

Question 1a)

$$E[Z] = E[(X-Y)^2] = E[X^2 + Y^2 + 2XY] = E[X^2] + E[Y^2] - E[2XY] \quad (1)$$

$$E[Y] = E[X] = \int_0^1 x dx = \left[ \frac{1}{2} x^2 \right]_0^1 = \frac{1}{2} \quad (2)$$

$$E[Y^2] = E[X^2] = \int_0^1 x^2 dx = \left[ \frac{1}{3} x^3 \right]_0^1 = \frac{1}{3} \quad (3)$$

$$E[Y^3] = E[X^3] = \int_0^1 x^3 dx = \left[ \frac{1}{4} x^4 \right]_0^1 = \frac{1}{4} \quad (4)$$

$$E[Y^4] = E[X^4] = \int_0^1 x^4 dx = \left[ \frac{1}{5} x^5 \right]_0^1 = \frac{1}{5} \quad (5)$$

Substitute (2) (3) (4) (5) into (1)

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

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

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

$$E[Z] = \frac{1}{6}$$

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

$$Z^2 = (X-Y)^4 = (X^2 + Y^2 - 2XY)(X^2 + Y^2 - 2XY) = X^4 + 6X^2Y^2 - 4X^3Y - 4XY^3 + Y^4 \quad (2)$$

substitute (2) into (1)

$$\text{Var}[Z] = E[X^4] + 6E[X^2]E[Y^2] - 4E[X^3]E[Y] - 4E[X]E[Y^3] + E[Y^4] - E[Z]^2$$

$$\text{Var}[Z] = 2 \times \frac{1}{5} + 6 \times \frac{1}{3} \times \frac{1}{3} - 8 \times \frac{1}{4} \times \frac{1}{2} - \frac{1}{36}$$

$$\text{Var}[Z] = \frac{7}{180}$$

Answer: The expectation is  $\frac{1}{6}$  and variance is  $\frac{7}{180}$

Question 1b)

$$\therefore R = z_1 + z_2 + \dots + z_d = \sum_1^d z_i$$

$$z_i = (x_i - y_i)^2$$

$$\therefore E[R] = \sum_1^d E[z_i] = \boxed{d E[z]} = \boxed{\frac{1}{6} d}$$

$$\therefore \text{Var}[R] = \sum_1^d \text{Var}[z_i] = \boxed{d \text{Var}[z]} = \boxed{\frac{7}{180} d}$$

Answer: The expectation is  $d E[z_i]$  which is  $\frac{1}{6} d$ .

The variance is  $d \text{Var}[z]$  which is  $\frac{7}{180} d$ .

Question 3a)

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

$$\text{Find limit: } \lim_{P(x) \rightarrow 0} P(x) \log_2 \left( \frac{1}{P(x)} \right) = 0$$

$$\therefore \sum_x P(x) \log_2 \left( \frac{1}{P(x)} \right) \geq 0$$

$$\therefore H(X) \geq 0$$

QED

Question 3b)

$$KL(P||q) = \sum_x P(x) \log_2 \frac{P(x)}{q(x)} = - \sum_x P(x) \log_2 \frac{q(x)}{P(x)} = E \left[ -\log_2 \frac{q(x)}{P(x)} \right] \quad (1)$$

$\therefore \phi(E\{x\}) \leq E\{\phi(x)\}$  from Jensen's inequality

$$\therefore -\log_2 E \left[ \frac{q(x)}{P(x)} \right] \leq E \left[ -\log_2 \frac{q(x)}{P(x)} \right] \quad (2)$$

$$-\log_2 E \left[ \frac{q(x)}{P(x)} \right] = -\log_2 \sum_x P(x) \frac{q(x)}{P(x)} = -\log_2 \sum_x q(x) = -\log_2 1 = 0 \quad (3)$$

substitute (3) into (2)

$$E \left[ -\log_2 \frac{q(x)}{P(x)} \right] \geq -\log_2 E \left[ \frac{q(x)}{P(x)} \right]$$

$$E \left[ -\log_2 \frac{q(x)}{P(x)} \right] \geq 0 \text{ substitute into (1)}$$

$$KL(P||q) = E \left[ -\log_2 \frac{q(x)}{P(x)} \right] \leftarrow$$

$$\therefore KL(P||q) \geq 0$$

Question 3c)

(pg 3)

$$KL[P(x,y) || P(x)P(y)] \quad (1)$$

$$= \sum_{x,y} P(x,y) \log \left[ \frac{P(x,y)}{P(x)P(y)} \right]$$

$$= \sum_{x,y} P(x,y) \log \left[ \frac{P(x)P(y|x)}{P(x)P(y)} \right]$$

$$= \sum_{x,y} P(x,y) [\log P(y|x) - \log P(y)]$$

$$= -\sum_y \log P(y) \sum_x P(x,y) + \sum_{x,y} P(x,y) \log P(y|x)$$

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

$$= H(y) - \sum_x P(x) H(Y|X=x) = H(y) - H(Y|X) = I(Y;X) \text{ by definition (2)}$$

$$\begin{aligned} \therefore (1) &= (2) \\ KL[P(x,y) || P(x)P(y)] \\ &= I(Y;X) \end{aligned}$$

### Question 2 b:

Output of the models accuracy scores with green highlight as the highest accuracy score:

```
{(3, 'gini'): 0.710204081632653,  
(3, 'entropy'): 0.6326530612244898,  
(5, 'gini'): 0.710204081632653,  
(5, 'entropy'): 0.7122448979591837,  
(10, 'gini'): 0.7326530612244898,  
(10, 'entropy'): 0.7306122448979592,  
(15, 'gini'): 0.736734693877551,  
(15, 'entropy'): 0.736734693877551,  
(20, 'gini'): 0.7551020408163265,  
(20, 'entropy'): 0.7448979591836735 }
```

### Question 2c:

First two layers of the decision tree with parameters max\_depth=20, and criterion = 'gini'

