Data Mining
Assignment-1

Ayush Garg

Ans: 1.

(a) (i) P(first card is a heart) = $\frac{13}{52}$.
After picking up one heart from the deck of cards, only 51 cards are left out of which 12 are hearts.
Probability (second card is heart | first card is a heart) = $\frac{12}{51}$.

(ii)

$$\begin{aligned} P(\text{None of cards is hearts}) &= \frac{39}{52} * \frac{38}{51} \\ &= \frac{19}{34} \end{aligned}$$

$$\begin{aligned} P(\text{Atmost one card is heart}) &= (\frac{39}{52} * \frac{38}{51}) + (\frac{2 * 13}{52} * \frac{39}{51}) \\ &= \frac{247}{1156} \end{aligned}$$

$$\begin{aligned} P(\text{none of the cards are hearts}|\text{ atmost one card is heart}) &= \frac{\frac{19}{34}}{\frac{893}{1156}} \\ &= \frac{34}{47} \end{aligned}$$

(b) (i) This case has two possibilities :

1. ace is drawn from the first deck and then ace is drawn from the second deck
2. ace is not drawn from the first deck and then ace is drawn from the second deck

$$
\begin{aligned}
P(\text{card drawn from the second deck is an ace}) &= (\frac{4}{52} * \frac{5}{53}) + (\frac{48}{52} * \frac{4}{53}) \\
&= \frac{5}{689} + \frac{48}{689} \\
&= \frac{53}{689}
\end{aligned}
$$

(ii) This case has also two possibilities :

1. ace is drawn from the first deck and then ace is drawn from the second deck
2. ace is not drawn from the first deck and then ace is drawn from the second deck

$$
\begin{aligned}
P(\text{card drawn from the second deck is an ace}) &= (\frac{4}{52} * \frac{5}{55}) + (\frac{48}{52} * \frac{4}{55}) \\
&= \frac{1}{143} + \frac{48}{715} \\
&= \frac{53}{715}
\end{aligned}
$$

(iii)

$$
\begin{aligned}
P(\text{ace was transferred from the first deck}|\text{ ace drawn from second deck}) &= \frac{\frac{1}{143}}{\frac{53}{715}} \\
&= \frac{5}{53}
\end{aligned}
$$

Ans: 2.
(a)

$$
\begin{aligned}
P(\text{System is infected with virus}) &= (0.3 * 0.4) + (0.5 * 0.76) + (0.2 * 0.55) \\
&= 0.12 + 0.38 + 0.11 \\
&= 0.61
\end{aligned}
$$

$$
\begin{aligned}
P(\text{Windows system is affected}|\text{System is infected}) &= \frac{0.5 * 0.76}{0.61} \\
&= 0.623
\end{aligned}
$$

(b)

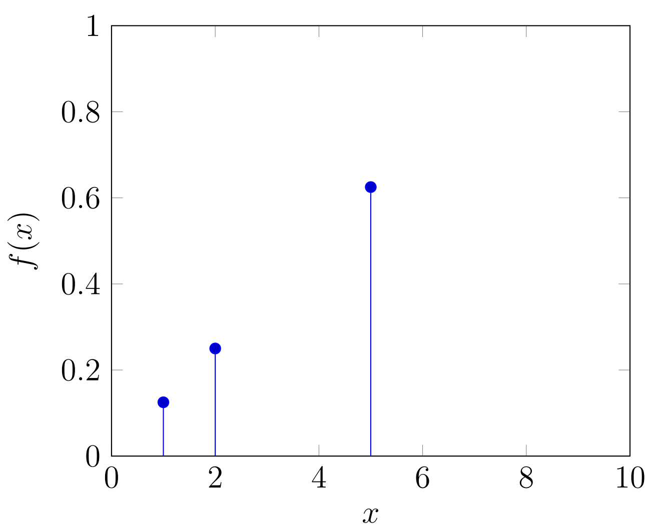
$$\begin{aligned} P(\text{Card has a green side}) &= \frac{1}{3} + (\frac{1}{3} * \frac{1}{2}) \\ &= \frac{1}{3} + \frac{1}{6} \\ &= \frac{1}{2} \end{aligned}$$

$$\begin{aligned} P(\text{Both the sides are green}|\text{ card has green side}) &= \frac{\frac{1}{3}}{\frac{1}{2}} \\ &= \frac{2}{3} \end{aligned}$$

Ans: 3.

(a) (i)

$$f(x) = \left\{ \begin{array}{ll} \frac{1}{8}, & x = \{1, 2, 5\} \\ 0, & \text{otherwise} \end{array} \right\}$$



(ii)

$$\begin{aligned} E[X] &= 1/8 + 2*(2/8) + 5*(5/8) \\ &= 30/8 \\ Var(X) &= E[X^2] - (E[X])^2 \\ &= 1*(1/8) + 4*(2/8) + 25*(5/8) - (30/8)^2 \\ &= 172/64 \end{aligned}$$

(iii)

$$\begin{aligned} E[2X+3] &= 2*E[X] + 3 \\ &= 2*(30/8) + 3 \\ &= 42/4 \end{aligned}$$

(b) For a distribution to be normalized :

$$\sum_{i=i}^{N} f(x_i) = 1$$

Here x has two values = -1 and 1.
Putting x = -1 and 1 in the above equation, we get :

$$\begin{aligned} \sum_{i=i}^{N} f(x_i) &= \frac{1-p}{2} + \frac{1+p}{2} \\ &= 1 \end{aligned}$$

Mean

$$\begin{aligned} E[X] = \sum_{i=i}^{N} xf(x_i) &= (-1)*(\frac{1-p}{2}) + 1*(\frac{1+p}{2}) \\ &= p \end{aligned}$$

Variance

$$\begin{aligned} Var(X) &= E[X^2] - (E[X])^2 \\ E[X^2] &= \sum_{i=i}^{N} x^2 f(x_i) \\ &= (-1)^2 * (\frac{1-p}{2}) + 1^2 * (\frac{1+p}{2}) \\ &= 1 \\ \implies Var(X) &= 1 - p^2 \end{aligned}$$

Ans: 4.

(a)

$$\begin{aligned} P(A|B) &= P(A \cap B)/P(B) \\ &= P(A)P(B)/P(B) \\ &= P(A) \end{aligned}$$

because A and B are independent events.

(b)

$$\begin{aligned} P(A, B|Z) &= P(A|Z) * P(B|Z) \\ \frac{P(A, B, Z)}{P(Z)} &= \frac{P(A, Z)}{P(Z)} * \frac{P(B, Z)}{P(Z)} \\ \frac{P(A, B, Z)}{P(B, Z)} &= \frac{P(A, Z)}{P(Z)} \\ P(A|B, Z) &= P(A|Z) \end{aligned}$$

(c) $P(A_1) = P(A_2) = P(A_3) = 1/2$
$P(A_1 \cap A_2) = 1/4 = P(A_1) * P(A_2)$
$P(A_1 \cap A_3) = 1/4 = P(A_1) * P(A_3)$
$P(A_2 \cap A_3) = 1/4 = P(A_2) * P(A_3)$
$P(A_1 \cap A_2 \cap A_3) = 1/4$
but $P(A_1) * P(A_2) * P(A_3) = 1/8$
$\implies P(A_1 \cap A_2 \cap A_3) \neq P(A_1) * P(A_2) * P(A_3)$

Ans: 5.

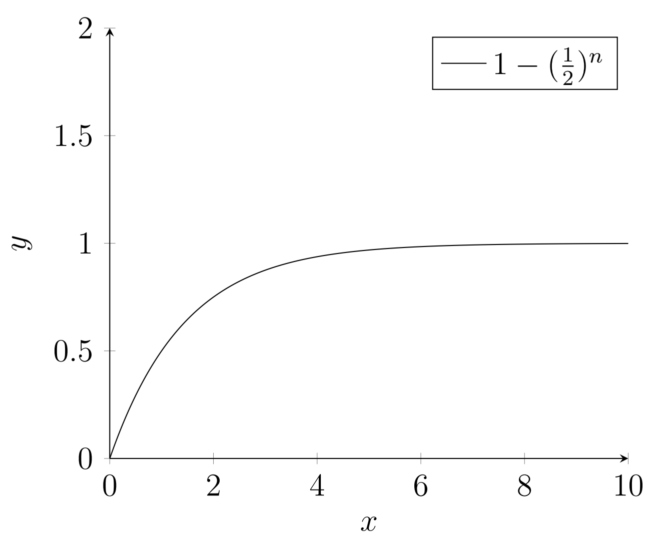
(a) (i) $Y_n = max\{X_1, X_2, ...X_n\}$
$Y_n = 0$ occurs only when all X_i are zero individually.

$$\begin{aligned} P(Y_n = 0) &= \prod_{i=i}^{n}(1 - p_i) \\ P(Y_n = 1) &= 1 - P(Y_n = 0) \\ &= 1 - \prod_{i=i}^{n}(1 - p_i) \end{aligned}$$

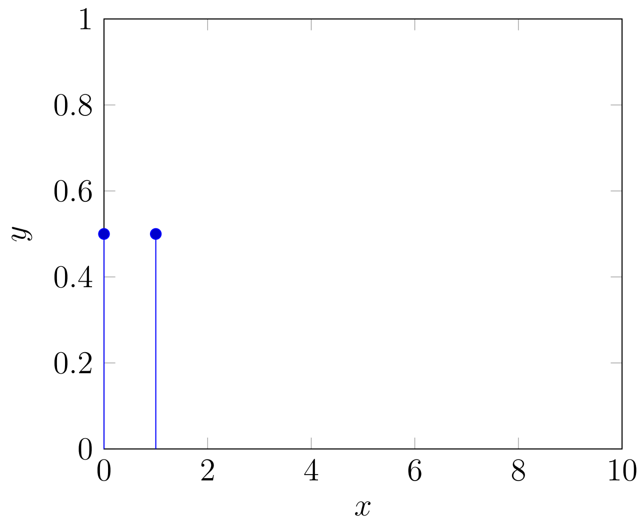
As Y_n can only take two values 0 and 1 with the probability of $Y_n = 1$ as shown above, so this implies that Y_n follows Bernoulli distribution.

$\implies E[Y_n] = 1 - \prod_{i=i}^{n}(1 - p_i)$
Here $p_i = \frac{1}{2}$, so this implies that :
$E[Y_n] = 1 - (\frac{1}{2})^n$

(ii) Plot of $E[Y_n]$:



(iii) The distribution for single Bernoulli random variable is shown here



We can see from the plot in the (ii) part that as the value of n increases, P(Y_n=1) approaches to 1

(b)

$$\begin{aligned}P(\text{Total is div by 4}) &= 9/36 \\ P(\text{Total is not div by 4}) &= 1 - P(\text{Total is div by 4}) \\ &= 27/36\end{aligned}$$

For expected winnings to be \$0, following condition should be met :

$$(\frac{9}{36}) * 12 = (\frac{27}{36}) * x \implies x = \$4$$

Ans: 6.

(a) The possible values for d are : 1,2,3,4,5,6.

(i) d=1

$$\begin{aligned}E[H|1] &= 0 * \frac{1}{2} + 1 * \frac{1}{2} \\ &= \frac{1}{2}\end{aligned} \qquad \begin{aligned}Var(H|1) &= (0^2 * \frac{1}{2} + 1^2 * \frac{1}{2}) - (\frac{1}{2})^2 \\ &= \frac{1}{4}\end{aligned}$$

(ii) d=2

$$E[H|2] = 0 * \frac{1}{4} + 1 * \frac{1}{2} + 2 * \frac{1}{4} \qquad Var(H|2) = (0^2 * \frac{1}{4} + 1^2 * \frac{1}{2} + 2^2 * \frac{1}{4}) - 1^2$$
$$= 1 \qquad = \frac{1}{2}$$

(iii) d=3

$$E[H|3] = 0 * \frac{1}{8} + 1 * \frac{3}{8} + 2 * \frac{3}{8} + 3 * \frac{1}{8}$$
$$= \frac{3}{2}$$

$$Var(H|3) = (0^2 * \frac{1}{8} + 1^2 * \frac{3}{8} + 2^2 * \frac{3}{8} + 3^2 * \frac{1}{8}) - (\frac{3}{2})^2$$
$$= \frac{3}{4}$$

(iv) d=4

$$E[H|4] = 0 * \frac{1}{16} + 1 * \frac{4}{16} + 2 * \frac{6}{16} + 3 * \frac{4}{16} + 4 * \frac{1}{16}$$
$$= 2$$

$$Var(H|4) = (0^2 * \frac{1}{16} + 1^2 * \frac{4}{16} + 2^2 * \frac{6}{16} + 3^2 * \frac{4}{16} + 4^2 * \frac{1}{16}) - (2)^2$$
$$= 1$$

(v) d=5

$$E[H|5] = 0 * \frac{1}{32} + 1 * \frac{5}{32} + 2 * \frac{10}{32} + 3 * \frac{10}{32} + 4 * \frac{5}{32} + 5 * \frac{1}{32}$$
$$= \frac{5}{2}$$

$$Var(H|5) = (0^2 * \frac{1}{32} + 1^2 * \frac{5}{32} + 2^2 * \frac{10}{32} + 3^2 * \frac{10}{32} + 4^2 * \frac{5}{32} + 5^2 * \frac{1}{32}) - (\frac{5}{2})^2$$
$$= \frac{5}{4}$$

(vi) d=6

$$\begin{aligned}E[H|6] &= 0*\frac{1}{64}+1*\frac{6}{64}+2*\frac{15}{64}+3*\frac{20}{64}+4*\frac{15}{64}+5*\frac{6}{64}+6*\frac{1}{64}\\&=3\end{aligned}$$

$$\begin{aligned}Var(H|4) &= (0^2*\frac{1}{64}+1^2*\frac{6}{64}+2^2*\frac{15}{64}+3^2*\frac{20}{64}+4^2*\frac{15}{64}+5^2*\frac{6}{64}+6^2*\frac{1}{64})-(3)^2\\&=\frac{3}{2}\end{aligned}$$

$\blacksquare$

(b)

$$\begin{aligned}E[H] &= \frac{E[H|1]+E[H|2]+E[H|3]+E[H|4]+E[H|5]+E[H|6]}{6}\\&=\frac{21}{12}\end{aligned}$$

$$\begin{aligned}Var(H) &= \frac{1}{6}(\frac{1}{2}+\frac{3}{2}+3+5+\frac{15}{2}+\frac{21}{2})-(\frac{21}{12})^2\\&=\frac{28}{6}-\frac{441}{144}\\&=\frac{231}{144}\end{aligned}$$

Ans: 7.

(a) Correlation coefficient formula:

$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$

Cov(X,Y) = E(XY)-E(X)E(Y)

Using rule of Iterated Expectations, if E[X|Y] is constant, then it must necessarily be equal to its mean E[E[X|Y]]=E[X].

$$\begin{aligned} E[X] &= E[E[X|Y]] \\ &= E[c] \\ &= c \end{aligned}$$

Also, using rule of Iterated Expectations

$$\begin{aligned} E(XY) &= E[E[XY|Y]] \\ &= E[YE(X|Y)] \\ &= E[Yc] \\ &= cE[Y] \\ &= E(X)E[Y] \end{aligned}$$

As E(XY)=E(X)E(Y), this means that X and Y are uncorrelated.

(b)

$$\begin{aligned} Cov(X, Y+Z) &= E[X(Y+Z)] - E[X]E[Y+Z] \\ &= E[XY + XZ] - E[X](E[Y] + E[Z]) \\ &= E[XY] - E[X]E[Y] + E[XZ] - E[X]E[Z] \\ &= Cov(X,Y) + Cov(X,Z) \end{aligned}$$

(c)

$$\begin{aligned} Cov(X_1 + X_2, Y_1 + Y_2) &= Cov(X_1 + X_2, Y_1) + Cov(X_1 + X_2, Y_2) \\ &= Cov(X_1, Y_1) + Cov(X_2, Y_1) + Cov(X_1, Y_2) + Cov(X_2, Y_2) \\ &= 5 + 1 + 2 + 8 \\ &= 16 \end{aligned}$$

$\blacksquare$

Ans: 8.
(a) Given $||x||_2 = ||y||_2 = 1$

$$\begin{aligned} cos\theta &= \frac{x \cdot y}{|x||y|} \\ &= x \cdot y \\ &= x^T y \end{aligned}$$

$$\begin{aligned} ||x-y||_2^2 &= (x-y)^T(x-y) \\ &= x^T x - 2x^T y + y^T y \\ &= 2 - 2x^T y \\ &= 2 - 2cos\theta \end{aligned}$$

(b)

$$\rho(X,Y) = \frac{\frac{1}{n}\sum_{i=i} x_i y_i - \mu_x \mu_y}{\sigma_x \sigma_y}$$

As the data points have been standardized, mean becomes zero and standard deviation becomes 1, so the formula now becomes like this :

$$\rho(X,Y) = \frac{1}{n}\sum_{i=i} x_i y_i$$

Euclidean distance can be calculated like this :

$$\begin{aligned} d(X,Y) &= \sqrt{\sum_{i=i}^{N}(x_i - y_i)^2} \\ &= \sqrt{\sum_{i=i}^{N} x_i^2 + \sum_{i=i}^{N} y_i^2 - 2\sum_{i=i}^{N} x_i y_i} \end{aligned}$$

As the data points have been standardized, $\sum_{i=i}^{N} x_i^2$ and $\sum_{i=i}^{N} y_i^2$ becomes equal to n each. So, the relation becomes :

$$\begin{aligned} d(X,Y) &= \sqrt{2n - 2n * \rho(X,Y)} \\ d^2(X,Y) &= 2n - 2n * \rho(X,Y) \\ \implies \rho(X,Y) &= 1 - \frac{d^2(X,Y)}{2n} \end{aligned}$$

Ans: 9.

(a) For a matrix to have an inverse, the determinant for that matrix has to be non zero.

$$A = \begin{bmatrix} 9 & 1 & 9 & 9 & 9 \\ 9 & 0 & 9 & 9 & 2 \\ 4 & 0 & 0 & 5 & 0 \\ 9 & 0 & 3 & 9 & 0 \\ 6 & 0 & 0 & 7 & 0 \end{bmatrix}$$

Finding the determinant along second column :

$$\det(A) = -1 * \det(\begin{bmatrix} 9 & 9 & 9 & 2 \\ 4 & 0 & 5 & 0 \\ 9 & 3 & 9 & 0 \\ 6 & 0 & 7 & 0 \end{bmatrix})$$

Finding determinant of 4*4 matrix along fourth column :

$$\det(A) = (-1) * (-2) * \det(\begin{bmatrix} 4 & 0 & 5 \\ 9 & 3 & 9 \\ 6 & 0 & 7 \end{bmatrix})$$

Finding determinant of 3*3 matrix along second column :

$$\det(A) = 2 * 3 * \det(\begin{bmatrix} 4 & 5 \\ 6 & 7 \end{bmatrix})$$

det(A) = 6 * (28-30) = -12 $\neq$ 0
$\implies$ A is invertible matrix.

(b) $A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}$

A - λ I = $\begin{bmatrix} 0-\lambda & 1 \\ -2 & -3-\lambda \end{bmatrix}$

det(A - λ I) = 0 for finding the eigenvalues
$\implies -\lambda(-3-\lambda) + 2 = 0$
$\implies \lambda^2 + 3\lambda + 2 = 0$
$\implies \lambda = -2 or \lambda =$ -1
These are the eigenvalues for A.

For eigenvectors :

(a) For λ = -2
$\begin{bmatrix} 2 & 1 \\ -2 & -1 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$
$\implies 2x_1 + x_2 = 0$
$\implies$ Eigenvector is of the form $\begin{bmatrix} k \\ -2k \end{bmatrix}$

(b) For λ = -1
$\begin{bmatrix} 1 & 1 \\ -2 & -2 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$
$\implies x_1 + x_2 = 0$
$\implies$ Eigenvector is of the form $\begin{bmatrix} k \\ -k \end{bmatrix}$

Ans: 10.
(a) Likelihood term is :

$$P(x_1, ..., x_N|\mu) = \prod_{i=i}^{N} P(x_i|\mu)$$
$$= \prod_{i=i}^{N} \frac{1}{\sqrt{2\pi\delta^2}} e^{-\frac{(x_i-\mu)^2}{2\delta^2}}$$

As log is a monotonically increasing function, we can calculate the maximum likelihood estimation of μ by maximising the log-likelihood

$\log(P(x_1, ..., x_N|\mu)) = \sum_{i=i}^{N} \log(\frac{1}{\sqrt{2\pi\delta^2}}) - \frac{(x_i-\mu)^2}{2\delta^2}$

Taking derivative w.r.t μ and setting the value to zero:

$$\sum_{i=i}^{N} \frac{(x_i - \mu)}{\delta^2} = 0$$

$$\sum_{i=i}^{N} (x_i - \mu) = 0$$

$$\sum_{i=i}^{N} x_i = N\mu$$

$$\implies \mu = \frac{(\sum_{i=i}^{N} x_i)}{N}$$

(b) By using Bayes rule:

$$P(\mu|x_1, ..., x_N) = \frac{P(x_1, ..., x_N|\mu)P(\mu)}{P(x_1, ..., x_N)}$$

We are given that :

P$(\mu) = (\frac{1}{\sqrt{2\pi\lambda^2}})e^{-\frac{(\mu-\eta)^2}{2\lambda^2}}$

From above part, we know that :

$P(x_1, ..., x_N|\mu) = \prod_{i=i}^{N} \frac{1}{\sqrt{2\pi\delta^2}} e^{-\frac{(x_i-\mu)^2}{2\delta^2}}$

We want to find that value of μ that maximizes the function P$(\mu|x_1, ..., x_N)$.

After taking log of this function and then derivative w.r.t μ, we get:

$\sum_{i=i}^{N} \frac{(x_i-\mu)}{\delta^2} - \frac{(\mu-\eta)}{\lambda^2} = 0$

$\frac{(\mu-\eta)}{\lambda^2} = \sum_{i=i}^{N} \frac{(x_i-\mu)}{\delta^2}$

$\frac{(\mu-\eta)}{\lambda^2} = \sum_{i=i}^{N} \frac{x_i}{\delta^2} - \frac{N\mu}{\delta^2}$

$\frac{\mu}{\lambda^2} + \frac{N\mu}{\delta^2} = \sum_{i=i}^{N} \frac{x_i}{\delta^2} + \frac{\eta}{\lambda^2}$

$$\frac{(\delta^2+N\lambda^2)\mu}{\delta^2\lambda^2} = \frac{\delta^2\eta+\lambda^2\sum_{i=i}^{N} x_i}{\delta^2\lambda^2}$$

$$\mu = \frac{\delta^2\eta+\lambda^2\sum_{i=i}^{N} x_i}{(\delta^2+N\lambda^2)}$$