Carson McLean
CSC 2515
HW 1
2018-09-26

1 (a)

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

$$\mathbb{E}(Z) = \mathbb{E}(X^2 - 2XY + Y^2)$$

$$= \mathbb{E}(X^2) - \mathbb{E}(2XY) + \mathbb{E}(Y^2)$$

$$= \mathbb{E}(X^2) - 2(\mathbb{E}(X)\mathbb{E}(Y)) + \mathbb{E}Y^2$$ 2 is constant, X, Y independant

$$= \int_0^1 x^2 \frac{1}{1-0} - \left(2 \cdot \int_0^1 x \frac{1}{1-0} \cdot \int_0^1 y \frac{1}{1-0}\right) + \int_0^1 y^2 \frac{1}{1-0}$$

$$= \frac{1}{3} - \left(2 \cdot \frac{1}{2} \cdot \frac{1}{2}\right) + \frac{1}{3} = \frac{2}{3} - \frac{1}{2} = \boxed{\frac{1}{6}}$$

$$\mathbb{E}(X) = \int_a^b x f_X(x)\,dx$$ continuous random variables

$$\mathbb{E}(g(X)) = \int_a^b g(x) f_X(x)\,dx$$

Uniform PDF $= \frac{1}{b-a} = f(x)$

$$\mathrm{Var}(Z) = \mathbb{E}(Z^2) - (\mathbb{E}(Z))^2 = \mathbb{E}(X^4 - 4X^3Y + 6X^2Y^2 - 4XY^3 + Y^4) - \left(\tfrac{1}{6}\right)^2$$

$$= \left[\int_0^1 x^4 \frac{1}{1-0} - \left(4 \cdot \int_0^1 x^3 \frac{1}{1-0} \cdot \int_0^1 y \frac{1}{1-0}\right) + \left(6 \cdot \int_0^1 x^2 \frac{1}{1-0} \cdot \int_0^1 y^2 \frac{1}{1-0}\right) - \left(4 \cdot \int_0^1 x \frac{1}{1-0} \int_0^1 y^3 \frac{1}{1-0}\right) + \int_0^1 y^4 \frac{1}{1-0}\right] - \frac{1}{36}$$

$$= \frac{1}{5} - \left(4 \cdot \frac{1}{4} \cdot \frac{1}{2}\right) + \left(6 \cdot \frac{1}{3} \cdot \frac{1}{3}\right) - \left(4 \cdot \frac{1}{2} \cdot \frac{1}{4}\right) + \frac{1}{5}\Big] - \frac{1}{36}$$

$$= \frac{1}{5} - \frac{1}{2} + \frac{2}{3} - \frac{1}{2} + \frac{1}{5} - \frac{1}{36} = \frac{1}{15} - \frac{1}{36} = \boxed{\frac{7}{180}}$$

1 b

$$R = Z_1 + \dots + Z_d = dZ \quad \left(= \sum_{i=1}^{d} Z_i = \right)$$ where Z is a random variable as seen in 1 a $Z_i = (X_i - Y_i)^2$

$\mathbb{E}(R) = \mathbb{E}(dZ) = d\,\mathbb{E}(Z)$ by removing the constant d From 1 a, $\mathbb{E}(Z) = 1/6$

$= \boxed{d/6}$

$Var(R) = Var(Z_1 + \dots + Z_d) = Var\left(\sum_{i=1}^{d} Z_i\right)$

By Bienaymé Formula

$Var\left(\sum_{i=1}^{d} Z_i\right) = \sum_{i=1}^{d} Var(Z_i) = d\,Var(Z_i)$ From 1 a, $Var(Z) = 7/180$

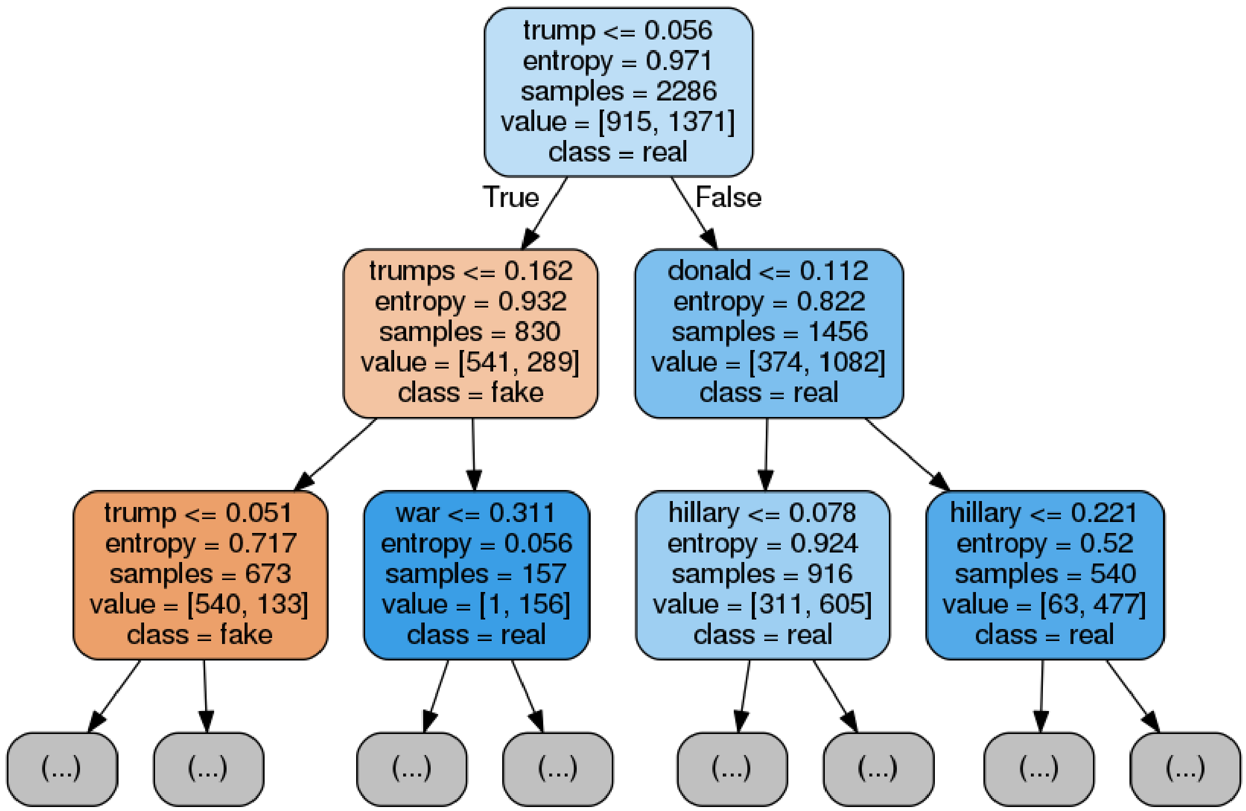
$= \boxed{\dfrac{d7}{180}}$

Q2b

##
Max Depth: 1 // Split Criteria: entropy
0.6959183673469388
Max Depth: 1 // Split Criteria: gini
0.6959183673469388
Max Depth: 4 // Split Criteria: entropy
0.7877551020408163
Max Depth: 4 // Split Criteria: gini
0.7816326530612245
Max Depth: 8 // Split Criteria: entropy
0.8244897959183674
Max Depth: 8 // Split Criteria: gini
0.8163265306122449
Max Depth: 12 // Split Criteria: entropy
0.8183673469387756
Max Depth: 12 // Split Criteria: gini
0.8040816326530612
Max Depth: 16 // Split Criteria: entropy
0.810204081632653
Max Depth: 16 // Split Criteria: gini
0.789795918367347

##
Best Results:
Max Depth: 8 // Split Criteria: entropy
0.8244897959183674

Q2c



Q2d

Top Most Split

xi = trump
Information Gain: 0.032391364179196636

Other Keywords

xi = trumps
Information Gain: 0.04736671065028819

xi = donald
Information Gain: 0.05227450569883951

xi = energy
Information Gain: 0.0006377983877460247

xi = hillary
Information Gain: 0.04058134341841224

xi = canada
Information Gain: 0.00027362801976738016

xi = somemadeupword
Information Gain: 0.0

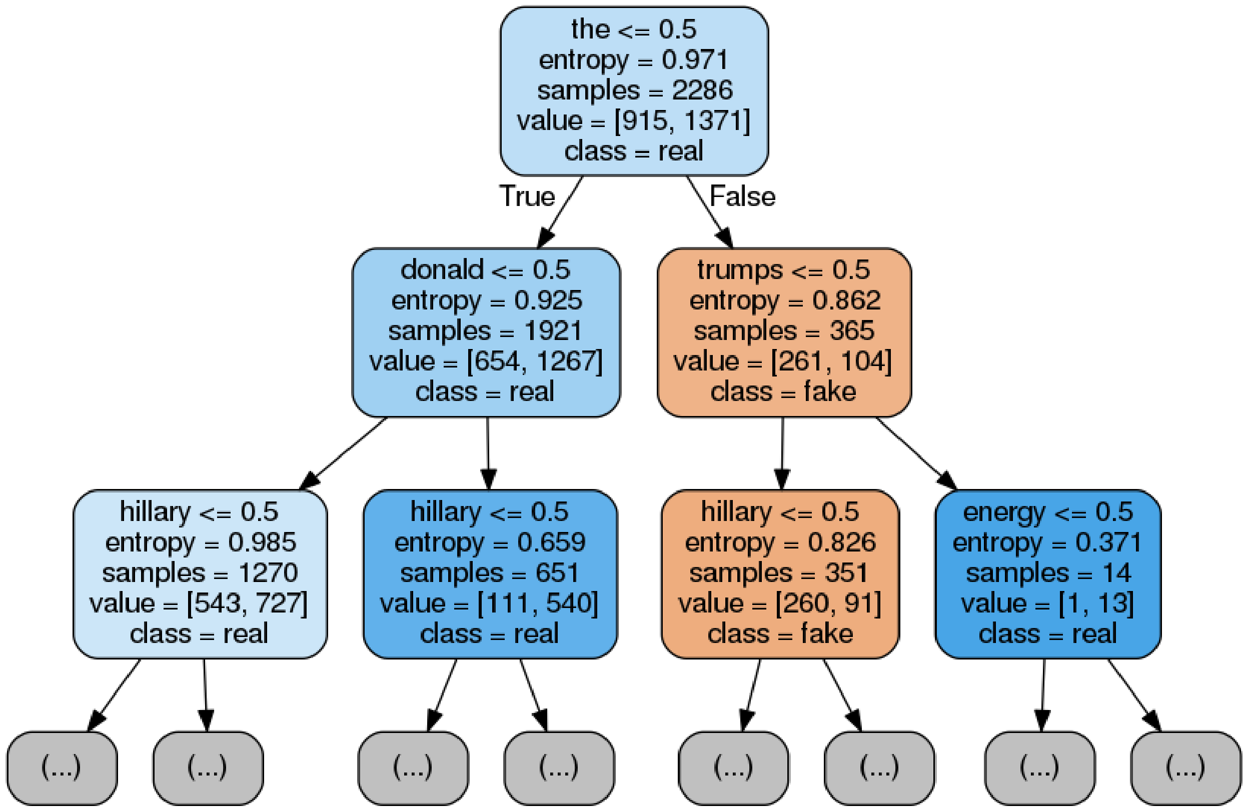
xi = the
Information Gain: 0.055942512911258624

xi = cat
Information Gain: 0.0

Explanation

In my code, I use a TfidfVectorizer() for the Classifier which in turn is used to create the Tree diagram. Yet, IG is calculated in Q2d using whether or not the x_i keyword is present or not. Therefore, there is a divergence in how the graph diagram gets generated, and which node appears at the top, versus how IG is calculated. This results in the root node word for Tfidf (ie, 'trump') to appear to not have the highest IG score. In a Decision Tree, we would expect that the root node have the highest IG so that we maximize the split and better fit the data.

So, in my code as provided, simply switching to a CountVectorizer() [line 32-33], which is more closely in line to the IG keyword present vs absent style calculation, we see the following new Tree graph:



Top Most Split

xi = the
Information Gain: 0.055942512911258624

Other Keywords

xi = trump
Information Gain: 0.032391364179196636

xi = trumps
Information Gain: 0.04736671065028819

xi = donald
Information Gain: 0.05227450569883951

xi = energy
Information Gain: 0.0006377983877460247

xi = hillary

Information Gain: 0.04058134341841224

xi = canada
Information Gain: 0.00027362801976738016

xi = somemadeupword
Information Gain: 0.0

xi = cat
Information Gain: 0.0

So, with a CountVectorizer(), the IG calculation is in line with what is seen in the decision tree graph. However, I have decided to leave Tfidf as the default. There was no requirement to use one or the other, and Tfidf achieves ~80% accuracy, while Count is 75%. Further, ‘the’ is a stop word that perhaps should have been removed.