)}) && \text{PMF Criteria: } \textstyle\sum_y p(y|x) = 1 \\
&= H(X) + H(Y|X) && \text{Definition of } H(X), H(Y|X)
\end{aligned}$$

Concludes.

Part D

Prove that $KL(p||q)$ is non-negative.
Assume $p > 0, q > 0$

By definition:

$$
\begin{aligned}
KL(p||q) &= \sum_x p(x) log_2(\frac{p(x)}{q(x)}) \\
&= \sum_x p(x)(-log_2(\frac{q(x)}{p(x)}) && \text{Logarithmic property} \\
&<= -log_2(\sum_x p(x)\frac{q(x)}{p(x)}) && \text{Log is a concave function, and -Log is convex, so by Jensen's inequality.} \\
&= -log_2(1) && \text{PMF criteria} \\
&= 0
\end{aligned}
$$

and 0 is non-negative.

Part E

Show that $I(Y;X) = KL(p(x,y)||p(x)p(y))$
By definition:

$$
\begin{aligned}
I(Y;X) &= H(Y) - H(Y|X) \\
&= H(X,Y) - H(X|Y) - H(Y|X) && \text{We know } H(Y) = H(X,Y) - H(X|Y) \\
&= \sum_x \sum_y p(x,y) log_2(\frac{1}{p(x,y)}) - \sum_x \sum_y p(x,y) log_2(\frac{1}{p(x|y)}) - \sum_x \sum_y p(x,y) log_2(\frac{1}{p(y|x)}) \\
&= \sum_x \sum_y p(x,y)(log_2(\frac{1}{p(x,y)}) - log_2(\frac{1}{p(x|y)}) - log_2(\frac{1}{p(y|x)})) && \text{Summation property} \\
&= \sum_x \sum_y p(x,y)(log_2(\frac{1}{p(x,y)} p(x|y)p(y|x))) && \text{Logrithmic property} \\
&= \sum_x \sum_y p(x,y)(log_2(\frac{1}{p(x,y)} \frac{p(x,y)}{p(y)} \frac{p(x,y)}{p(x)})) && \text{by definition} \\
&= \sum_x \sum_y p(x,y) log_2(\frac{p(x,y)}{p(x)p(y)}) \\
&= KL(p(x,y)||p(x)p(y)) && \text{by definition}
\end{aligned}
$$

Concludes.

Question 3

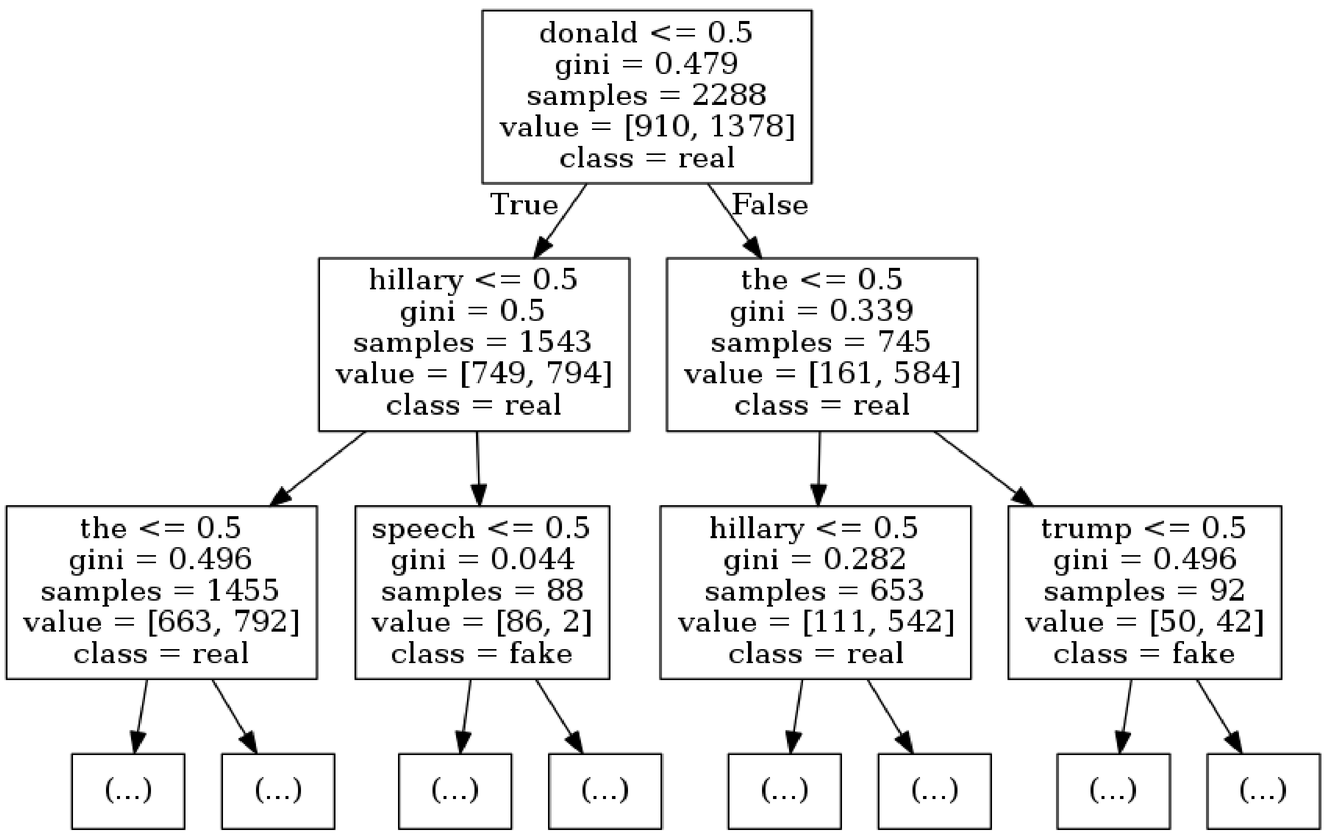
Part A

Included in python source file.

Part B

```
Tree with depth 15 and criterion gini : validation score = 0.7198364008179959
Tree with depth 30 and criterion gini : validation score = 0.7566462167689162
Tree with depth 45 and criterion gini : validation score = 0.754601226993865
Tree with depth 60 and criterion gini : validation score = 0.7648261758691206
Tree with depth 75 and criterion gini : validation score = 0.7730061349693251
Tree with depth 15 and criterion entropy : validation score = 0.7116564417177914
Tree with depth 30 and criterion entropy : validation score = 0.7443762781186094
Tree with depth 45 and criterion entropy : validation score = 0.7607361963190185
Tree with depth 60 and criterion entropy : validation score = 0.7607361963190185
Tree with depth 75 and criterion entropy : validation score = 0.7730061349693251
Best Tree Model is given by criterion: gini, max depth: 75, validation score: 0.7730061349693251
```

Part C



Part D

```
Computed Information gain from the word 'the': 0.05760902430425707
Computed Information gain from the word 'hillary': 0.03764013901583019
Computed Information gain from the word 'donald': 0.049373316707295944
Computed Information gain from the word 'trump': 0.0
Computed Information gain from the word 'war': 0.000891963349783475
```

Part E

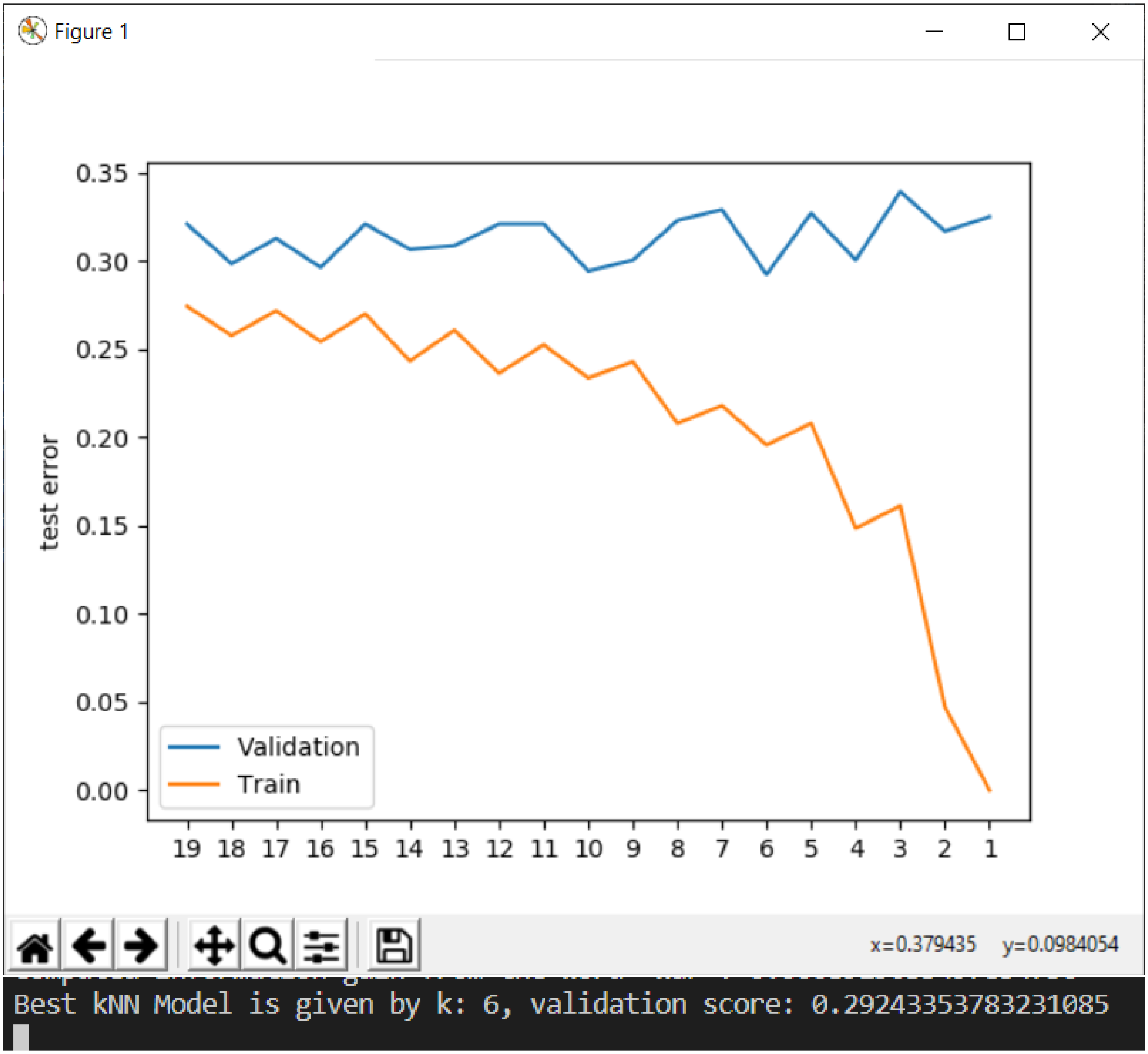


Figure 1
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
test error
Validation
Train
19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
x=0.379435 y=0.0984054
Best kNN Model is given by k: 6, validation score: 0.29243353783231085