CS57300: Data Mining Homework

Praneeth Sai Valiveti (pvalivet@purdue.edu)

1 Counting

1.1 Two cards are drawn from a deck of 52 cards without replacement:

1.1.1 What is the probability that the second card is a heart, given that the first card is a heart?

Ans. Considering the fact that we have already selected one **heart**, we are left with a total of 51 cards among which 12 are **hearts**.
Hence, the probability of selecting a **heart** as the second card becomes:

$$\text{P(second card is } \textbf{heart}) = \frac{12}{51}$$

1.1.2 What is the probability that both cards are hearts, given that at least one card is a heart?

Ans The probability that both cards are **hearts**

$$\text{P(Both are } \textbf{hearts}) = \frac{\binom{13}{2}}{\binom{52}{2}}$$

The probability of all scenarios where we have at least one **heart**

$$\text{P(At least one } \textbf{heart}) = 1 - \text{P(Neither of the two are } \textbf{hearts})$$

$$\text{P(At least one } \textbf{heart}) = 1 - \frac{\binom{39}{2}}{\binom{52}{2}}$$

Hence

$$\text{P(Both cards are } \textbf{heart}|\text{At least one is a } \textbf{heart}) = \frac{\frac{\binom{13}{2}}{\binom{52}{2}}}{1-\frac{\binom{39}{2}}{\binom{52}{2}}}$$

Hence

$$\text{P(Both cards are } \textbf{heart}|\text{At least one is a } \textbf{heart}) = \frac{2}{15}$$

1.2 One card is selected from a deck of 52 cards and placed in a second deck containing 52 cards. A card is then selected from the second deck.

1.2.1 What is the probability that a card drawn from the second deck is an ace?

Ans An **ace** can be drawn from the second deck in two scenarios.
1. The card drawn from the first deck was an **ace** and the second card is an **ace**.
2. The card drawn from the first deck was not an **ace** and the second card is an **ace**.

Hence

$$\text{P(Card drawn from second deck is an } \textbf{ace}) = \text{P(Case 1)} + \text{P(Case 2)}$$

$$\text{P(Card drawn from second deck is an } \textbf{ace}) = \frac{4\times5}{52\times53} + \frac{48\times4}{52\times53}$$

$$= \frac{1}{13}$$

1.2.2 If the first card is placed into a deck of 54 cards containing two jokers, then what is the probability that a card drawn from the second deck is an ace?

Ans Reformulating the above question according to new numbers, we get

$$\begin{aligned}\text{P(Card drawn from second deck is an } \mathbf{ace}) &= \tfrac{4\times5}{52\times55} + \tfrac{48\times4}{52\times55} \\ &= \tfrac{212}{2860}\end{aligned}$$

1.2.3 Given that an ace was drawn from the second deck in (ii), what is the conditional probability that an ace was transferred from the first deck?

Ans
We need to find P(First selected is **ace**| Second is **ace**)

$$\begin{aligned}P(card_1 = \mathbf{ace}|card_2 = \mathbf{ace}) &= \tfrac{P(card_2=\mathbf{ace}|card_1=\mathbf{ace})\times P(card_1=\mathbf{ace})}{P(card_2=\mathbf{ace})} \\ &= \frac{\frac{5}{55}\times\frac{4}{52}}{\frac{53}{715}} \\ &= \frac{5}{53}\end{aligned}$$

2 Probability and conditional probability

2.1 Suppose that 30 percent of computer owners use an Apple machine, 50 percent use a Windows machine, and 20 percent use Linux. Suppose that 65 percent of Apple users have succumbed to a computer virus, 82 percent of Windows users get the virus, and 50 percent of Linux users get the virus. We select a person at random and learn that their system was infected with the virus. What is the probability that the person is a Windows user?

Ans

$$P(Apple) = 0.3$$
$$P(Windows) = 0.5$$
$$P(Linux) = 0.2$$
$$P(Virus|Apple) = 0.65$$
$$P(Virus|Windows) = 0.82$$
$$P(Virus|Linux) = 0.50$$

$$\text{P(Windows|Virus)} = \frac{P(Windows)*P(Virus|Windows)}{P(Windows)*P(Virus|Windows)+P(Apple)*P(Virus|Apple)+P(Linux)*P(Virus|Linux)}$$
$$\text{P(Windows|Virus)} = \frac{0.5\times0.82}{(0.5\times0.82)+(0.3\times0.65)+(0.2\times0.50)}$$

$$= \frac{0.41}{0.705}$$
$$= 0.581$$

2.2 There are three cards. The first is green on both sides, the second is red on both sides, and the third is green on one side and red on the other. Consider the scenario where a card is chosen at random and one side is shown (also chosen at random). If the side shown is green, what is the probability that the other side is also green?

Ans We know that one side of the selected card is **green**. There are three such cases, one with **red** on one side, and two where the other side are **green**. The number of favorable cases would be where the double **green** card is selected, hence two cases.

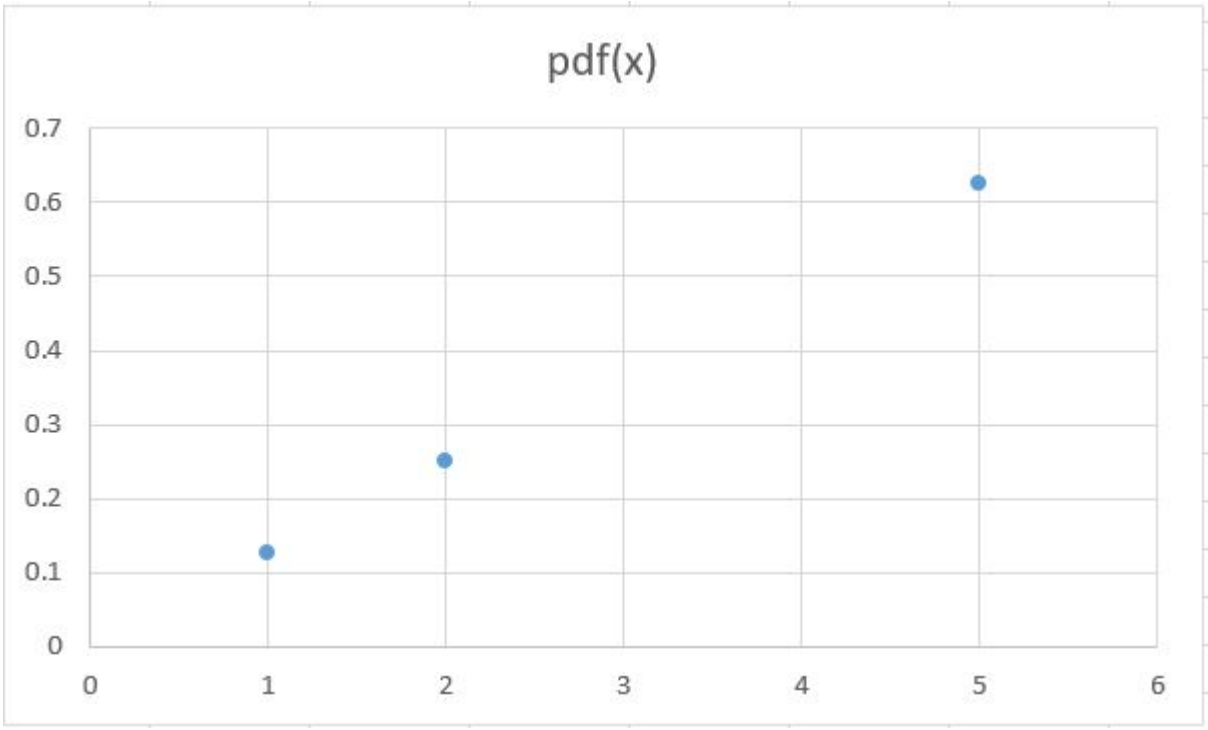
$$\text{P(Other side is } \textbf{green}) = \frac{2}{3}$$

3 Probability distributions

3.1 Let X be a random variable with discrete pdf $f(x) = \frac{x}{8}$ if x = 1, 2, or 5 and zero otherwise.

3.1.1 Sketch the graph of the discrete pdf f(x).

Ans



3.1.2 Find E[X] and Var(X).

Ans

$$\text{E[X]} = \Sigma[x \times f(x)]$$
$$\text{E[X]} = 1 \times \tfrac{1}{8} + 2 \times \tfrac{2}{8} + 5 \times \tfrac{5}{8}$$
$$\text{Hence, E[X]} = 3.75$$

$$\text{Var[X]} = E[X^2] - E[X]^2$$
$$= \Sigma[x^2 \times f(x)] - E[X]^2$$
$$= [1^2 \times \tfrac{1}{8} + 2^2 \times \tfrac{2}{8} + 5^2 \times \tfrac{5}{8}] - 3.75^2$$
$$= 2.6875$$

3.1.3 Find E[2X + 3]

Ans

$$\text{E[2X+3]} = 2 \times E[X] + 3$$
$$\text{Hence, E[2X+3]} = 10.5$$

3.2 The form of the Bernoulli(p) distribution is not symmetric between the two values of X. In some situations, it will be more convenient to use an equivalent formulation for which x lies between -1 and 1, in which case the distribution can be written as:
$P(x|p) = (\frac{1-p}{2})^{\frac{1-x}{2}}(\frac{1+p}{2})^{\frac{1+x}{2}}$
Show that this distribution is normalized (i.e., sums to 1) and evaluate its mean and variance.

Ans For normality, $\Sigma[P(x|p)]$ must be equal to 1.

$$\begin{gathered}\Sigma[P(x|p)] = P(-1|p) + P(1|p) \\ = \frac{1-p}{2} + \frac{1+p}{2} \\ = 1\end{gathered}$$

To find the **Mean**

$$\begin{gathered}E[X] = \Sigma[x \times P(x)] \\ = (-1 \times \frac{1-p}{2}) + (1 \times \frac{1+p}{2}) \\ = p\end{gathered}$$

To find the **Variance**

$$\begin{gathered}Var[X] = E[X^2] - E[X] \\ = [(-1^2 \times \frac{1-p}{2}) + (1^2 \times \frac{1+p}{2})] - p^2 \\ = 1 - p^2\end{gathered}$$

4 Independence

4.1 Prove the following: If A and B are independent events, then P(A|B) = P(A). Also, for any pair of events A and B, P(AB) = P(A|B)P(B) = P(B|A)P(A)

Ans Definition of Independence : "In probability theory, two events are independent if the occurrence of one does not affect the probability of occurrence of other."
In other words, two events A and B are said to be independent, if information about the occurrence of B gives us absolutely no information about the occurrence of A.
Hence, since the occurrence of B changes nothing about the probability of occurrence of A, we can say that

$$P(A|B) = P(A)$$

Method 2:
For independent events A and B, $P(AB) = P(A) \times P(B)$

$$\begin{aligned} P(A|B) &= \frac{P(AB)}{P(B)} \\ &= \frac{P(A)\times P(B)}{P(B)} \\ &= P(A) \end{aligned}$$

Part 2:
If we know that B has occurred, we have effectively reduced the sample space to the area wherein B has occurred for certain. Hence, the effective sample space here has become the probability that B has occurred.
Now, we want to find inside this new sample space, the probability that A has occurred. This is $P(A \cap B)$. Hence

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

The same principle, when used on P(B|A), gives us

$$P(B|A) = P(BA) \times P(A)\ P(A \cap B) = P(B \cap A)$$ Hence, $$P(B|A) = P(AB) \times P(A)$$

Hence,

$$P(A \cap B) = P(A|B) \times P(B) = P(B|A) \times P(A)$$

4.2 A box contains the following four slips of paper, each having exactly the same dimensions: (1) win prize 1, (2) win prize 2, (3) win prize 3, (4) win prizes 1, 2, and 3. One slip will be randomly selected. Let A1 = win prize 1, A2 = win prize 2, and A3 = win prize 3. Show that A1, A2, and A3 are pairwise independent, but that the three events are not mutually independent - i.e. $P(A1 \cap A2 \cap A3) \neq P((A1)P(A2)P(A3))$.

Ans We know that

$$P(A_1) = \frac{2}{4} = \frac{1}{2}$$

Similarly,

$$\begin{aligned} P(A_2) &= \frac{1}{2} \\ P(A_3) &= \frac{1}{2} \end{aligned}$$

And by looking at the sets,

$$\begin{aligned} P(A_1 \cap A_2) &= \frac{1}{4} \\ P(A_1 \cap A_3) &= \frac{1}{4} \\ P(A_2 \cap A_3) &= \frac{1}{4} \end{aligned}$$

We see that

$$P(A_1 \cap A_2) = P(A_1) \times P(A_2)$$
$$P(A_1 \cap A_3) = P(A_1) \times P(A_3)$$
$$P(A_2 \cap A_3) = P(A_2) \times P(A_3)$$

Hence, we can confirm that they are independent pairwise.
But,

$$P(A_1 \cap A_2 \cap A_3) = \frac{1}{4}$$
$$P(A_1) \times P(A_2) \times P(A_3) = \frac{1}{8}$$
$$P(A_1 \cap A_2 \cap A_3) \neq P(A_1) \times P(A_2) \times P(A_3)$$

Hence, we can conclude that they are not mutually independent.

5 Expectation

5.1 Let $X_1, ..., X_n$ be Bernoulli(p=0.5). Let $Y_n = max(X_1, ..., X_n)$

5.1.1 Find $E[Y_n]$

Ans The distribution of Y_n would be as follows:

$$P(Y_n) = \frac{1}{2^n} \text{ when } Y_n = 0$$
$$P(Y_n) = 1 - \frac{1}{2^n} \text{ when } Y_n = 1$$

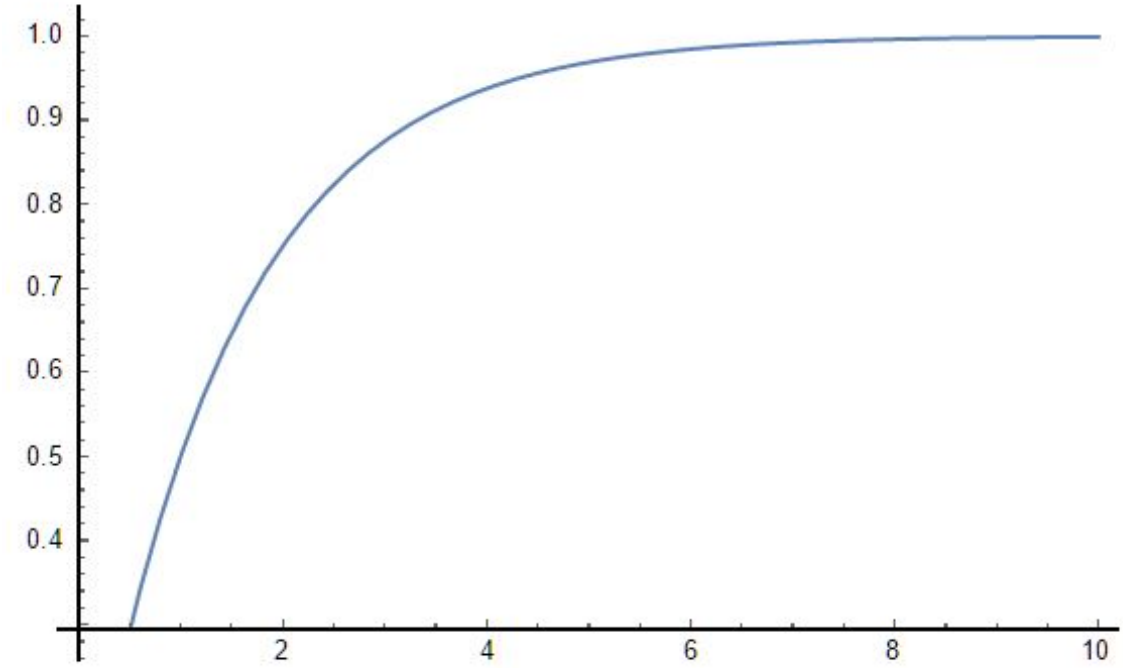
Hence, the distribution of $P(Y_n)$ can be written as

$$P_{Y_n}(x) = (\frac{1}{2^n})^{1-x} \times (1 - \frac{1}{2^n})^x$$

$E[Y_n] = \Sigma[x \times P_{Y_n}(x)] = 1 - \frac{1}{2^n}$

5.1.2 Plot $E[Y_n]$ as a function of n

Ans



Above is the graph of Y_n vs n. As expected the value drastically increases and becomes asymptotic to 1.

5.1.3 How is the distribution of the max (Y_n) different from that of a single Bernoulli (X_i)?

Ans We have found that the distribution of Y_n is

$$P_{Y_n}(x) = (\frac{1}{2^n})^{1-x} \times (1 - \frac{1}{2^n})^x$$

We know that the distribution of a Bernoulli(p) Distribution is

$$P_X(x) = p^x \times (1-p)^{1-x}$$

From this we can see that Y_n is a Bernoulli(p=$1 - \frac{1}{2^n}$) Random Variable.
The maximum value of Y_n is 1 and the probability of this is $1 - \frac{1}{2^n}$ while that for each X_i is 1 with a probability of p.

5.2 You and your friend are playing the following game: two dice are rolled; if the total showing is divisible by 3, you pay your friend $6. If you want to make the game fair, how much should she pay you when the total is not divisible by 3? A fair game is one in which your expected winnings are $0.

Ans The probability that the sum is divisible by 3 is $\frac{1}{3}$ and not divisible by 3 is $\frac{2}{3}$. If **x** is the amount she should pay you,

$$\frac{1}{3} \times (-6) + \frac{2}{3} \times x = 0$$
$$x = 3$$

Hence she should pay $3 when the total is not divisible by 3.

6 Conditional Expectation

6.1 Consider the setting where you first roll a fair 6-sided die, and then you flip a fair coin the number of times shown by the die. Let D refer to the outcome of the die roll (i.e., number of coin flips) and let H refer to the number of heads observed after D coin flips.

6.1.1 Determine E[H|d] and Var(H|d).

Ans We know that $f(D,H) = \frac{1}{6} \times \binom{d}{h} \times \frac{1}{2^d}$ (f(X,Y) is the probability distribution) $E[X|Y = y] = \Sigma[x \times f(X|Y = y)]$

$$\begin{aligned} f(H|D = d) &= \frac{f(H,D)}{f(D)} \\ &= \frac{\frac{1}{6} \times \binom{d}{h} \times \frac{1}{2^d}}{\frac{1}{6}} \\ &= \binom{d}{h} \times \frac{1}{2^d} \end{aligned}$$

Which is a Binomial Distribution $(d, \frac{1}{2})$ where **h** is the number of heads (successes), giving rise to:

$$E[H|D = d] = \Sigma[h \times \binom{d}{h} \times \frac{1}{2^d}]$$

For a binomial distribution(n,p) , E[X]=$n \times p$
Hence,

$$E[H|D = d] = \frac{d}{2}$$

Similarly, the variance of a Bin(n,p) is $Var[X] = n \times p \times (1 - p)$

$$Var[H|D = d] = \frac{d}{4}$$

6.1.2 Determine E[H] and Var(H)

$$E[H] = E[H|D = d] = E[\frac{d}{2}] = \frac{1}{2} \times \Sigma(d \times \frac{1}{6}) = \frac{7}{4}$$

Var(H)=E[Var(H|D)] + Var(E[H|D]) (**Law of Total Variance**)

$$= E[\frac{d}{4}] + var(\frac{d}{2}) = \frac{77}{48}$$

7 Covariance and Correlation

7.1 Show that if E[X|Y = y] = c for some constant c, then X and Y are uncorrelated

Ans

$$E[X|Y=y] = c$$
$$E[X|Y=y] = E[E[XY/Y=y]$$
$$= E[X] \times E[Y]$$

Generally, $E[XY] = E[X] \times E[Y] + cov(X,Y)$
But, here we see that

$E[XY] = c \times E[Y]$ Which means that cov(X,Y) = 0

Hence, they must be uncorrelated.

7.2 Show Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z).

Ans

$$Cov(X,Y) = E[XY] - E[X] \times E[Y]$$
$$Cov(X,Y+Z) = E[X \times (Y+Z)] - E[X] \times E[Y+Z]$$
$$= E[XY + XZ] - E[X] \times (E[Y] + E[Z])$$
$$= (E[XY] - E[X] \times E[Y]) + (E[XZ] - E[X] \times E[Z])$$
$$= cov(X,Y) + cov(X,Z)$$

7.3 Let X1 and X2 be quantitative and verbal scores on one aptitude exam and Y1 and Y2 be corresponding scores on another exam. If Cov(X1, Y1) = 5, Cov(X1, Y2) = 1, Cov(X2, Y1) = 2, and Cov(X2, Y2) = 8, what is the covariance between the two total scores X1 +X2 and Y1 +Y2?

Ans Using the above established rule,

$$cov(X_1 + X_2, Y_1 + Y_2) = cov(X_1, Y_1) + cov(X_1, Y_2) + cov(X_2, Y_1) + cov(X_2, Y_2) = 16$$

8 Distance and Correlation Measures

8.1 Show how Euclidean distance can be expressed as a function of cosine similarity when each data vector has an L_2 length of 1.

Ans Consider two vectors x and y, both having L_2 length of 1. In such a scenario,
Variance = n $\times$ the sum of the square of their values.

$$\begin{aligned} d(x,y) &= \sqrt{\Sigma[x_i + y_i]^2} \\ &= \sqrt{\Sigma x_i^2 + y_i^2 + 2 \times x_i \times y_i} \\ &= \sqrt{2(1 - cos(x,y))} \end{aligned}$$

8.2 Show how Euclidean distance can be expressed as a function of correlation when each data point has been standardized by subtracting its mean and dividing by its standard deviation.

Ans Let x and y be standardized vectors. (Mean= 0, Variance =1)
Correlation between x and y = dot product(x,y) divided by n.

$$\begin{aligned} d(x,y) &= \sqrt{\Sigma(x_i + y_i)^2} \\ &= \sqrt{\Sigma(x_i^2 + y_i^2 + 2 \times x_i \times y_i)} \\ &= \sqrt{2n(1 - corr(x,y))} \end{aligned}$$

9 Linear Algebra

9.1 Verify directly that A(AB) $= A^2B$.

Ans

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

$$B = \begin{bmatrix} 2 & -2 \\ 1 & 3 \\ -4 & 4 \end{bmatrix}$$

Calculating A(AB)

$$AB = \begin{bmatrix} -3 & -1 \\ 0 & 0 \\ 2 & -2 \end{bmatrix}$$

$$A(AB) = \begin{bmatrix} -1 & -3 \\ -4 & -4 \\ -7 & -5 \end{bmatrix}$$

Calculating A^2B

$$A^2 = \begin{bmatrix} 2 & -1 & 1 \\ 5 & -2 & 3 \\ 6 & -3 & 4 \end{bmatrix}$$

$$A^2B = \begin{bmatrix} -1 & -3 \\ -4 & -4 \\ -7 & -5 \end{bmatrix}$$

9.2 Specify whether the following matrix has an inverse without trying to compute the inverse:

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

Ans Given matrix

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

Taking determinant along the 2^{nd} column, we get

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

Taking along the 4^{th} column, we get

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

Taking along the 2^{nd} column, we get

$$(-1 \times 2 \times -3) \times \begin{bmatrix} 4 & 5 \\ 6 & 7 \end{bmatrix}$$

The determinant hence becomes

$$(-1 \times 2 \times -3 \times (4 \times 7 - 6 \times 5)) = -12$$

Since the determinant is not 0, we can conclude that the inverse for the given matrix exists.