## CSC311 HOMEWORK 1

Xinyi Liu 1003792773

Q1.

(a)

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

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

Therefore, the expectation of Z is and the variance of Z is .

(b)

Since all random variables and are independently and uniformly form [0, 1], using the answer from part (a), we get

(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)).

Compare , to MED, as dimension goes to infinity:

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

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

Q2.

(a)

Since p(x) is a probability mass function, .

Then and by the property of log function, .

So for all possible x.

Since where x is all possible values, .

Hence H(X) is non-negative.

(b)

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

Then

Since

(c)

By the definition of conditional entropy,

Hence,

(d)

From the question, we know

To prove ,

we want to prove .

Since is a concave function, using Jensen’s inequality , we get

Since q is a probability distribution,

So

Therefore i.e. is non-negative.

(e)

By the definition of KL divergence,

By the definition of information gain and entropy

Hence,

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

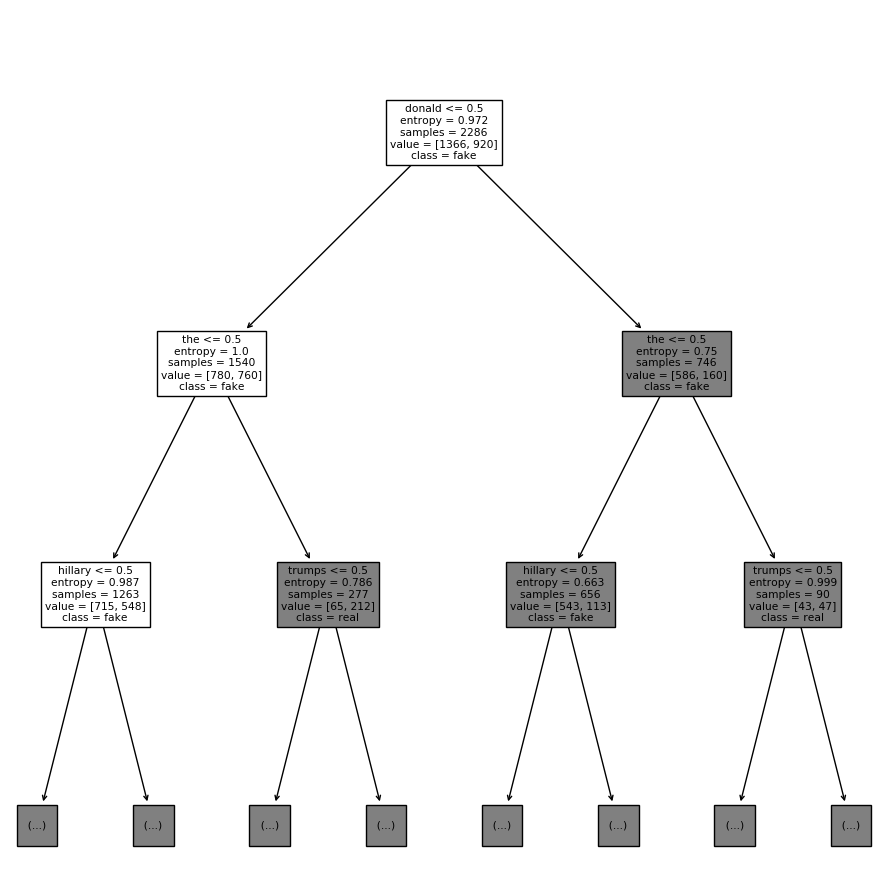
Accuracy for depth 128 and criteria gini : 0.7489795918367347

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

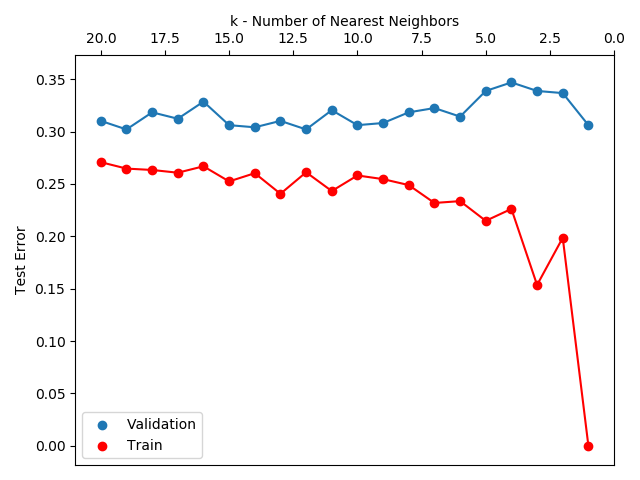
The information gain of 'trumps': 0.04240472003947697

The information gain of 'china': 0.0031134148983391124

The information gain of 'good': 0.0004942945179704354

(e)

The relationship between number of nearest neighbours and test error



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