**CSC420 Homework 1**

**Question 1: Nearest Neighbours and the Curse of Dimensionality**

**Question 1a: Expectation and Variance of V**

**Final Answer (work below)**:

We know:

1. are independent

Using additive property of expectation to obtain

Since are independent:

Since are sampled from the identical distributions, all their moments must be the same

Recall properties of uniform distributions:

Thus for

Creating an equation for ’s n-th moment about 0 ()

Using the above

**Question 1a, continued…**

Definition of variance in terms of expectations

Using the additive property of expectation to obtain

Since are independently sampled:

Since are sampled from identical distributions

Using the previous moment equation

Using all prior results to compute

**Question 1b: Expectation and Variance of R**

**Final Answer (work below):**

Define to be the squared Euclidean distance between two unit-uniformly distributed -dimensional points

Expectation of (simplify using the additive property of expectation)

Since all are drawn from identical distributions, and all have the same relationship with and , all must also be drawn from identical distributions. Therefore, all have identical expectations:

According to the above, we can turn the summation in into a multiplication of

Since is distributed the same as in part (a),

Since each is independent of its peers (i.e. ), all variances can simply be summed together

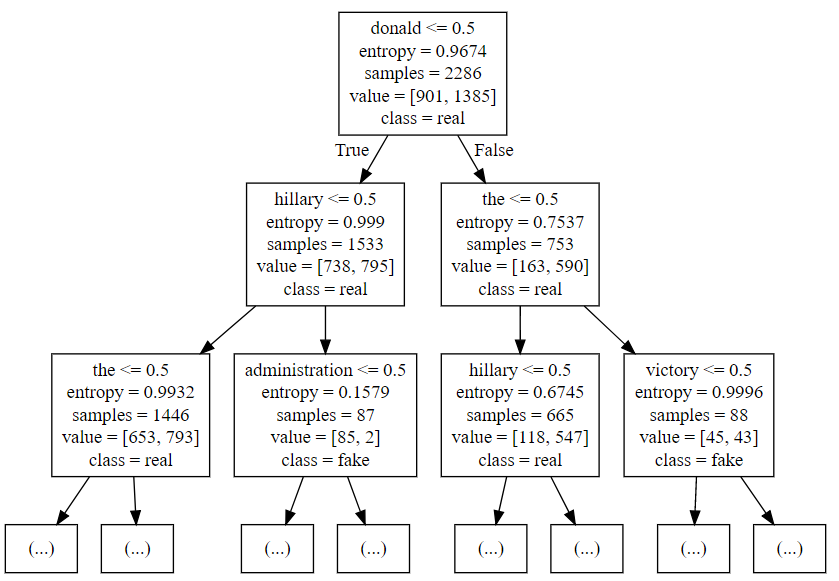
Since is distributed identically, we can turn the above summation into a multiplication of , and use the previous result

**Question 2: Decision Trees**

**Question 2b: Output of select\_model (tested over both criteria, and depths 2-6, inclusive)**

|  |
| --- |
| Model(depth=2, criteria=entropy): score = 0.6346938775510204  Model(depth=3, criteria=entropy): score = 0.6857142857142857  Model(depth=4, criteria=entropy): score = 0.7061224489795919  Model(depth=5, criteria=entropy): score = 0.689795918367347  Model(depth=6, criteria=entropy): score = 0.689795918367347  Model(depth=2, criteria=gini): score = 0.7040816326530612  Model(depth=3, criteria=gini): score = 0.7040816326530612  Model(depth=4, criteria=gini): score = 0.7040816326530612  Model(depth=5, criteria=gini): score = 0.6877551020408164  Model(depth=6, criteria=gini): score = 0.7 |

**Question 2c: Visualization of Best Model**



**Question 2d: Output from compute\_information\_gain**

Bolded is the topmost label (“donald”)

|  |
| --- |
| **Information Gain in Label by splitting on donald: 0.03412305015881989**  Information Gain in Label by splitting on hillary: 0.026697700148983317  Information Gain in Label by splitting on clinton: 0.007298694213891288  Information Gain in Label by splitting on korea: 0.011570446505243082  Information Gain in Label by splitting on america: 0.00849756958300818  Information Gain in Label by splitting on putin: 0.0017367934909430227 |