CSC311 HW1

1/24/20 Due: 2/3/20

CSC 311 - Homework 1

1. a) Given: $Z = |X-Y|^2$

Then: $E(Z) = E(|X-Y|^2)$

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

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

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

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

↳ b/c X & Y are from same distribution

$\Rightarrow E(X^n) = E(Y^n) \quad \forall n \in R$ ①

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

• $X \sim$ uniform $\Rightarrow f(x) = \frac{1}{b-a}$

$= \frac{1}{1-0} = 1$

• $E(Z) = 2(E(X^2) - E(X)^2)$

$= 2(\int_0^1 x^2 \cancel{f(x)}\,dx - (\int_0^1 x \cancel{f(x)}\,dx)^2)$

$= 2(\frac{1}{3}x^3|_0^1 - (\frac{1}{2}x^2|_0^1)^2)$

$= 2(\frac{1}{3} - \frac{1}{4})$

$= 2(\frac{4}{12} - \frac{3}{12})$

$= \frac{2}{12} = \frac{1}{6}$ *

• $Var(Z) = E(Z^2) - (E(Z))^2$

• $E(Z^2) = E((|X-Y|^2)^2)$

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

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

$+ Y^2X^2 - 2XY^3 + Y^4)$

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

$+ E(Y^4)$

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

$- 4E(X)E(X^3) + E(X^4)$ by ①

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

$= 2\int_0^1 x^4 dx - 8(\int_0^1 x^3 dx)(\int_0^1 x\,dx) + 6(\int_0^1 x^2 dx)(\int_0^1 x^2 dx)$

$= 2(\frac{1}{5}) - 8(\frac{1}{4})(\frac{1}{2}) + 6(\frac{1}{3})(\frac{1}{3})$

$= 2/5 - 1 + 6/9$

$= 18/45 - 45/45 + 30/45 = 3/45 = 1/15$ *

$\therefore\ E(Z) = \boxed{1/6}$

$var(Z) = E(Z^2) - (E(Z))^2$

$= 1/15 - (1/6)^2$

$= 1/15 - 1/36 = 36/540 - 15/540$

$= \boxed{7/180}$

b) WTF: $E(|X-Y|_2)$ and $var(|X-Y|_2)$ in d dimension

sol:

- Let $R_d = \|X-Y\|_2^2 = \sum_{i=1}^{d} Z_i$

Then $E(R_d) = E\left(\sum_{i=1}^{d} Z_i\right)$

$= \sum_{i=1}^{d} E(Z_i)$

$= \sum_{i=1}^{d} E(Z)$

b/c $E(Z_i) = E(|X_i - Y_i|^2) = E(|X_j - Y_j|^2) = E(Z_j)$ $\forall i \neq j$

$= d\,E(Z)$ — $E(Z) = 1/6$ from 1a)

$\boxed{= d/6}$.

- $Var(R_d) = var\left(\sum_{i=1}^{d} Z_i\right)$

$= \sum_{i=1}^{d} var(Z_i)$

$= \sum_{i=1}^{d} var(Z)$

$= d\,var(Z)$

$\boxed{= 7d/180}$

$\therefore\ E(R_d) = d/6,\ var(R_d) = 7d/180$

c) Let $\sqrt{d}$ be max possible distance between 2 points in d-dimensional hypercube

① $$\lim_{d\to\infty} \frac{E(R_d)}{\sqrt{d}} = \lim_{d\to\infty} \frac{d/6}{\sqrt{6}} = \infty$$

↳ This shows that as $d \to \infty$, the average distance between 2 points scales faster than the max possible distance. This supports "most points are far away."

② $$\lim_{d\to\infty} \frac{var(R_d)}{E(R_d)} = \frac{\frac{7d}{180}}{\frac{d}{6}} = \frac{42}{180}$$

↳ The variance scales slower than expected value. Thus "most points are approx. the same distance" is justified.

2. a) Given: • $0 < P(x) \leq 1 \quad \forall x \in X$

• $\log_2(x) \geq 0 \quad \forall x \geq 1$

Pf: $H(X) = \sum_x P(x) \log_2\left(\frac{1}{P(x)}\right)$

$\Rightarrow \sum_x 1 \cdot \log_2\left(\frac{1}{1}\right) \leq H(x) < \sum_x 1 \cdot \left(\frac{1}{0^+}\right)$

$\Rightarrow 0 \leq H(x) < \infty$

$\therefore H(X) \geq 0 \quad \forall x \in X$

b) $H(X,Y) = -\sum_x \sum_y p(x,y) \log_2 P(x,y)$

$= -\sum_x \sum_y p(x,y) \cdot \log_2 P(x) P(Y|X)$

$= -\sum_x \sum_y (p(x,y) \log_2(P(x) P(y)))$ — X & Y independent

$= -\sum_x \sum_y p(x,y) \log_2 P(x)$

$- \sum_x \sum_y p(x,y) \log_2 P(y)$

$= -\sum_x P(x) \log_2 P(x) - \sum_y P(y) \log_2 P(y)$

$= H(X) + H(Y)$

Given : $H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 P(y|x)$

c) Pf: $H(X, Y)$

$= -\sum_{x \in X} \sum_{y \in Y} P(X,Y) \log_2 P(X,Y)$

$= -\sum_x \sum_y P(x,y) \log_2 (P(x) P(y|x))$

$= -\sum_x \sum_y p(x,y) \log_2 p(x)$

$\quad - \sum_x \sum_y p(x,y) \log_2 p(y|x)$

$= -\sum_x p(x) \log_2 p(x)$

$\quad - \sum_x \sum_y p(x,y) \log_2 p(y|x)$

$= H(X) - H(Y|X)$

d) WTS : $KL(p||q) \geq 0$

Given: $p(x) > 0$, $q(x) > 0$

Pf: $KL(p||q) = \sum_x p(x) \log_2 \left(\frac{p(x)}{q(x)}\right)$

$= \sum_x p(x) \left(-\log_2 \frac{q(x)}{p(x)}\right)$

$\geq -\log_2 \sum_x \cancel{p(x)} \frac{q(x)}{\cancel{p(x)}}$ (By Jensen's Ineq.)

$= -\log_2 (1)$

$= 0$

e) WTS: $I(Y;X) = KL(P(X,Y) || P(X)P(Y))$

Given: $I(Y;X) = H(Y) - H(Y|X)$ ①

$H(Y) = -\sum_{x \in X} P(x) \log_2 P(x)$ ②

$H(Y|X) = -\sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 P(y|x)$ ③

Pf: $KL(P(x,y) || P(x)P(y))$

$= \sum_x \sum_y P(x,y) \log_2 \left(\frac{P(x,y)}{P(x)P(y)}\right)$

$= \sum_x \sum_y P(x,y) \log_2 \left(\frac{\cancel{P(x)} P(y|x)}{\cancel{P(x)} P(y)}\right)$

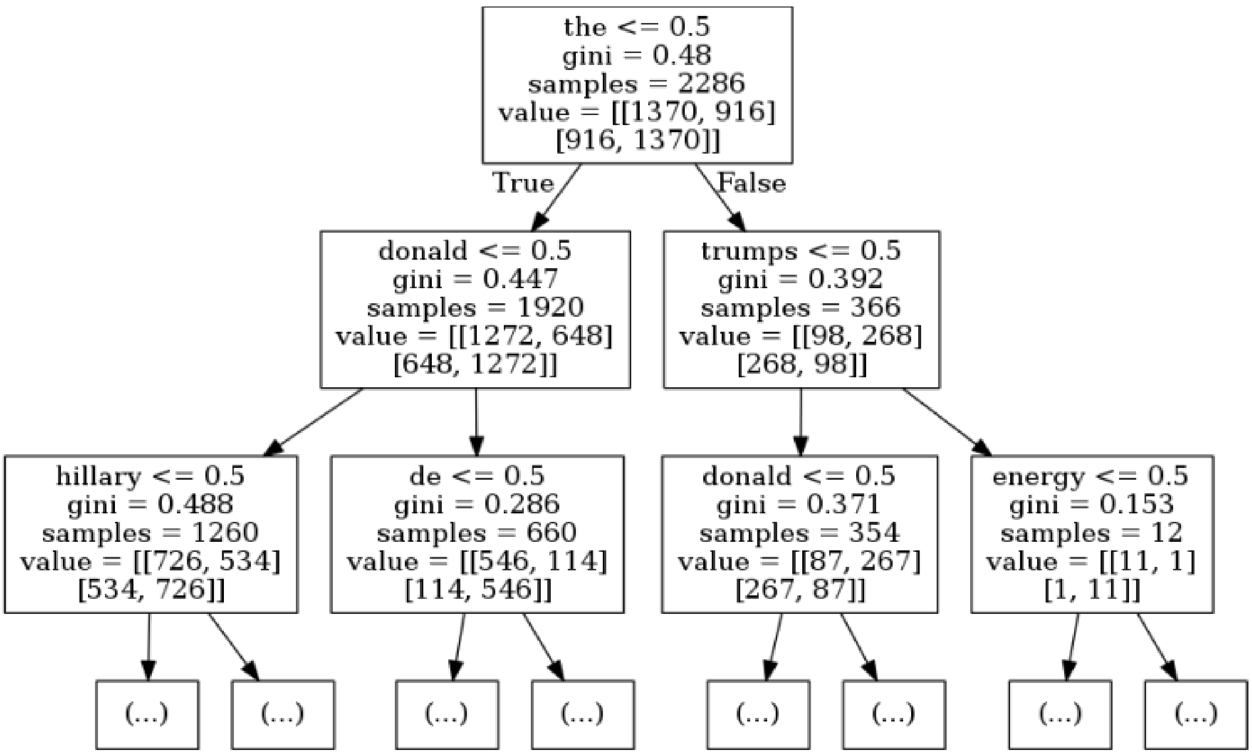
$= \sum_x \sum_y P(x,y) \log_2 P(y|x) - \sum_x \sum_y P(x,y) \log_2 P(y)$

$= -H(Y|X) - \sum_y P(y) \log_2 P(y)$

$= -H(Y|X) + H(Y)$

$= I(Y;X)$

Decision Tree With Depth 50 and Criterion Gini (only first 3 layer shown)



Prediction accuracy on test set: 0.7448979591836735

Sample Keywords:

information gain for the: 0.9042833975606189

information gain for donald: 0.834642716332616

information gain for trumps: 0.1929506731451265

information gain for hillary: 0.518867440494653

information gain for de: 0.881657886020225

information gain for energy: 0.9709392360332805

information gain for wtf: 0.00042251892876064265

information gain for riots: 0.6508499405789779

information gain for berserk: 0.00042251892876064265

K Nearest Neighbor Hyperparameter Performance