1. (a)

Expectation value of Z

since X and Y are independent univariate random variables, we know that:

$$E(Z) = E((X-Y)^2)$$
$$= E(X^2 \; - 2XY + \; Y^2)$$
$$= E(X^2) - 2E(X)E(Y) + E(Y^2)$$

To calculate $E(X^2)$, using the *Law of unconscious statistician* we can tell that

$$E(X^2) = \; \int_0^1 X^2 p(X)\, dx$$

where $p(X)$ is the density function. For a uniform distribution in [0, 1],

$$p(X) = \; \frac{1}{b-a} = \frac{1}{1-0} = 1$$

Where [a, b] represents the interval for the distribution, which is [0, 1] in our case. Then we can continue with our calculation,

$$E(X^2) = \; \int_0^1 X^2 \, dx \;\; = \frac{1^3 - \; 0^3}{3} = \frac{1}{3}$$
$$E(X) = \; \int_0^1 X \, dx \;\; = \frac{1^2 - \; 0^2}{2} = \frac{1}{2}$$

Since X and Y are two independent identical distributions, we know

$$E(X) = E(Y)$$
$$E(X^2) = E(Y^2)$$

Finally ,

$$E(Z) = E(X^2) - 2E(X)E(Y) + E(Y^2) = 2E(X^2) - 2E(X)^2$$
$$\boldsymbol{E(Z) = \; \frac{2}{3} - \; \frac{1}{2} \; = \; \frac{1}{6}}$$

Variance of Z:

$$Var(Z) = E(Z^2) - E(Z)^2$$
$$= E(Z^2) - \left(\frac{1}{6}\right)^2$$
$$= E((X^2 \; - 2XY + \; Y^2)^2) - \frac{1}{36}$$
$$E((X^2 \; - 2XY + \; Y^2)^2) = E(X^4 - 4X^3Y + 6X^2Y^2 - \; 4XY^3 + Y^4)$$
$$= E(X^4) - 4E(X^3)E(Y) + 6E(X^2)E(Y^2) - \; 4E(X)E(Y^3) + E(Y^4)$$
$$E(X^3) = \; \int_0^1 X^3 \, dx \;\; = \frac{1^4 - \; 0^4}{4} = \frac{1}{4}$$

$$E(X^4) = \int_0^1 X^4 \, dx \; = \frac{1^5 - 0^5}{5} = \frac{1}{5}$$

$$E(X^4) - 4E(X^3)E(Y) + 6E(X^2)E(Y^2) - 4E(X)E(Y^3) + E(Y^4) =$$

$$2E(X^4) - 8E(X^3)E(X) + 6E(X^2)^2 =$$

$$2 * \frac{1}{5} - 8 * \frac{1}{4} * \frac{1}{2} + 6 * \frac{1}{9} =$$

$$\frac{1}{15} =$$

Lastly,

$$\boldsymbol{Var(Z)} = \frac{\mathbf{1}}{\mathbf{15}} - \frac{\mathbf{1}}{\mathbf{36}} = \frac{\mathbf{7}}{\mathbf{180}}$$

1. (b) Given that d is the dimension for $R = Z_1 + \cdots + Z_d$, and using $E(Z)$ and $Var(Z)$ we computed in 1. (a), we have:
Expectation:

$$R = Z_1 + \cdots + Z_d$$

$$E(R) = E(Z_1 + \cdots + Z_d)$$

$$E(R) = E(Z_1) + \cdots + E(Z_d)$$

$$\boldsymbol{E(R)} = \frac{\mathbf{1}}{\mathbf{6}} + \cdots + \frac{\mathbf{1}}{\mathbf{6}} = \frac{\boldsymbol{d}}{\mathbf{6}}$$

Variance:

$$Var(R) = Var(Z_1 + \cdots + Z_d)$$

Since each Z_i is independent from all $Z_k, i \neq k$, $Cov(Z_i, Z_k) = 0$ for all $i, k \text{ in } d$. We have:

$$\boldsymbol{Var(R)} = \boldsymbol{Var(Z_1)} + \cdots + \boldsymbol{Var(Z_d)} = \frac{\mathbf{7}\boldsymbol{d}}{\mathbf{180}}$$

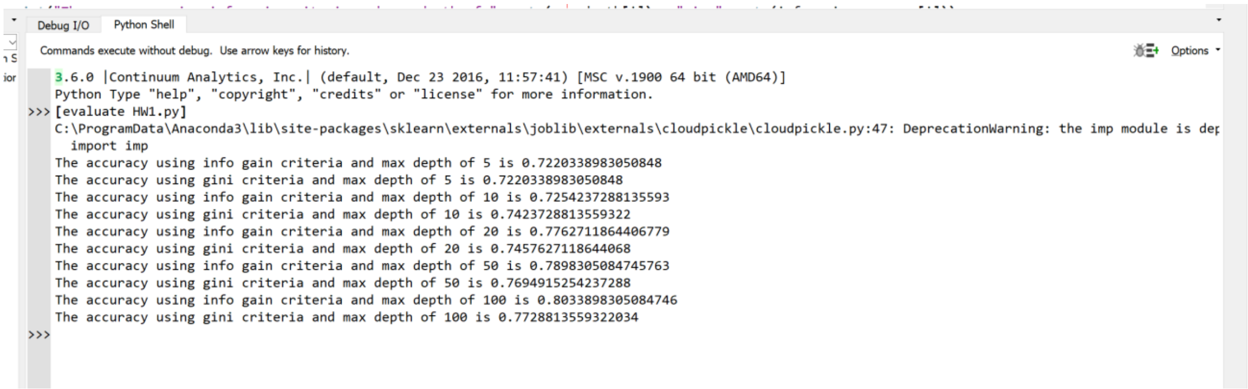
1. (c) The maximum Euclidean distance in d-dimensional space can be represented as the distance between point $(0, \ldots, 0)$ and $(1, \ldots, 1)$, which will be:

$$\sqrt{(1-0)_1^2 + (1-0)_2^2 + \cdots + (1-0)_d^2} = \sqrt{d}$$

Therefore the squared maximum Euclidean distance will simply be d.

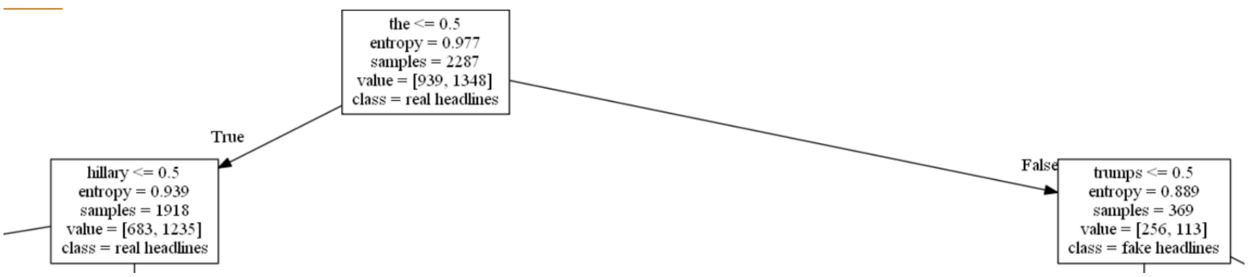
From 1. (b) we know that the variance of squared Euclidean distance between two points in unit cube with d dimensions is $Var(R) = \frac{7d}{180}$, therefore $Std(R) = \sqrt{\frac{7d}{180}}$ which is proportional to $\sqrt{d}$, where as the expectational value $E(R) = \frac{d}{6}$ is proportional to d. From this we can conclude that in higher dimensions, $Std(R)$ grows slower than $E(R)$, and is relatively smaller than $E(R)$, thus “most points are approximately the same distance”. On the contrast $E(R)$ grows linearly with d and as d gets larger, $E(R)$ gets relatively large, thus “most points are far away”.

2. (b): This is the performance from the function select_model(), using max_depth of (5, 10, 20, 50, 100) and split criteria of information gain and gini-coefficient. Performances of all models have accuracy rate of 70% to 80%, with the best performance 80% from the classifier with 100 layer and information-gain criteria.

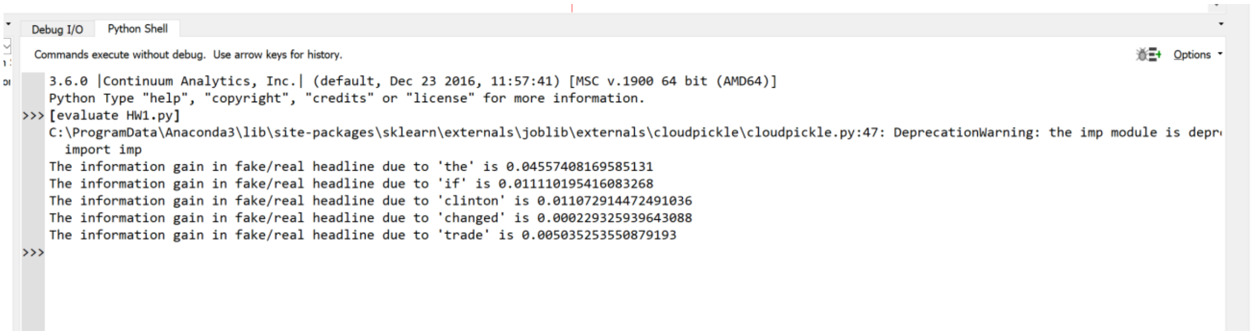


```
  3.6.0 |Continuum Analytics, Inc.| (default, Dec 23 2016, 11:57:41) [MSC v.1900 64 bit (AMD64)]
  Python Type "help", "copyright", "credits" or "license" for more information.
>>> [evaluate HW1.py]
  C:\ProgramData\Anaconda3\lib\site-packages\sklearn\externals\joblib\externals\cloudpickle\cloudpickle.py:47: DeprecationWarning: the imp module is dep
    import imp
  The accuracy using info gain criteria and max depth of 5 is 0.7220338983050848
  The accuracy using gini criteria and max depth of 5 is 0.7220338983050848
  The accuracy using info gain criteria and max depth of 10 is 0.7254237288135593
  The accuracy using gini criteria and max depth of 10 is 0.7423728813559322
  The accuracy using info gain criteria and max depth of 20 is 0.7762711864406779
  The accuracy using gini criteria and max depth of 20 is 0.7457627118644068
  The accuracy using info gain criteria and max depth of 50 is 0.7898305084745763
  The accuracy using gini criteria and max depth of 50 is 0.7694915254237288
  The accuracy using info gain criteria and max depth of 100 is 0.8033898305084746
  The accuracy using gini criteria and max depth of 100 is 0.7728813559322034
>>>
```

2 (c): This is the visualization from the first two layers of the decision tree, using the classifier with 100 decision-tree layer and information-gain split criteria.



2. (d): The is the output from the compute_information_gain() function. The five different words chosen to calculate information gain are [“the”, “if”, “Clinton”, “changed”, “trade”], with “the” being the top-most word from the decision tree in 2 (c) , the the rest of the words randomly chosen from the data set.


Debug I/O
Python Shell
Commands execute without debug. Use arrow keys for history.
Options

```
    3.6.0 |Continuum Analytics, Inc.| (default, Dec 23 2016, 11:57:41) [MSC v.1900 64 bit (AMD64)]
    Python Type "help", "copyright", "credits" or "license" for more information.
>>> [evaluate HW1.py]
    C:\ProgramData\Anaconda3\lib\site-packages\sklearn\externals\joblib\externals\cloudpickle\cloudpickle.py:47: DeprecationWarning: the imp module is depr
      import imp
    The information gain in fake/real headline due to 'the' is 0.04557408169585131
    The information gain in fake/real headline due to 'if' is 0.011110195416083268
    The information gain in fake/real headline due to 'clinton' is 0.011072914472491036
    The information gain in fake/real headline due to 'changed' is 0.000229325939643088
    The information gain in fake/real headline due to 'trade' is 0.005035253550879193
>>>
```