EXAMPLES SHEET 2

Probability

Christopher Yau

November 27, 2020

These are some exercises designed to practice your knowledge and ability to apply some of the concepts of probability you learnt in lectures.

Answers are given at the end but try to attempt the questions first before looking at the solutions.

1. (a) Bayes Theorem states that, for two events A and B, the conditional probability:

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

Given that $P(A) = 0.1$ and the conditional probability of B given A is 0.5 and A given B is 0.2.

What is the conditional probability $P(B|\overline{A})$?

Answers:

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

so

$$0.2 = \frac{0.5 \times 0.1}{0.5 \times 0.1 + P(B|\overline{A}) \times 0.9} \Rightarrow P(B|\overline{A}) = \frac{1}{0.9}\left(\frac{0.5 \times 0.1}{0.2} - 0.1 \times 0.5\right) = 0.222$$

(b) In a randomized control trial, breast cancer patients are randomly assigned with probability 0.6 and 0.4 to one of two groups who will receive a standard treatment and a novel experimental agent respectively.

i. Historical data indicates that the standard treatment would clear the disease in 70% of patients. However, in preclinical trials, 85% of those treated with the experimental agent showed no signs of disease.

A. What is the probability that a cancer patient recruited to the trial will be cleared of disease?

Answers:

Apply Total Probabilty:

$$\begin{aligned} P(\text{clear}) &= P(\text{clear}|\text{exp})P(\text{exp}) + P(\text{clear}|\text{standard})P(\text{standard}), \\ &= 0.85 \times 0.4 + 0.7 \times 0.6, \\ &= 0.76 \end{aligned}$$

B. What is the probability that a patient received the standard treatment, given that they were given an all-clear after treatment?

Answers:

Apply Bayes' Theorem:

$$\begin{aligned} P(\text{standard}|\text{clear}) &= \frac{P(\text{clear}|\text{standard})P(\text{standard})}{P(\text{clear})}, \\ &= 0.7 \times 0.6/0.76, \\ &= 0.553 \end{aligned}$$

ii. The trial involves patients recruited from a number of hospitals. In one local regional hospital, 4 patients were recruited.

A. What is the probability that at least one of these 4 patients would get an all-clear after treatment?

Answers:

Let X denote the number of patients with an all-clear:

$$\begin{aligned} P(X > 0) &= 1 - P(X = 0), \\ &= 1 - (1 - 0.76)^4, \\ &= 0.9966822 \end{aligned}$$

i.e. this is one minus the probability that all four patients are not cleared of disease.

B. The local regional hospital only has limited capacity for supporting recurrent disease. What is the probability that 2 or more patients would get disease recurrence?

Answers:

Let X denote the number of recurrences

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1), \\ &= 1 - 0.76^4 - 4 \times 0.24^1 \times 0.76^3, \\ &= 1 - 0.334 - 0.421 \qquad\qquad = 0.245 \end{aligned}$$

The probability of at least two recurrences is one minus the probability that there are no recurrences plus the probability of one recurrence.

The following expression for the probability mass function of the Binomial distribution for n trials and success probability p maybe useful:

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$$

and the binomial coefficient is given by:

$$\binom{n}{x} = \frac{n!}{x!(n - x)!}$$

2. A GP practice receives patients from two local villages. Each village has 500 and 2,000 residents respectively. All residents use the GP practice as it is the only one located in close proximity.

   (a) Historical averages, suggest that 1 in 5 people from the smaller village will use the GP practice in any single year whilst the rate is 2 in 5 from the larger village.

      i. What is the expected number of people who will use the GP practice in any one year?

      Answers:

      The probability that a patient comes from the small village is $500/(500+2000) = 1/5$

      The probability that a patient comes from the large village is therefore $4/5$

      Apply Total Probability to get expected rate of patients in a year:

$$\begin{aligned} P(\text{GP}) &= P(\text{GP}|\text{Small})P(\text{Small}) + P(\text{GP}|\text{Large})P(\text{Large}), \\ &= 0.2 \times 0.2 + 0.4 \times 0.8, \\ &= 0.36 \end{aligned}$$

      Therefore, expected number of patients is $0.36 \times 2500 = 900$

      ii. A patient attends a GP appointment. What is the probability that they come from the small village?

      Answers:

      Apply Bayes' Theorem:

$$\begin{aligned} P(\text{Small}|\text{GP}) &= \frac{P(\text{GP}|\text{Small})P(\text{Small})}{P(\text{GP})}, \\ &= 0.2 \times 0.2/0.36, \\ &= 0.111 \end{aligned}$$

   (b) The GP practice randomly selects 10 individuals from the area it serves to take part in a trial of a new online appointment booking service.

      i. What is the probability that at least one individual will make use of the service during the year?

      Answers:

      The probability of at least one individual making use of the service is one minus the probability that no one uses the service.

      Let X denote the number of trial individuals using the GP:

$$\begin{aligned} P(X > 0) &= 1 - P(X = 0), \\ &= 1 - (1 - 0.36)^5, \\ &= 0.988 \end{aligned}$$

ii. What is the probability that there is at least one individual from each village included in the trial?

Answers:

Let X denote the number of people from village 1 and Y the number from village 2 and $X + Y = 10$.

We want X to be between 1 and 9 since this implies that at least one person from X and Y is represented.

We want $P(X > 0 \cap X < 10) = 1 - P(X = 0 \cup X = 10) = 1 - P(X = 0) - P(X = 10)$

$$\begin{aligned} P(X > 0 \cap X < 10) &= 1 - P(X = 0) - P(X = 10), \\ &= 1 - 0.2^{10} - 0.8^{10}, \\ &= 0.893 \end{aligned}$$

3. (a) A discrete random variable X can take on the values 0, 1, 2, 3 and 4. Its probability mass function has the form:

$$p(X = x) = \frac{1}{Z}\exp(-x)$$

Compute the value of Z then find the expectation $E[X]$ and variance $V[X]$.

Answers:

Since the pmf must sum to 1:

$$\sum_{x=0}^{4} \frac{1}{Z}\exp(-x) = 1$$

then

$$Z = \sum_{x=0}^{4} \exp(-x) = 1.57$$

This means:

$$E[X] = \sum_{x=0}^{4} xp(x) = \sum_{x=0}^{4} x\frac{1}{Z}\exp(-x) = 0.548$$

$$V[X] = \sum_{x=0}^{4} x^2 p(x) - E[X]^2 = \sum_{x=0}^{4} x^2 \frac{1}{Z}\exp(-x) - E[X]^2 = 0.75$$

(b) Two identical six-sided die are rolled. What is the expected value and variance of the difference in the two values?

Answers:

Since the two die are identical and independently thrown. Let X_1 and X_2 denote the values of the two die then:

$$E[X_1 - X_2] = E[X_1] - E[X_2] = 3.5 - 3.5 = 0$$

where $E[X_1] = E[X_2] = (1/6)(1 + 2 + 3 + 4 + 5 + 6) = 3.5$

Recall $V[aX + bY] = a^2V[X] + b^2V[Y]$ so:

$$V[X_1 - X_2] = V[X_1] + (-1)^2V[X_2] = V[X_1] + V[X_2] = 5.84$$

where $V[X_1] = V[X_2] = (1/6)(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) - 3.5^2 = 2.92$.

(c) A random variable Y has the following pmf:

y	0	1	2
$P(Y = y)$	0.5	0.3	0.2

Compute the expected value of $2Y - 1$ and its variance.

Answers:

$$E[2Y - 1] = 2E[Y] - 1 = 0.4$$

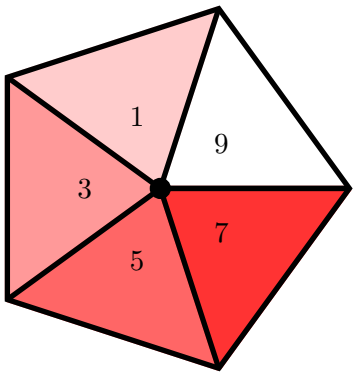
where $E[Y] = 0.5(0) + 0.3(1) + 0.2(2) = 0.7$.

Recall $V[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$:

$$V[2Y - 1] = 2^2 V[Y] = 2.45$$

where $V[Y] = 0.5(0^2) + 0.3(1^2) + 0.2(2^2) - 0.7^2 = 0.61$.

(d) A five sided spinner is spun twice:



What is the probability that the sum of the scores is greater than 10?

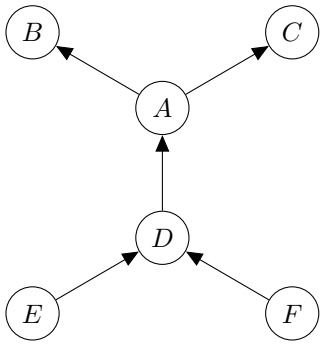
Answers:

There are $5 \times 5 = 25$ possible combinations.

Possible combinations greater than 10: $(3,9)(9,3)(5,9)(9,5)(5,7)(7,5)(7,7)(7,9)(9,7)(9,9)$

Therefore, the probability that the sum of scores is greater than 10 is $10/25 = 2/5$

(e) Given the following directed acyclic graph (DAG):

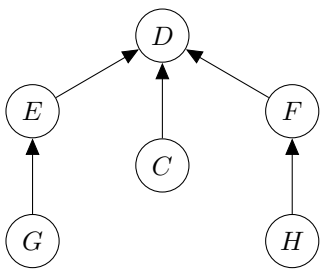


Factorise the joint distribution $P(A, B, C, D, E, F)$.

Answers:

$$P(A, B, C, D, E, F) = P(B|A)P(C|A)P(A|D)P(D|E, F)P(E)P(F)$$

(f) Given the following directed acyclic graph (DAG):



Factorise the joint distribution $P(C, D, E, F, G, H)$.

Answers:

$$P(C, D, E, F, G, H) = P(D|C, E, F)P(E|G)P(F|H)P(G)P(H)P(C)$$

4. (a) A function is given by:

$$h = 2x^2 + 6x + y^2 - 4y + 5$$

i. Find the partial derivatives of h with respect to x and y.

Answers:

$$\frac{\partial h}{\partial x} = 4x + 6,$$
$$\frac{\partial h}{\partial y} = 2y - 4.$$

ii. We apply gradient descent to find the values of (x, y) values that minimise h. Show the form of the gradient descent updates.

Answers:

$$x' = x - \lambda \frac{\partial h}{\partial x} = x - \lambda(4x + 6),$$
$$y' = y - \lambda \frac{\partial h}{\partial y} = y - \lambda(2y - 4).$$

iii. Determine the values of x and y at convergence and the value of h at its minimal value.

Answers:

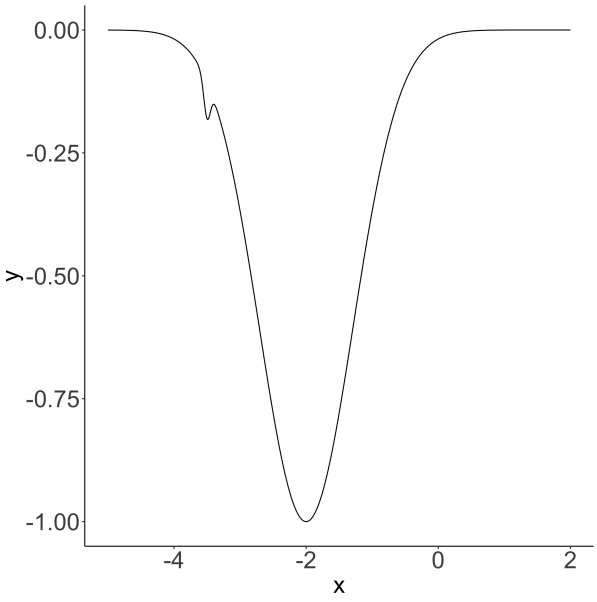
The derivatives at convergence must be zero at convergence so:

$$4x + 6 = 0 \Rightarrow x = -1.5,\ 2y - 4 = 0 \Rightarrow y = 2$$

therefore

$$h = -3.5$$

(b) The following shows a function $y = f(x)$:



Explain why stochastic gradient descent might be preferable to gradient descent when attempting to find the value of x that minimises this function.

Answers:

The function contains a local minima located around $x \approx -3.5$. If gradient descent is initialised from values of $x < -3.5$ then it will converge to this local minimum. Random restarts of the algorithm from different initialisation points might allow the true global minimum to be found. Alternatively, stochastic gradient descent might be another solution since its intrinsic stochasticity could allow it to "jump" out of small local minima like this. However, it may take many iterations before a suitable update step occurs which allows the SGD to escape.

5. In a recycling plant, the proportion of refuse which is Class A or Class B recyclable material occurs with probability θ and $1-\theta$ respectively. The probability θ will be used to determine how many workers to assign to process each type of item.

   You are given a data set which consists of a random sample of 20 independent refuse items that were manually classified and contained 14 Class As and 6 Class Bs.

   (a) If $X_1, X_2, \ldots, X_{20}$ are binary random variables which represent the events that each of the 20 objects belong to Class A (1) or not (0).

   Show that the likelihood $p_\theta(X_1, \ldots, X_{20})$ of observing the sample data given θ is given by:

$$p_\theta(X_1, \ldots, X_{20}) = \theta^{14}(1-\theta)^6$$

   Hint: You may use the fact that the probability mass function for a random variable Z that follows a Bernoulli distribution with parameter h is given by:

$$p(Z = z) = h^z(1-h)^{1-z}$$

   Answers:

   The objects are classified independently so the likelihood is given by:

$$p_\theta(X_1 = A, \ldots, X_{14} = A, X_{15} = B, \ldots, X_{20} = B) = \prod_{i=1}^{20} p(X_i|\theta)$$

   where the individual object assignment probability is given by a Bernoulli distribution:

$$p(X_i|\theta) = \theta^{X_i}(1-\theta)^{1-X_i}$$

   This means:

$$p_\theta(X_1 = A, \ldots, X_{14} = A, X_{15} = B, \ldots, X_{20} = B) = \theta^{14}(1-\theta)^6$$

   since there are 14 Class As and 6 Class Bs.

   (b) Design a gradient descent approach to find the value of θ that maximises the likelihood. What is the form of the gradient descent update expression?

   Answers:

   Take the negative log-likelihood:

$$-\log p_\theta(X_1, \ldots, X_{20}) = -14\log\theta - 6\log(1-\theta)$$

   Differentiate wrt to θ:

$$\frac{d}{d\theta} - \log p_\theta(X_1, \ldots, X_{20}) = -\frac{14}{\theta} + \frac{6}{1-\theta}$$

   Therefore the update is:

$$\theta' = \theta - \lambda\left(\frac{6}{1-\theta} - \frac{14}{\theta}\right)$$

(c) Show that the algorithm converges when the value of the parameter θ reaches 0.7.

Answers:

At the maximum likelihood, the derivative is 0 so solve for $\hat{\theta}$:

$$\begin{aligned} \frac{14}{\hat{\theta}} - \frac{6}{1-\hat{\theta}} &= 0, \\ \Rightarrow\ 14(1-\hat{\theta}) &= 6\hat{\theta}, \\ \Rightarrow\ \hat{\theta} &= 14/20 = 0.7. \end{aligned}$$

Alternatively, plug in the value $\theta = 0.7$ into the derivative and show it is equal to zero.

(d) At a certain point in the gradient descent algorithm, $\theta = 0.65$. What is the value of θ after the next update with step length 0.1? What does this suggest about the step length?

Answers:

The update gives:

$$\theta' = \theta - \lambda \left(\frac{6}{1-\theta} - \frac{14}{\theta} \right) = 0.65 - 0.1 \left(\frac{6}{1-0.65} - \frac{14}{0.65} \right) = 1.08956$$

The update has overshot the minimum and has generated an invalid value of θ (since θ is a probability it must be between 0 and 1).

This implies the step length is too large and should be reduced.

6. The personal salary of online shoppers is known to determine the probability that shoppers will actually make a purchase after browsing an online store.

   You are given the following data from six independent shoppers which includes their relative (to the population average) salary and whether they made a purchase on their last visit to an online store:

Data Item	1	2	3	4	5	6
Relative Salary (x)	0.376	0.918	-0.527	0.557	0.4667	-1.791
Purchase (y)	1	1	0	1	0	0

   You are asked to devise a binary classification algorithm to estimate the probability that a shopper will make a purchase given their salary (x) as an input.

   (a) Suppose you decide to approach this problem using logistic regression. Show that the likelihood of the data is given by:

$$\prod_{i=1,2,4} \frac{1}{1+\exp(-z_i)} \prod_{i=3,5,6} \frac{\exp(-z_i)}{1+\exp(-z_i)}$$

   where $z_i = b_0 + b_1 x_i$ and (b_0, b_1) are regression parameters.

   Answers:

   The likelihood is given by:

$$p(y_1 = 1, y_2 = 1, y_3 = 0, y_4 = 1, y_5 = 0, y_6 = 0) = \prod_{i=1,2,4} p(y_i = 1) \prod_{i=3,5,6} p(y_i = 0)$$

   For logistic regression:

$$p(y_i = 1) = \frac{1}{1+\exp(-z_i)}, \; p(y_i = 0) = 1 - p(y_i = 1) = \frac{\exp(-z_i)}{1+\exp(-z_i)}$$

   Hence,

$$p(y_1 = 1, y_2 = 1, y_3 = 0, y_4 = 1, y_5 = 0, y_6 = 0) = \prod_{i=1,2,4} \frac{1}{1+\exp(-z_i)} \prod_{i=3,5,6} \frac{\exp(-z_i)}{1+\exp(-z_i)}$$

   (b) Show that the update expressions for an online stochastic gradient descent algorithm to estimate the parameters b_0 and b_1 using the given dataset are given by:

$$b_0' = b_0 - \lambda(p(y_i) - y_i),$$
$$b_1' = b_1 - \lambda x_i(p(y_i) - y_i)$$

   Answers:

   The negative log-likelihood (loss function) is given by:

$$L(b_0, b_1) = \sum_{i=1,2,4} \log(1+\exp(-z_i)) + \sum_{i=3,5,6} [z_i + \log(1+\exp(-z_i))]$$

   Online Stochastic Gradient Descent estimates the gradient from one randomly chosen sample so if we pick samples 1, 2, or 4 ($y_1 = 1$) we need the derivatives:

$$\frac{\partial L}{\partial b_0} = \frac{1}{1+\exp(-z_i)} \frac{\partial}{\partial b_0}(1+\exp(-z_i)) = -\frac{\exp(-z_i)}{1+\exp(-z_i)} = -(1 - p(y_i))$$

and

$$\frac{\partial L}{\partial b_1} = -\frac{1}{1+\exp(-z_i)}\frac{\partial}{\partial b_1}(1+\exp(-z_i)) = -\frac{x_i \exp(-z_i)}{1+\exp(-z_i)} = -x_i(1-p(y_i))$$

And if we pick samples 3, 5 or 6 ($y_i = 0$) we need the derivatives:

$$\frac{\partial L}{\partial b_0} = 1 + \frac{1}{1+\exp(-z_i)}\frac{\partial}{\partial b_0}(1+\exp(-z_i)) = 1 - \frac{\exp(-z_i)}{1+\exp(-z_i)} = p(y_i)$$

and

$$\frac{\partial L}{\partial b_1} = x_i + \frac{1}{1+\exp(-z_i)}\frac{\partial}{\partial b_1}(1+\exp(-z_i)) = x_i + \frac{-x_i \exp(-z_i)}{1+\exp(z_i)} = x_i p(y_i)$$

This is consistent with the given updates:

$$b_0' = b_0 - \lambda(p(y_i) - y_i),$$
$$b_1' = b_1 - \lambda x_i(p(y_i) - y_i)$$

(c) You decide to compare the performance of the logistic regression algorithm against a deterministic perceptron algorithm. Given initial weights $(w_0, w_1) = (1, 1)$, where w_0 is the bias term. Determine the classification of each of the data items.

Answers:

Need to compute $z = w_0 + w_1 x_i$ for each item and then take the sign to get class prediction $\hat{y}$:

Data Item	1	2	3	4	5	6
Relative Salary (x)	0.376	0.918	-0.527	0.557	0.4667	-1.791
Purchase (y)	1	1	0	1	0	0
z	1.376	1.918	0.4731	1.557	1.4667	-0.7915
$\hat{y}$	1	1	1	1	1	-1

(d) Now, apply one iteration of the perceptron update, compute the new weights and compute the updated classification.

Answers:

Perceptron update is given by:

$$w_0' = w_0 + y_i,\ w_1' = w_1 + y_i x_i$$

Items 3 and 5 are misclassified (should be -1) so we need to pick one of these to use for the update, lets choose 3 (but we could choose 5 too):

$$w_0' = 1 + (-1) = 0,\ w_1' = 1 + (-1)(-0.527) = 1.527$$

Recomputing the classification:

Data Item	1	2	3	4	5	6
Relative Salary (x)	0.376	0.918	-0.527	0.557	0.4667	-1.791
Purchase (y)	1	1	0	1	0	0
z	0.575	1.402	-0.805	0.851	0.713	-2.735
$\hat{y}$	1	1	-1	1	1	-1

Only item 5 is misclassified.

(e) Explain why the perceptron algorithm will not be able to classify all items correctly.

Answers:

The perceptron is a linear classifier and this data set is not linearly separable. If the decision boundary is placed betwee items 1 and 3, at least one of the data items will have to be misclassified, in this case item 5.