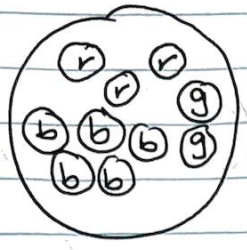
Courtney Hodge

Bayesian ML - Homework 1 - Probability and Priors

Problem 1: Basic Probability



10 balls total 5/10 blue
3/10 red
2/10 green

(a) The probability of drawing a red ball from the bag is (3/10)

~~5/10 blue balls~~
(b) ~~(4/9 blue balls) Probability that~~

5/10 probability that the first ball is blue
4/9 probability that the next ball is also blue
$\frac{5}{10} \times \frac{4}{9} = \frac{20}{90} = \left(\frac{2}{9}\right)$, probability that the next ball is also blue

Problem 2: Independent Events

P(Server being down) = 0.05, server failures are independ.

Probability of 2 independent events $P(A \text{ and } B) = P(A) \times P(B)$

P(server 1 being down AND server 2 being down) =
$P(0.05) \times P(0.05) = 0.0025$

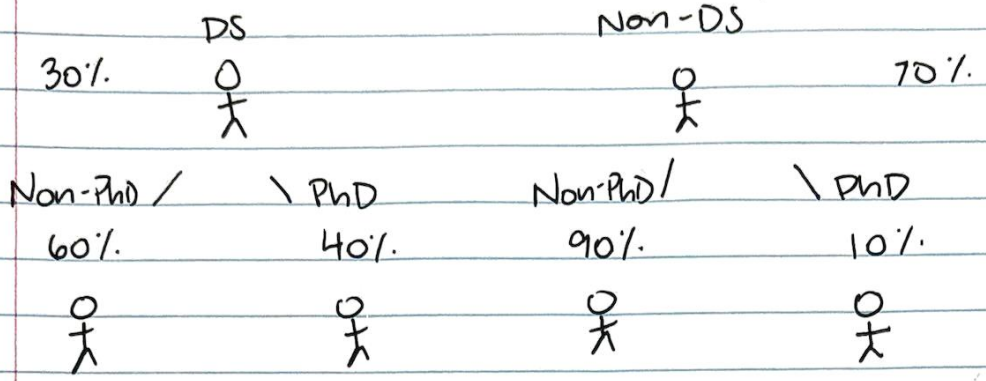
(a) (0.0025)

Probability of at least 1 being down
P(server 1 being down OR server 2 being down) =
$P(0.05) + P(0.05) = 0.1$

(b) (0.1)

Courtney Hodge

Problem 3: Conditional Probability



(a) If an employee is chosen randomly, what is the probability that the employee is a Data Scientist?

$$P(DS) \times P(PhD|DS) = \left(\frac{3}{10}\right) \times \left(\frac{4}{10}\right) = \boxed{\frac{12}{100}}$$

(b) Given that an employee has a PhD, what is the probability that the employee is a Data Scientist?

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \rightarrow P(DS|PhD) = \frac{P(PhD|DS)P(DS)}{P(PhD)}$$

#1 Calculate PhD using the law of total probability:

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

$$P(PhD) = P(PhD|DS)P(DS) + P(PhD|Non\text{-}DS)P(Non\text{-}DS)$$

$$= \left(\frac{4}{10}\right)\left(\frac{3}{10}\right) + \left(\frac{1}{10}\right)\left(\frac{7}{10}\right)$$

$$= \frac{12}{100} + \frac{7}{100} = \frac{19}{100}$$

#2 Solve: $P(DS|PhD) = \dfrac{\left(\frac{4}{10}\right)\left(\frac{3}{10}\right)}{\left(\frac{19}{100}\right)} = \boxed{0.63}$

Courtney Hodge

Problem 4: Law of Total Probability

Law of Total Probability

$$P(A) = \sum_{i=1}^{n} P(A|B_i)\,P(B_i)$$

P(Tested Positive | Has Disease) = 0.95

P(Tested Positive | Doesn't Have Disease) = 0.1

P(Has Disease) = 0.5% = 0.005

P(Doesn't Have Disease) = 0.995

(a) If a person tested positive in the test, what is the probability that the person actually has the disease?

$$P(\text{Has Disease} \mid \text{Tested Positive}) = \frac{P(\text{Has Disease})\,P(\text{Tested Positive} \mid \text{Has Disease})}{P(\text{Tested Positive})}$$

$$P(A \mid B) = P(B \mid A)$$

$$P(A|B) = \frac{0.005\,(0.95)}{P(B)} = \frac{0.005(0.95)}{0.10425} = 0.0456$$

$$P(B) = P(\text{Tested Positive}) = P(\text{Tested Pos} \mid \text{Has Disease})\,P(\text{Has Disease}) + P(\text{Tested Pos} \mid \text{Doesn't Have Disease})\,P(\text{Doesn't Have Disease})$$

$$P(\text{Tested Positive}) = 0.95\,(0.005) + (0.1)\,0.995 = 0.10425$$

(b) What is the total probability of a person testing positive?

$$P(\text{Tested Positive}) = 0.10425$$

Courtney Hodge

Problem 5

Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(\text{Spam}) = .9$

$P(\text{Not-Spam}) = 0.1$ $\quad P(\text{Identified Not-Spam} \mid \text{Spam}) = 0.05$

$P(\text{Identifies Spam} \mid \text{Spam}) = 0.95$

$P(\text{Identifies Not-Spam} \mid \text{Not-Spam}) = 0.85$

$P(\text{Identified Spam} \mid \text{Not-Spam}) = [redacted]\ 0.15$

(a) If an email is picked at random, and the filter classifies it as spam, what is the probability that it is actually spam?

~~$P(\text{Spam}) \times P(\text{Identifies Spam} \mid \text{Spam}) = 0.9 \times 0.95 = 0.855$~~

$P([redacted]\ B \mid A)$ $\quad P(A)$

$P(\text{Spam} \mid \text{Identified Spam}) = \frac{P(\text{Identified Spam} \mid \text{Spam})P(\text{Spam})}{P(\text{Identified Spam})}$

$P(A \mid B) = \frac{0.95(0.9)}{P(B)} = \frac{0.95(0.9)}{0.87} = 0.983$

Law of Total Probability

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

$P(\text{Identified Spam}) = P(\text{Identified Spam} \mid \text{Spam})P(\text{Spam}) + P(\text{Identified Spam} \mid \text{Not-Spam})P(\text{Not-Spam})$

$= 0.95(.9) + ([redacted]\ 0.15)0.1$

$= 0.87$

$P([redacted]\ \text{Identified Spam} \mid \text{Not-Spam}) = 0.15$

Courtney Hodge

Problem 5

(b) If an email is classified as "not spam", what is the probability that it is actually spam?

$P(\text{Spam} \mid \text{Identified Not-Spam})$

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

$$= \frac{P(\text{Spam})\,P(\text{Identified Not-Spam} \mid \text{Spam})}{P(\text{Identified Not-Spam})}$$

$$= \frac{0.9\,(0.05)}{P(\text{Identified Not-Spam})} = \frac{0.9(0.05)}{0.13} = 0.346$$

Law of Total Probability

$$P(A) = \sum_{i=1}^{n} P(A|B_i)\,P(B_i)$$

$$P(\text{Identified}^{\text{NOT}} \wedge \text{Spam}) = P(\text{Identified}^{\text{NOT}} \wedge \text{Spam} \mid \text{Spam})\,P(\text{Spam}) + P(\text{Identified Not Spam} \mid \text{Not-Spam}) * P(\text{Not-Spam})$$

$$= 0.05\,(0.9) + (0.85)\,0.1$$

$$= 0.13$$

Courtney Hodge

Problem 6: Expectation of a Discrete Random Variable

(a) Define the random variable x that models this game.

$\frac{1}{6}$ = 6 , win \$10

$\frac{5}{6}$ = 1, 2, 3, 4 or 5 , lose \$2

~~$P(\frac{1}{6}) + P(\frac{5}{6})$~~

X = 10, if you roll a 6
X = -2, if you roll 1 through 5

$$X = \begin{cases} 10, & \text{if a 6 appears} \\ -2, & \text{if any number 1 to 5 appears} \end{cases}$$

(b) Compute the expected value of X
P(if a 6 appears)X + P(if you roll 1→5)X

To compute the expected value of X, we use the expected value of a discrete random variable

$$E(X) = \sum_i P(x_i) \cdot x_i$$, where x_i are the possible outcomes and $P(x_i)$ are the corresponding probabilities.

Prob of rolling a 6 is $P(X=10) = 1/6$
Prob of rolling anything else $P(X=-2) = 5/6$

$$E(X) = P(X=10)(10) + P(X=-2)(-2)$$

$$\frac{-10}{6} = 0$$

$$(\tfrac{1}{6})(10) + (5/6)(-2) = \frac{10}{6} + \left(\frac{-10}{6}\right) = 0$$

E(X) = 0

Courtney Hodge

Problem 7: Expectation of a Continuous Random Variable

(a) Compute the expected value $E[X]$ of X

$f(x) = 2e^{-2x}$ for $x \geq 0$

Probability Density Function

PDF

$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$

$E(X) = \int_{0}^{\infty} x\left(2e^{-2x}\right)dx$

Integration By Parts

$u = x \quad du = dx$

$v = -e^{-2x} \quad dv = 2e^{-2x}$

$\int 2e^{-2x}\,dx \quad u = -2x$

$du = -2dx$

$-\int e^{u}\,du$ (exponential rule)

$= -e^{u} = -e^{-2x}$

$\int u\,dv = uv - \int v\,du$ (take neg out)

$E(X) = \left[-xe^{-2x}\right]_0^{\infty} + \int_0^{\infty} e^{-2x}\,dx$

Solving, we get... At $x=0$, $-xe^{-2x} = 0$

At $x=\infty$, $-xe^{-2x}$ ~~goes to~~ $= 0$

So $\left[-xe^{-2x}\right]_0^{\infty}$ is $= 0$

Solving $\int_0^{\infty} e^{-2x}\,dx = \left[\frac{-1}{2}e^{-2x}\right]_0^{\infty} = \left(0 - \left(\frac{-1}{2}\right)\right) = \frac{1}{2}$

Therefore, $\boxed{E(X) = 0 + \frac{1}{2} = \frac{1}{2}}$

(b) Compute the variance $Var[X]$ of X

Computing $E(X^2)$: $E(X^2) = \int_0^{\infty} x^2 \cdot f(x)\,dx$

Since we know that the PDF of an exponential distribution is $f(x) = \lambda e^{-\lambda x}$, then...

$E(X^2) = \int_0^{\infty} x^2\left(2e^{-2x}\right)dx$

$E(X) = 1/\lambda$

$E(X^2) = 2/\lambda^2$ ... with $\lambda = 2$

Variance of a continuous random variable X is

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

$E(X^2) = \frac{2}{\lambda^2} = \frac{2}{2^2} = \frac{2}{4} = \frac{1}{2}$ → Now we can compute the Variance

$V(X) = E(X^2) - (E(X))^2 = 0.5 - (0.5)^2 = 0.5 - 0.25 = 0.25$

$V(X) = 0.25\ yrs^2$

(c) Interpret your findings for parts a and b in the context of the server's processing time.

$E(X) = 0.5$ means that on average, it will take servers 30 minutes to process a query of this type.

$V(X) = 0.25\ hrs^2$ is the spread of variability of the server's processing times around the average, meaning processing times can take significantly longer or shorter times to process based on the query.