Homework 1
Submitted By- Pooja Bhatia

Solution

Q1(a)

Probability Distribution Function (PDF)

$X \in [a, b]$

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

in the question $x \in [0, 1]$

$$f(x) = \begin{cases} \frac{1}{1-0} & \text{for } 0 \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

similarly $Y \in [0, 1]$

$$f(y) = \begin{cases} 1 & Y \in [0, 1] \\ 0 & \text{otherwise.} \end{cases}$$

Now,

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

By using the below mentioned properties

Property 1

$E[aX] = aE[X]$; where X is random variable and $a \in R$ is constant

Property 2

$E[X_1 + X_2 \dots X_K] = E[X_1] + E[X_2] \dots E[X_K]$

if $X_1, X_2 \dots X_K$ are K random variables.

$$E[z] = E[X^2] - 2E[XY] + E[Y^2]$$

By using property 3:

$E[XY] = E[X]E[Y]$ as X and Y are independent.

$$E[z] = E[X^2] - 2E[X]E[Y] + E[Y^2]$$

Since X and Y are continuous radom variable ranging from $0 \le X \le 1$ and $0 \le Y \le 1$

$$E[z] = \int_0^1 x^2 f(x) \cdot dx - 2 \int_0^1 x f(x) dx \int_0^1 y f(y) dy + \int_0^1 y^2 f(y) dy$$

$$= \left[\frac{x^3}{3}\right]_0^1 - 2\left\{\left[\frac{x^2}{2}\right]_0^1 \left[\frac{y^2}{2}\right]_0^1\right\} + \left[\frac{y^3}{3}\right]_0^1$$

$$E[Z] = \frac{1}{3} - 2\left\{\frac{1}{2} \times \frac{1}{2}\right\} + \frac{1}{3}$$

$$= \frac{2}{3} - \frac{1}{2} = \frac{1}{6} = 0.1667 \text{ (approximately)}$$

Note: calculation shown by using scipy.integrate.quad and scipy.integrate.nquad (ahead)

Ans, Expectation of Random Variable Z is $\boxed{0.1667}$

Variance of random variable Z

$$Var[Z] = E[Z^2] - E[Z]^2$$

$$= E[((X-Y)^2)^2] - (1/6)^2$$

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

Now, using property 1, 2, 3 as earlier

$$Var[Z] = E[X^4] - 4E[X^3]E[Y] + 6E[X^2]E[Y^2] - 4E[X]E[Y^3] + E[Y^4] - \left(\frac{1}{6}\right)^2$$

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

$$- 4\int_0^1 x f(x)\,dx \int_0^1 y^3 f(y)\,dy + \int_0^1 y^4 f(y)\,dy - \left(\frac{1}{6}\right)^2$$

$$Var[z] = \left[\frac{x^5}{5}\right]_0^1 - 4\left\{\left[\frac{x^4}{4}\right]_0^1 \left[\frac{y^2}{2}\right]_0^1\right\} + 6\left\{\left[\frac{x^3}{3}\right]_0^1 \left[\frac{y^3}{3}\right]_0^1\right\}$$

$$- 4\left\{\left[\frac{x^2}{2}\right]_0^1 \left[\frac{y^4}{4}\right]_0^1\right\} + \left[\frac{y^5}{5}\right]_0^1 - \left(\frac{1}{6}\right)^2$$

$$= \frac{1}{5} - 4\left\{\frac{1}{4}\times\frac{1}{2}\right\} + 6\left\{\frac{1}{3}\times\frac{1}{3}\right\} - 4\left\{\frac{1}{2}\times\frac{1}{4}\right\}$$

$$+ \frac{1}{5} - \left(\frac{1}{6}\right)^2$$

$$= \frac{1}{5} - \frac{1}{2} + \frac{2}{3} - \frac{1}{2} + \frac{1}{5} - \left(\frac{1}{6}\right)^2$$

$$= \frac{2}{5} + \frac{2}{3} - 1 - \left(\frac{1}{6}\right)^2$$

$$= \frac{16-15}{15} - \left(\frac{1}{6}\right)^2$$

$$= \frac{1}{15} - \frac{1}{36} = 0.03889 \text{ (approximately)}$$

Note: Calculation also shown using scipy.integrate.quad and scipy.integrate.nquad (ahead)

Ans, Variance of random Variable Z is $\boxed{0.03889}$

```
#HomeWork 1 # Q 1 (a)
# Evaluating Integral numerically using scipy.integrate.quad and scipy.integrate.nquad

#Calculating Expectation of Z
from scipy import integrate
import numpy as np
from scipy.integrate import quad

def f2(x, y):
    return x*y
def bounds_y():
    return [0, 1]
def bounds_x(y):
    return [0, 1]
I2=(integrate.nquad(f2,[bounds_x, bounds_y]))

def f1(x):
    return x**2
def f3(y):
    return y**2

I1=quad(f1,0,1)
I3=quad(f3,0,1)

E_Z=I1[0]-2*I2[0]+I3[0]
print ("Expectation of Z is", E_Z)
```

Expectation of Z is 0.16666666666666668

```
#Calculating Variance of Z
def f2(x, y):
    return x**3*y
def bounds_y():
    return [0, 1]
def bounds_x(y):
    return [0, 1]
I2=(integrate.nquad(f2,[bounds_x, bounds_y]))
def f3(x, y):
    return x**2*y**2
I3=(integrate.nquad(f3,[bounds_x, bounds_y]))
def f4(x, y):
    return x*y**3
I4=(integrate.nquad(f4,[bounds_x, bounds_y]))

def f1(x):
    return x**4
def f5(y):
    return y**4

I1=quad(f1,0,1)
I5=quad(f5,0,1)

Var_Z= I1[0]-4*I2[0]+6*I3[0]-4*I4[0]+I5[0]-E_Z**2
print ("Variance of Z is", Var_Z)
```

Variance of Z is 0.038888888888888945

Solution

Q1(b)

Squared Euclidean distance

$R = Z_1 + \ldots\ldots + Z_d$, where

$Z_i = (X_i - Y_i)^2$

Since points independently from a unit cube in d ~~md~~ imension each coordinate is sampled independently from $[0,1]$

i.e

$X_1 \ldots\ldots X_d$ and $Y_1 \ldots\ldots Y_d$

$E[R] = E[Z_1 + Z_2 + Z_3 \ldots Z_d]$

$= E[Z_1] + E[Z_2] \ldots\ldots E[Z_d]$

Since from the previous question we can say that $E[Z_1] = E[Z_2] \ldots = E[Z_d] = E[Z]$

$E[R] = d\,E[Z] = d \times \frac{1}{6} = 0.1667 \times d$ //substituting value from previous part

Ans, $E[R]$ is $\boxed{0.1667 \times d}$ (approximately)

i.e ~~E[Z]~~ $\boxed{E[R] = d\,E[Z]}$ 1

Now, similarly

$Var[R] = E[R^2] - E[R]^2$

$Var[R] = Var[Z_1 + \cdots Z_d]$

$= Var[(X_1 - Y_1)^2 + \cdots (X_d - Y_d)^2]$

Now since X and Y are independent

Property $Var(X+Y) = Var(X) + Var(Y)$
If X and Y are independent

$Var[R] = Var[(X_1 - Y_1)^2] + \cdots Var[(X_d - Y_d)^2]$
(d times)

$Var[R] = d\, Var[Z]$

// since $Z_i = (X_i - Y_i)^2$

Ans $\boxed{Var[R] = d\, Var[Z]}$

Please Turn Over
(PTO)

Ans, $\boxed{Var[R] = d \; Var[Z]}$ ②

$\boxed{Var[R] = d \times 0.03889}$

// substituting value of Var[Z] from previous question

\# Just as an Example

Now say we take $d=3$ (3 dimension)

$Var[R] = (3) \times 0.03889$ // from 2

$= 0.11667$

$E[R] = 0.1667 \times 3 = 0.5001$

Output of Q 2 (b)

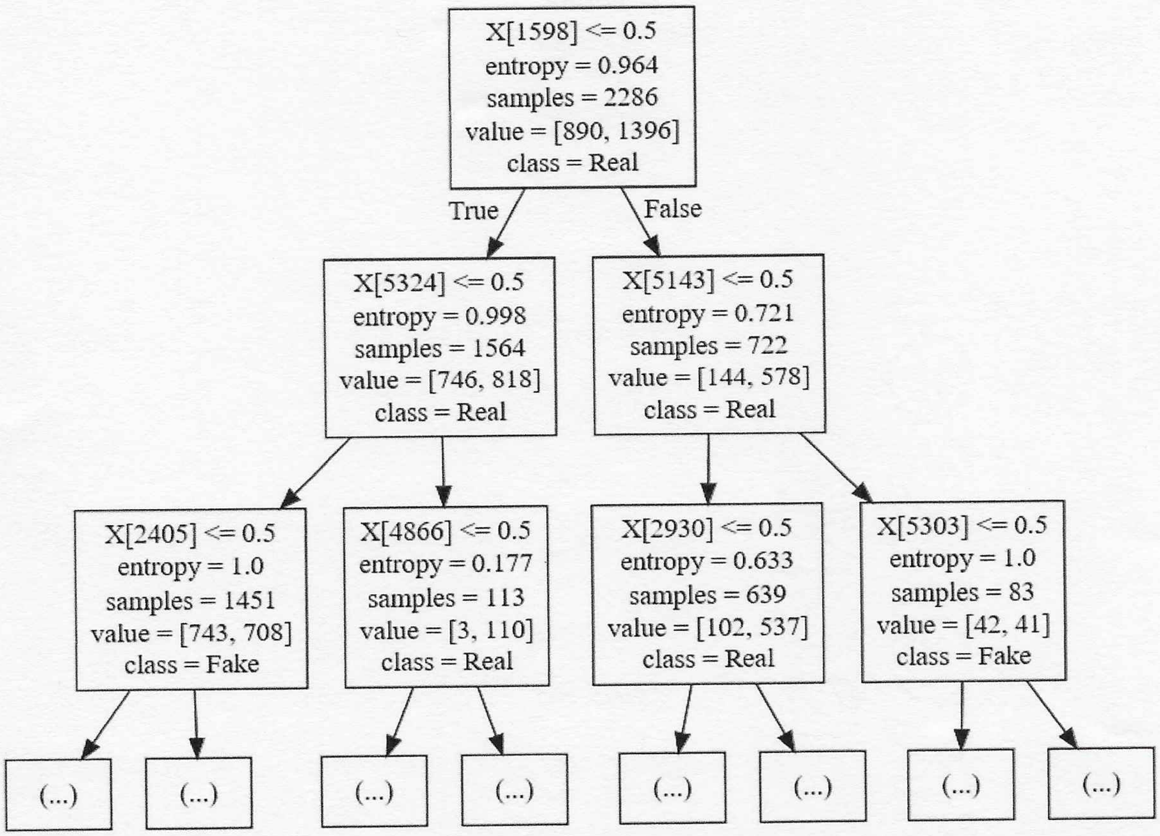
```
Dataset clean_real Lenght::  1968
Dataset clean_real Shape::  (1968, 1)
Dataset clean_fake Lenght::  1298
Dataset clean_fake Shape::  (1298, 1)
Accuracy for split criteria gini coefficient and max depth 15 is 76.12244897959184 %
Accuracy for split criteria gini coefficient and max depth 30 is 74.6938775510204 %
Accuracy for split criteria gini coefficient and max depth 45 is 73.87755102040816 %
Accuracy for split criteria gini coefficient and max depth 20 is 75.3061224489796 %
Accuracy for split criteria gini coefficient and max depth 50 is 73.26530612244898 %
Accuracy for split criteria information gain and max depth 15 is 74.08163265306122 %
Accuracy for split criteria information gain and max depth 30 is 75.3061224489796 %
Accuracy for split criteria information gain and max depth 45 is 76.53061224489795 %
Accuracy for split criteria information gain and max depth 20 is 75.3061224489796 %
Accuracy for split criteria information gain and max depth 50 is 77.9591836734694 %
```

Output of Q 2 (c)

```
#output in test form
digraph Tree {
node [shape=box] ;
0 [label="X[1598] <= 0.5\nentropy = 0.964\nsamples = 2286\nvalue = [890, 1396]\nclass = Real"] ;
1 [label="X[5324] <= 0.5\nentropy = 0.998\nsamples = 1564\nvalue = [746, 818]\nclass = Real"] ;
0 -> 1 [labeldistance=2.5, labelangle=45, headlabel="True"] ;
2 [label="X[2405] <= 0.5\nentropy = 1.0\nsamples = 1451\nvalue = [743, 708]\nclass = Fake"] ;
1 -> 2 ;
3 [label="(...)"] ;
2 -> 3 ;
342 [label="(...)"] ;
2 -> 342 ;
351 [label="X[4866] <= 0.5\nentropy = 0.177\nsamples = 113\nvalue = [3, 110]\nclass = Real"] ;
1 -> 351 ;
352 [label="(...)"] ;
351 -> 352 ;
357 [label="(...)"] ;
351 -> 357 ;
358 [label="X[5143] <= 0.5\nentropy = 0.721\nsamples = 722\nvalue = [144, 578]\nclass = Real"] ;
0 -> 358 [labeldistance=2.5, labelangle=-45, headlabel="False"] ;
359 [label="X[2930] <= 0.5\nentropy = 0.633\nsamples = 639\nvalue = [102, 537]\nclass = Real"] ;
358 -> 359 ;
360 [label="(...)"] ;
359 -> 360 ;
519 [label="(...)"] ;
359 -> 519 ;
520 [label="X[5303] <= 0.5\nentropy = 1.0\nsamples = 83\nvalue = [42, 41]\nclass = Fake"] ;
358 -> 520 ;
521 [label="(...)"] ;
520 -> 521 ;
522 [label="(...)"] ;
520 -> 522 ;
}
```

Output of Q 2 (c)

#Output in Image form



Output of Q 2 (d)

```
Information gain for topmost split in the decision tree of previous part is 0.053585095629213875
Information gain for topmost split in the decision tree which was generated again is 0.05407624529829791
Information gain for topmost split in the decision tree which was generated again is 0.051125248744604
```