Homework 1 by Pavel Golikov. I am working in a group with Yash Kant and Barza Nisar.

1 Question 1

a) First, consider the definition of absolute value function and the expression inside the function:

$$|X-Y| = \begin{cases} X-Y & \text{if } X \geq Y \\ Y-X & \text{if } Y < X \end{cases}$$

and notice that we are squaring this expression, so $(Y-X)^2 = (X-Y)^2$ which means that in our case (because we have X and Y defined over real (and not complex) numbers), we have $|X-Y|^2 = (X-Y)^2$, which means that we can remove absolute value from Z and consider that simplified $Z = (X-Y)^2$. First, we compute the expectation of Z. According to definition of expectation:

$$E[(X-Y)^2] = \int_0^1 \int_0^1 (x-y)^2 p(x)p(y)\, dx\, dy \;= \int_0^1 \int_0^1 (x-y)^2\, dx\, dy$$

since X and Y are sampled from uniform distribution over $[0,1]$. Computing the integral directly:

$$\int_0^1 \int_0^1 (x-y)^2\, dx\, dy$$

$$= \int_0^1 \int_0^1 x^2 - 2xy + y^2\, dx\, dy$$

$$= \int_0^1 (\int_0^1 x^2 - 2xy + y^2\, dx)\, dy$$

$$= \int_0^1 (\int_0^1 x^2\, dx - \int_0^1 2xy\, dx + \int_0^1 y^2\, dx)\, dy$$

$$= \int_0^1 (\frac{1}{3} - y + y^2)\, dy$$

$$= \int_0^1 \frac{1}{3}\, dy - \int_0^1 y\, dy + \int_0^1 y^2\, dy$$

$$= \frac{1}{3} - \frac{1}{2} + \frac{1}{3} = \frac{1}{6}$$

Second, we compute the variance of random variable Z:

$$Var(Z) = Var(|X-Y|^2) = E[(X-Y)^4] - E[(X-Y)^2]^2 = \frac{1}{15} - \frac{1}{36} = \frac{7}{180}$$

Second term in the difference is a square of expectation of Z, and the first term is computed similarly to expectation of Z, i.e. as a direct integral - we expand the product and integrate each term directly, noting that $p(X)$ and $p(Y)$ are equal to 1 since we are sampling from uniform distribution over $[0,1]$. I omit the computations here since they are very similar to those in previous part of the question.

b) First, let us expand R:

$$R = Z_1 + Z_2 + ... + Z_d = \sum_{n=1}^{d} |X_i - Y_i|^2$$

Now, we compute the expectation of R as follows:

$$E[R] = E[Z_1 + Z_2 + ... + Z_d] = \sum_{n=1}^{d} E[\|X_i - Y_i\|^2]$$

since we know that we can view each variable X_i and Y_i as drawn independently and uniformly from $[0, 1]$, i.e. all variables are mutually independent. We thus get:

$$E[R] = E[Z_1 + Z_2 + ... + Z_d] = \sum_{n=1}^{d} E[\|X_i - Y_i\|^2] = \frac{d}{6}$$

Now let us compute the variance of R. We know that Z_i is independent of Z_j for all $0 < i, j < d$, which means that we can use the fact that variance of sum of independent variables is the sum of variances of each variable. Thus we have:

$$Var(R) = Var(\sum_{n=1}^{d} |X_i - Y_i|^2) = \sum_{n=1}^{d} Var(|X_i - Y_i|^2) = \frac{7d}{180}$$

2 Question 2

a) First, let us expand the definition of $H(X)$:

$$H(X) = \sum_x p(x) log_2(\frac{1}{p(x)}) = -\sum_x p(x) log_2(p(x))$$

At this point it is left to notice that p(x) is always positive, $log_2(p(x))$ is always negative (because function $p(x)$'s range is between 0 and 1, and logarithm function is negative on that range), which makes the entire expression non negative.

b) We will expand the definition of $H(X, Y)$, noting that since X and Y are independent, we have $p(x, y) = p(x)p(y)$:

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

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

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

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

At this point, note that $\sum_x p(x) = 1$ since the summation is over all x and likewise for y. This means we can simplify the two terms in our expression by collapsing the summation of $p(y)$ in the first part and summation of $p(x)$ to get:

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

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

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

c)

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

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

$$-\sum_x \sum_y p(y|x)p(x) log_2(p(y|x)p(x))$$

$$-\sum_y p(y|x) log_2(p(y|x)) - \sum_x p(x) log_2(p(x))$$

$$H(Y|X) + H(X)$$

d)

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

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

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

Now, let Y be a random variable s.t. $Y = g(X)$, where $g(x) = \frac{q(x)}{p(x)}$. Since log is a concave function, we have, by Jensen's inequality:

$$log(E[g(x)]) \geq E[log(g(x))]$$

Now, expanding $g(x)$ and definition of expectation:

$$log(\sum_x p(x)g(x)) \geq \sum_x log(g(x))$$

$$log(\sum_x p(x)\frac{q(x)}{p(x)}) \geq \sum_x p(x) log(\frac{q(x)}{p(x)})$$

$$log(\sum_x q(x)) \geq \sum_x p(x) log(\frac{q(x)}{p(x)})$$

$$log(1) \geq \sum_x p(x) log(\frac{q(x)}{p(x)})$$

$$0 \geq \sum_x p(x) log(\frac{q(x)}{p(x)})$$

which proves that $KL(p||q)$ is non negative.

e)

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

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

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

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

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

Notice that the first term is $-H(Y|X)$ and in the second term, we can collapse on the variable x since the sum $\sum_x \sum_y p(x,y) log_2(p(y))$ is marginalization of x, so we get:

$$= -H(Y|X) - \sum_y p(y) log_2(p(y))$$

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

3 Question 3

a) Our code is contained in file hw1_code.py

b) According to our calculations, the 5 "sensible" sizes for tree depth were 64, 128, 256, 512, and 1024.

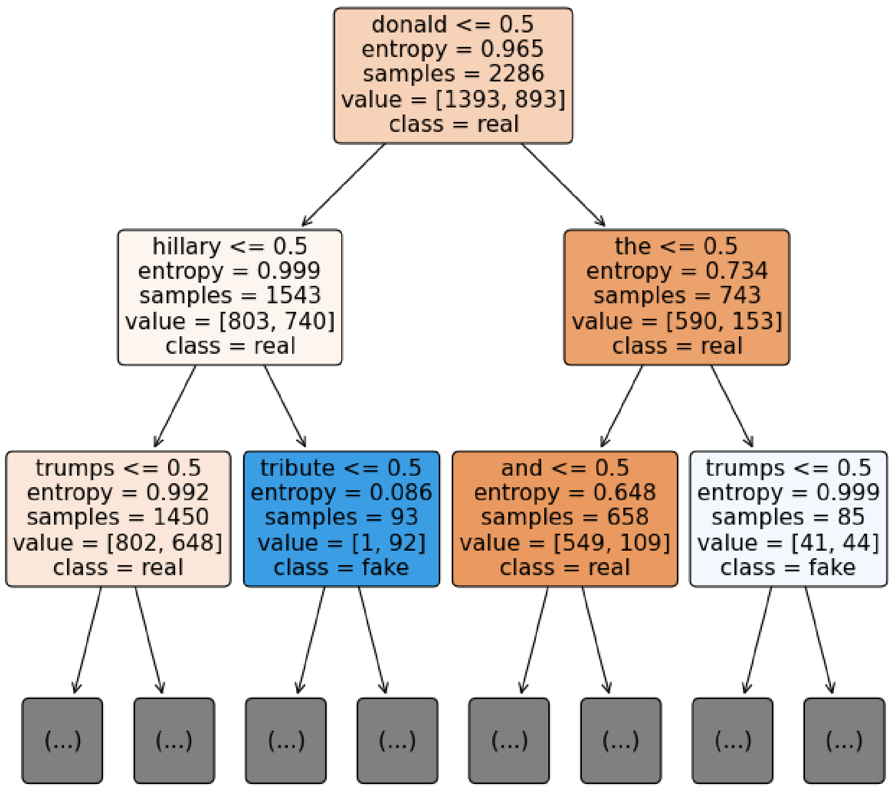
Validation accuracy with depth 64 and criterion entropy = 0.7244897959183674
Validation accuracy with depth 128 and criterion entropy = 0.7244897959183674
Validation accuracy with depth 256 and criterion entropy = 0.7428571428571429
Validation accuracy with depth 512 and criterion entropy = 0.7428571428571429
Validation accuracy with depth 1024 and criterion entropy = 0.7428571428571429

Validation accuracy with depth 64 and criterion gini = 0.7244897959183674
Validation accuracy with depth 128 and criterion gini = 0.736734693877551
Validation accuracy with depth 256 and criterion gini = 0.736734693877551
Validation accuracy with depth 512 and criterion gini = 0.736734693877551
Validation accuracy with depth 1024 and criterion gini = 0.736734693877551

Validation accuracy was calculated in our method select_tree_model.

c) Best DT model has depth 256 and criterion entropy with validation accuracy = 0.7428571428571429
Best DT model has test accuracy: 0.7612244897959184

We provide calculation of Information Gain for 4 keywords. Please see the illustration and Information Gain calculation results below:



Ours: Information Gain for splitting at feature: donald is: [0.05260332]
Sklearn's: Mutual Information for splitting at feature: donald is: 0.036562781102445̲46
Ours: Information Gain for splitting at feature: hillary is: [0.04127666]
Sklearn's: Mutual Information for splitting at feature: hillary is: 0.028610797475487986
Ours: Information Gain for splitting at feature: trump is: [0.02758086]
Sklearn's: Mutual Information for splitting at feature: trump is: 0.026673266691983714
Ours: Information Gain for splitting at feature: plotting is: [0.00059343]
Sklearn's: Mutual Information for splitting at feature: plotting is: 0.000411335795472314

e)

With K = 1, val error: 0.326530612244898 and train error: 0.0
With K = 2, val error: 0.3877551020408163 and train error: 0.17104111986001747
With K = 3, val error: 0.36530612244897964 and train error: 0.1697287839020123
With K = 4, val error: 0.3795918367346939 and train error: 0.22878390201224852
With K = 5, val error: 0.3408163265306122 and train error: 0.19335083114610674
With K = 6, val error: 0.36734693877551017 and train error: 0.24190726159230092
With K = 7, val error: 0.3224489795918367 and train error: 0.22003499562554685
With K = 8, val error: 0.35918367346938773 and train error: 0.2497812773403325
With K = 9, val error: 0.34693877551020413 and train error: 0.21959755030621175
With K = 10, val error: 0.36734693877551017 and train error: 0.26115485564304464
With K = 11, val error: 0.34285714285714286 and train error: 0.23797025371828517
With K = 12, val error: 0.35918367346938773 and train error: 0.2637795275590551
With K = 13, val error: 0.3326530612244898 and train error: 0.2432195975503062
With K = 14, val error: 0.34285714285714286 and train error: 0.26596675415573057
With K = 15, val error: 0.3346938775510204 and train error: 0.2502187226596675
With K = 16, val error: 0.3571428571428571 and train error: 0.2755905511811023
With K = 17, val error: 0.33061224489795915 and train error: 0.25896762904636916
With K = 18, val error: 0.34693877551020413 and train error: 0.2839020122484689
With K = 19, val error: 0.3346938775510204 and train error: 0.26596675415573057
With K = 20, val error: 0.35306122448979593 and train error: 0.2970253718285214

Best KNN model has K = 7 with val: 0.6775510204081633 and train: 0.7799650043744532
Best KNN model has test accuracy: 0.7081632653061225

Please see below the graph requested in the assignment: