Paper Code

Coláiste na Tríonóide, Baile Átha Cliath
Trinity College Dublin
Ollscoil Átha Cliath | The University of Dublin

Faculty of Engineering, Mathematics and Science

School of Computer Science & Statistics

Integrated Computer Science Programme **Hilary Term 2017**
Year 3

ST3009: Statistical Methods for Computer Science

DD MMM YYYY **Venue** **00.00 – 00.00**

Doug Leith

Instructions to Candidates:

Attempt **all** questions.

You may not start this examination until you are instructed to do so by the invigilator.

Materials Permitted for this examination:

Non-programmable calculators are permitted for this examination – please indicate the make and model of your calculator on each answer book used.

1. Suppose five exams are to be scheduled. There are three possible slots per day when an exam can be held, and six possible days. There can be at most one exam per slot.

   a. In how many different ways can the five exams be scheduled over the slots in the six days ? [5 marks]

   Solution 18 possible slots, 5 exams. Number of ways is 18x17x16x15x14=1028160

   b. In how many different ways can the five exams be scheduled so that each falls on a different day. [5 marks]

   Solution 6 days, 5 exams. 3 slots per day. Number of ways is 18x15x12x9x6=174960

   c. What is the probability that the five exams fall onto different days ? Explain how you arrived at your answer using the axioms of probability. [5 marks]

   Solution Probability is 174960/1028160=0.17. Let S={S_1,S_2,…S_n} be the sample space of all possible schedules of 5 exams over 18 slots and 6 days where n=1028160 is the number of possible schedules. Schedules are mutually exclusive events, so by axioms, P(S)=1=P(S_1)+P(S_2)+…+P(S_n). Suppose each schedule is equally likely and has probability p. Then p=1/n. Let E be the event that exams fall on separate days. Since each schedule in E is mutually exclusive P(E)=|E|p=174960/1028160.

   d. State the definition of conditional probability. [5 marks]

   Solution For events E and F the conditional probability P(E|F)=P(E∩F)/P(F)

   e. Suppose the first exam is scheduled for the first slot of day one. Conditioned on this, what is the probability that the second exam is also scheduled on day one ?

   [5 marks]

   Solution Let F be event that first exam scheduled on slot 1 of day 1 and E the event that the second exam is scheduled on day 1.
   P(F)=1x17x16x15x14/1028160=0.055 (which equals 1/18 as we might expect).
   P(E∩F)= 1x2x16x15x14/1028160=0.0065. P(E|F)=P(E∩F)/P(F)= 0.118

2. On a given day of the week the weather can be either rainy or dry. Suppose the weather on each day of the week is independent of the weather on the other days of the week and that on a given day it rains with probability 0.1.

   a. State the definition of independence of random events. [5 marks]

   Solution Events E and F are independent if P(E∩F)=P(E)P(F).

b. Using independence of the weather on each day calculate the probability that it rains only one day in a given week (i.e. out of seven days) ? [5 marks]

<u>Solution</u> $\binom{7}{1}0.1\text{x}(1\text{-}0.1)^6$

c. What is the probability that it rains four days in a week? [5 marks]

<u>Solution</u> $\binom{7}{4}0.1^4\text{x}(1\text{-}0.1)^3$

d. Let X be a random variable whose value equals the number of rainy days in a week.

i. Derive the expected value E[X]. Hint: use the linearity of the expectation. [5 marks]

<u>Solution</u> Let RV X_i=1 if it rains on day i and 0 otherwise. $P(X_i=1)=0.1$ and so $E(X_i)=0\text{x}P(X_i=0)+1\text{x}P(X_i=1)=0.1$. $E(X)=E(\sum_{i=1}^{7}X_i)=\sum_{i=1}^{7}E(X_i)=7\times 0.1=0.7$

ii. Derive the variance var(X). Hint: use the linearity of the variance of independent random variables. [5 marks]

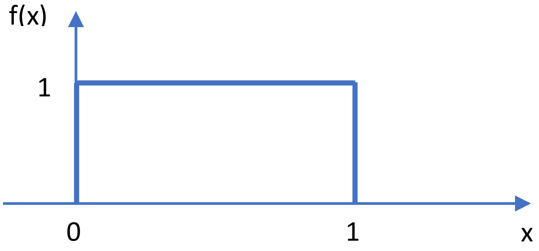
<u>Solution</u> $Var(X_i)=E(X_i^2)-E(X_i)^2$. $E(X_i)^2=0.1^2$ and $E(X_i^2)=0^2\text{x}P(X_i=0)+1^2\text{x}P(X_i=1)=0.1$ so $Var(X_i)=0.1-0.1^2=0.09$. $Var(X)=Var(\sum_{i=1}^{7}X_i)=\sum_{i=1}^{7}Var(X_i)$=7x0.09=0.63

e. Write a short matlab program that draws random values for X. [5 marks]

3. Suppose a continuous random variable X has probability density function f(x)=0 when x≤0 and x≥1 and f(x)=1 when 0<x<1 i.e. the pdf is uniformly distributed between 0 and 1.

a. Calculate the probability that X lies between 0 and 0.25. [5 marks]

<u>Solution</u> Let’s plot f(x) to help us:



So $P(0\le X\le 0.25)=\int_0^{0.25}f(x)dx=0.25$ is the area under the rectangle between 0 and 0.25.

b. What is the probability that X=0.3 ? Explain your answer. [5 marks]

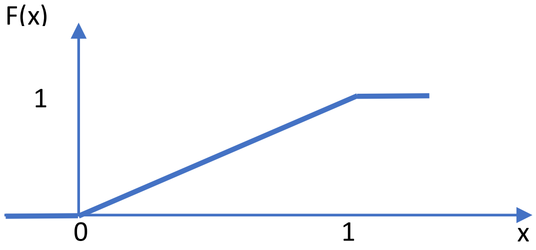
Solution The probability is 0. Since X is continuous valued it is only intervals such as 0≤X≤0.25 that have non-zero probability.

c. State the definition of the cumulative distribution function. [5 marks]

Solution CDF is F(x)=P(X≤x)

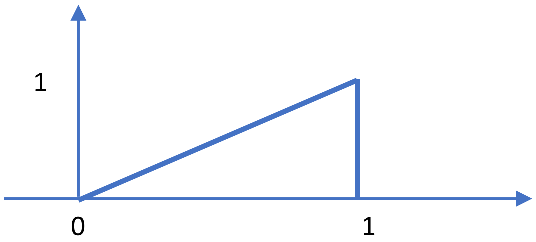
d. Calculate the cumulative distribution function of X. Illustrate with a graph. [5 marks]

Solution F(x) is the area under the pdf f between −∞ and x. So F(x)=0 for x≤0, F(x)=x for 0≤x≤1 and F(x)=1 for x>1.



e. Calculate the expected value E[X]. Hint: recall that for continuous random variables $E[X] = \int_{-\infty}^{\infty} xf(x)dx$, which in this example equals $E[X] = \int_0^1 xf(x)dx$ since f(x)=0 for x≤0 and x≥1. [5 marks]

Solution $E[X] = \int_0^1 xf(x)dx = \int_0^1 xdx$. That is, the area of the triangle:



The area of a triangle is half the base times the height, which in this case is 0.5x1x1=0.5. So E[X]=0.5

4. Suppose we are given patient data consisting of n pairs (x_1,y_1), (x_2,y_2), … (x_n,y_n) where x_i is the measured blood sugar level and y_i equals 1 if the patient has been diagnosed diabetic and 0 otherwise.

   a. With reference to Bayes Rule explain what is meant by the likelihood, prior and posterior. Explain how the maximum likelihood estimate of a model parameter differs from the maximum a posteriori (MAP) estimate. [5 marks]

Solution For events E and F, Bayes rule is P(E|F) = P(F|E)P(E)/P(F). P(F|E) is the likelihood, P(E) the prior, P(E|F) the posterior. A maximum likelihood estimate of parameter θ maximises P(F|θ) whereas a MAP estimate maximises P(θ|F)

Suppose we decide to use a logistic regression model for this patient data.

b. Give the expression for p(x_i | y_i) used in a logistic regression model? [5 marks]

Solution p(x_i | y_i,θ)=1/(1+exp(-z)) with z= θ y_i

c. Now give the expression for the likelihood of the observed data (x_1,y_1), (x_2,y_2), … (x_n,y_n) [5 marks]

Solution Let D={(x_1,y_1), (x_2,y_2), … (x_n,y_n)}. Then the likelihood is:

$$P(D|\theta) = \prod_{i=1}^{n} p(x_i|y_i,\theta) = \prod_{i=1}^{n} \left(\frac{1}{1+\exp{(-\theta y_i}}\right)^{x_i} \left(1 - \frac{1}{1+\exp{(-\theta y_i}}\right)^{(1-x_i)}$$

d. Discuss why a logistic regression model requires linear separability. [5 marks]