CSC411H1 , HW1

Lino Lastella
1001237654

1

a. Let $X, Y \sim \text{Uniform}(0,1)$

Then, $E(X) = E(Y) = 1/2$.
$Var(X) = Var(Y) = 1/12$.

Let $Z = (X-Y)^2$.

Then, $E(Z) = E([X-Y]^2) = E(X^2+Y^2-2XY) =$

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

$$= \underbrace{(E(X))^2 + Var(X)}_{\text{From formula for variance}} + \underbrace{(E(Y))^2 + Var(Y)}_{\text{Same as} \leftarrow} \underbrace{-2E(X)E(Y)}_{X,Y \text{ are ind.}} =$$

$$= \frac{1}{4} + \frac{1}{12} + \frac{1}{4} + \frac{1}{12} - 2 \cdot \frac{1}{2} \cdot \frac{1}{2} = 1/6.$$

$$Var(Z) = E([Z-E(Z)]^2) = E([Z-\tfrac{1}{6}]^2) =$$

$$= E\left(Z^2 + \frac{1}{36} - \frac{2}{6}Z\right) = \frac{1}{36} - \frac{2}{6}E(Z) + E(Z^2) =$$

$$= -\frac{1}{36} + E(([X-Y]^2)^2) =$$

$$= -\frac{1}{36} + E(X^4 - 4X^3Y + 6X^2Y^2 - 4XY^3 + Y^4) =$$

$$= -\frac{1}{36} + E(X^4) - 4E(X^3Y) + 6E(X^2Y^2) - 4E(XY^3) + E(Y^4)$$

By properties of expectations,

$$E(X^4) = E(Y^4) = \int_{-\infty}^{\infty} x^4 f(x)\,dx, \quad \text{where } f(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & \text{otherwise} \end{cases}$$

therefore, $E(X^4) = \int_0^1 x^4\,dx = \frac{x^5}{5}\Big|_0^1 = \frac{1}{5}$.

Similarly, $E(X^3) = E(Y^3) = \int_0^1 x^3\,dx = \frac{x^4}{4}\Big|_0^1 = \frac{1}{4}$

The finaly expression can be reduced to

$$Var(Z) = -\frac{1}{36} + \frac{1}{5} - 4\frac{1}{4}\cdot\frac{1}{2} + 6\cdot\frac{1}{3}\cdot\frac{1}{3} - 4\cdot\frac{1}{2}\cdot\frac{1}{4} + \frac{1}{5} =$$

$$= \frac{7}{180}$$

Then $E(Z) = 1/6$ and $Var(Z) = 7/180$

b. Let X_i, Y_i be independent uniform random variable sampled from $[0,1]$, for $i = 1, 2, \dots, d$.

Let $R = Z_1 + Z_2 + \dots + Z_d$, where $Z_i = (X_i - Y_i)^2$.

Then, $E(R) = E\left(\sum_{i=1}^{d} Z_i\right) = \sum_{i=1}^{d} E(Z_i) = \underbrace{\frac{1}{6} + \frac{1}{6} + \dots + \frac{1}{6}}_{d \text{ times}} =$

$= d/6$.

$$Var(R) = E\left(\left[R - E(R)\right]^2\right) = E\left(R^2 + \frac{d^2}{36} - \frac{2d}{6} R\right) =$$

$$= \frac{d^2}{36} - \frac{d^2}{18} + E(R^2) = -\frac{d^2}{36} + E(R^2) =$$

$$= -\frac{d^2}{36} + E\left(\left[Z_1 + \dots + Z_d\right]^2\right) =$$

$$= -\frac{d^2}{36} + \underbrace{\sum_{i=1}^{d} E(Z_i^2) + 2 \sum_{i \neq j} E(Z_i Z_j)}_{\text{Expanding the square of a poly.}}$$

$$= -\frac{d^2}{36} + \underbrace{\sum_{i=1}^{d}\left(\left[E(Z_i)\right]^2 + Var(Z_i)\right)}_{\text{Formula for variance}} + \underbrace{2 \sum_{i \neq j} E(Z_i) E(Z_j)}_{\text{Independence.}}$$

$$= -\frac{d^2}{36} + \sum_{i=1}^{d}\left(\frac{1}{36} + \frac{7}{180}\right) + 2 \sum_{i \neq j}^{d} \frac{1}{36}$$

$$= -\frac{d^2}{36} + \frac{d}{15} + 2 \sum_{i \neq j} \frac{1}{36}$$

Now notice that $\sum_{i \neq j}^{d}$ just means to select two different indices from 1 to d. We can do this in $\binom{d}{2}$ ways.

$$Var(R) = \frac{d}{15} - \frac{d^2}{36} + 2\binom{d}{2}\cdot\frac{1}{36} = \frac{d}{15} - \frac{d^2}{36} + \frac{d!}{(d-2)!\,2!}\cdot\frac{2}{36} =$$

$$= \frac{d}{15} - \frac{d^2}{36} + \frac{d(d-1)}{36} = \frac{12d - 5d^2 + 5d^2 - 5d}{180} = \frac{7d}{180}$$

Then, $E(R) = d/6$ and $Var(R) = 7d/180$

Criterion: entropy

max_depth: 1

Accuracy for this test is: 66.598778 %

max_depth: 3

Accuracy for this test is: 67.820774 %

max_depth: 6

Accuracy for this test is: 71.486762 %

max_depth: 9

Accuracy for this test is: 73.523422 %

max_depth: 12

Accuracy for this test is: 74.338086 %

Criterion: gini

max_depth: 1

Accuracy for this test is: 66.598778 %

max_depth: 3

Accuracy for this test is: 70.875764 %

max_depth: 6

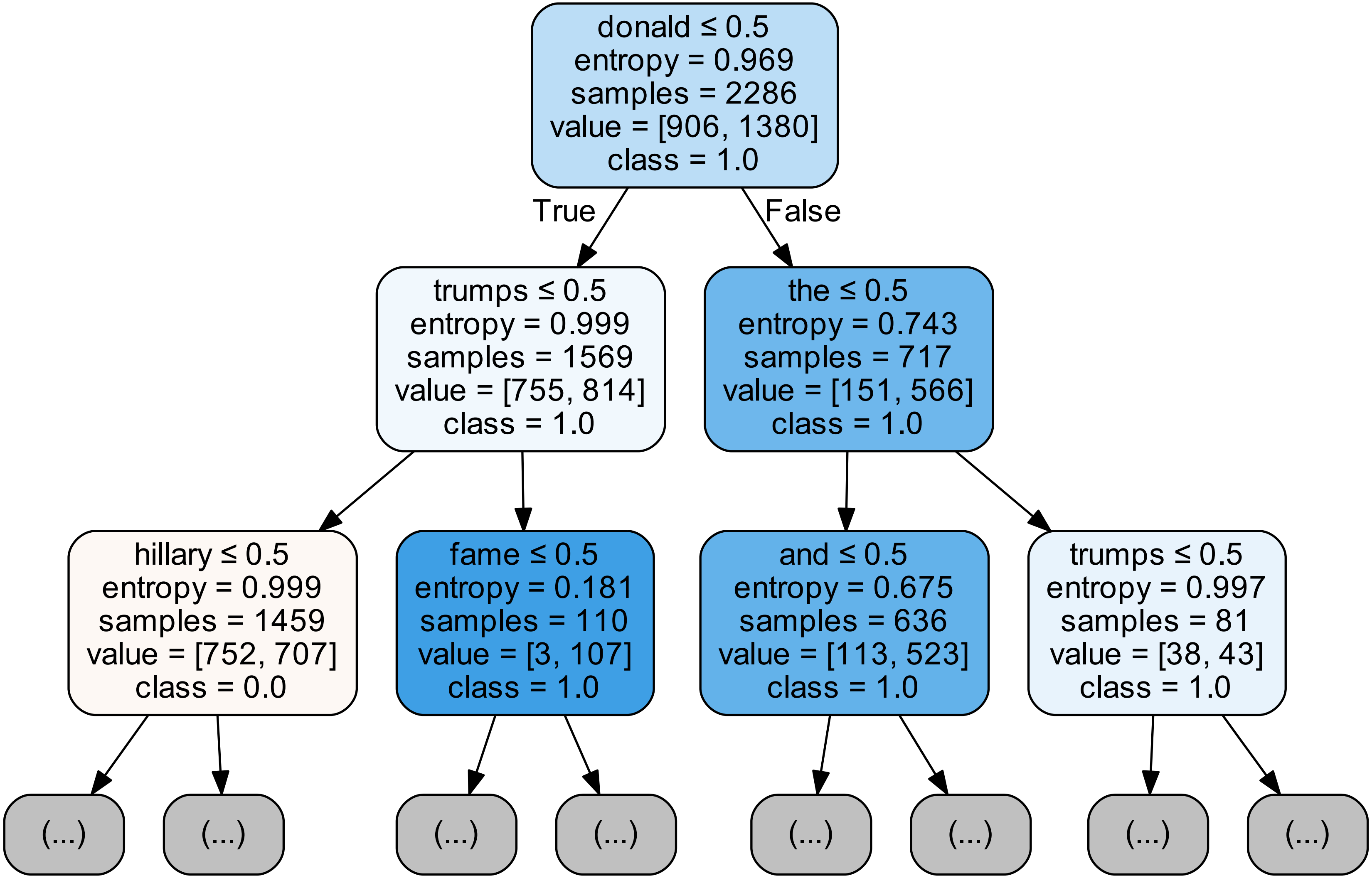
Accuracy for this test is: 71.690428 %

max_depth: 9

Accuracy for this test is: 73.116090 %

max_depth: 12

Accuracy for this test is: 73.523422 %


donald ≤ 0.5
entropy = 0.969
samples = 2286
value = [906, 1380]
class = 1.0
True
False
trumps ≤ 0.5
entropy = 0.999
samples = 1569
value = [755, 814]
class = 1.0
the ≤ 0.5
entropy = 0.743
samples = 717
value = [151, 566]
class = 1.0
hillary ≤ 0.5
entropy = 0.999
samples = 1459
value = [752, 707]
class = 0.0
fame ≤ 0.5
entropy = 0.181
samples = 110
value = [3, 107]
class = 1.0
and ≤ 0.5
entropy = 0.675
samples = 636
value = [113, 523]
class = 1.0
trumps ≤ 0.5
entropy = 0.997
samples = 81
value = [38, 43]
class = 1.0
(...)
(...)
(...)
(...)
(...)
(...)
(...)
(...)

"the"	0.067717
"and"	0.014410
"donald"	0.042960
"trumps"	0.044293
"fame"	0.002266
"hillary"	0.043895
"sledgehammer"	0.001200