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 2018**
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. You buy one share of stock in company C for €10. Each day the price of C either increases by €1 with probability p or decreases by €1 with probability 1-p. These changes from day to day are statistically independent. You decide to sell your share if it gains €2 (i.e. reaches a price of €12).

   (i) What is the probability that you will sell your share exactly 4 days after you buy it ? [5 marks]

   (ii) What is the probability that you sell your share at least 4 days after you buy it ? [5 marks]

   Suppose now that the daily change in the price of stocks in company C is observed to be related to the change in price of stocks in company D. Namely, the probability that stock in C increases by €1 is equal to 0.2 when the price of stock in company D increases that day, and is equal to 0.1 otherwise.

   (iii) State the definition of conditional probability. [5 marks]

   (iv) Describe how marginalisation can be used to calculate the probability of an event E based on knowledge of the conditional probabilities $P(E|F_1)$, $P(E|F_2)$ and $P(E|F_3)$ plus the probabilities $P(F_1)$, $P(F_2)$ and $P(F_3)$ when events F_1, F_2, F_3 are mutually exclusive and $F_1 \cup F_2 \cup F_3$ equals the sample space. [5 marks]

   (v) Suppose that the probability that stock in company D increases on a given day is 0.5. Calculate the probability that stock in company C increases that day. [5 marks]

Model Solution

(i) Prob sell in 4 days = Prob of one increase and decrease followed by two increases = $2p^3(1-p)$. Note that if had four increases then would have sold on day 2 (when had increased by €2).

(ii) Prob sell in 4 or more days = 1-(Prob sell in day 1 + Prob sell in day 2+Prob sell in day 3). Prob sell in day 1=0 (since need increase of €2 at least). Prob sell in day 2 = p^2. Prob sell in day 3 = 0 (if have 3 wins then sell on day 2, if a loss and 2 wins then don't sell). So answer is $1-p^2$

(iii) For random events E and F, $P(E|F) = P(E \cap F)/P(F)$

(iv) $P(E) = P(E|F_1)P(F_1)+ P(E|F_2)P(F_2) + P(E|F_3)P(F_3)$

(v) P(C increases) = P(C increases|D increases)P(D increases) + P(C increases|D doesn't increase)P(D doesn't increase) = $0.2 \times 0.5+ 0.1\times(1-0.5)$

© Trinity College Dublin, The University of Dublin 2018

2. Suppose you play a game where four 6-sided fair dice are rolled. Let X be equal to the minimum of the four values rolled (it is ok if more than one dice has the minimal value). It costs €2 to play the game and you win €X.

   (i) Calculate $P(X \geq k)$ as a function of $k=1,2,\ldots,6$. [5 marks]

   (ii) Assuming you know $P(X \geq k)$ for $k=1,2,\ldots,6$, show how to calculate the PMF of X. [5 marks]

   (iii) State the definition of the expected value. [5 marks]

   (iv) Calculate E[X]. [5 marks]

   (v) If you play the game many times do you expect to make a profit (win more than you pay to play the game) ? Explain your reasoning. What is the amount cost to play that would make you break even (i.e. have an expected profit of zero) ? [5 marks]

Model Solution

(i) $P(X \geq 1) = 1$ since dice must come up one or higher. $P(X \geq 2) = (5/6)^4$ since probability of one dice rolling 2 or higher is 5/6, and since dice rolls are independent the probability that all four dice at 2 or greater is $(5/6)^4$. By similar reasoning $P(X \geq 3) = (4/6)^4$, $P(X \geq 4) = (3/6)^4$, $P(X \geq 5) = (2/6)^4$, $P(X \geq 6) = (1/6)^4$.

(ii) $P(X=6)=P(X \geq 6) = (1/6)^4$. $P(X=5)= P(X \geq 5)- P(X \geq 6)$, $P(X=4)= P(X \geq 4)- P(X \geq 5)$, etc.

(iii) For RV X taking values $x_1, x_2, \ldots, x_n$ then $E[X] = x_1P(X=x_1)+x_2P(X=x_2)+\ldots+x_nP(X=x_n)$

(iv) $E[X] = 1.P(X=1)+2P(X=2)+\ldots+6P(X=6)$ using values from (ii) above.

(v) The expected profit is E[X]-2. If we play the game N times, for N sufficiently large then our profit is N(E[X]-2) with high probability. To break even the cost to play would be E[X].

© Trinity College Dublin, The University of Dublin 2018

3. A survey is carried out by selecting n people from the population and asking each person to answer either "yes" or "no" to a question. Let random variable Y_i take value 1 when the i'th respondent answers "yes" and 0 otherwise. The random variables Y_i i=1,2,..,n are independent and identically distributed with $E[Y_i]= \mu$.

   (i) Let random variable $Z = \sum_{i=1}^{n} Y_i$. Write an expression for E[Z] in terms of $E[Y_i]$. Explain your answer. Hint: use the linearity of the expected value.
   [5 marks]

   (ii) Using the definition of expectation prove that E[Z/n]=E[Z]/n for n>0.
   [5 marks]

   (iii) Using Chebyshev's inequality explain the weak law of large numbers and the behaviour of $|Z/n - \mu|$ as n becomes large. Recall that for random variable X Chebyshev's inequality is: $P(|X - \mu|\geq k) \leq E[(X- \mu)^2]/k^2$ for any k and μ.
   [5 marks]

   (iv) Explain what a confidence interval is, using Z/n as an estimate of μ as an example. [5 marks]

   (v) Describe how to use bootstrapping to estimate a confidence interval for Z/n.
   [5 marks]

Model Solution

(i) $E[Z] = E[\sum_{i=1}^{n} Y_i] = \sum_{i=1}^{n} E[Y_i] = n E[Y_i]$ where we use linearity of the expectation to move the E[] inside the sum

(ii) $E[Z/n] = \sum_{i=1}^{n} i/n\, P(Z=i) = (\sum_{i=1}^{n} i\, P(Z=i))/n = E[Z]/n$

(iii) $E[Z/n]=E[Z]/n= E[Y_i]= \mu$. $Var(Z/n)=1/n^2 Var(Z) = nVar(Y_i)/n^2$. Plugging these values into Chebyshev's inequality, $P(|Z/n - \mu|\geq k) \leq Var(Yi)/nk^2$. For any value of k>0, as n goes to infinity then $Var(Yi)/nk^2$ goes to zero.

(iv) A confidence interval is an interval [a,b] within which a RV X lies with a specified probability e.g. with probability at least 0.95. This can be written $P(a \leq X \leq b) \geq 0.95$. In the case of Z/n as an estimate of μ we might consider the interval $P(\mu-\varepsilon\leq Z/n\leq \mu+\varepsilon)$. From (iii) we know that $P(\mu-\varepsilon\leq Z/n\leq \mu+\varepsilon)$ goes to 1 for any $\varepsilon>0$ as n grows large.

(v) From the observed data Y*i*, i=1,..,n, draw a sample of m points uniformly at random with replacement. Using this sample calculate estimate Z/n. Repeat to obtain a number of estimates. From the distribution of these estimates we can then estimate a confidence interval for Z/n.

© Trinity College Dublin, The University of Dublin 2018

4. Suppose we mark the answers of 200 students to each of 10 exam questions. Let S_{ij} be an indicator variable which is 1 if student i answered question j correctly and -1 otherwise. You observe all of the answers for all students. Assume that

$$P(S_{ij}=y \mid a_i, d_j) = 1/(1+\exp(-y(a_i-d_j))$$

where a_i is a parameter that represents the students ability and d_j is a parameter which represents the questions difficulty.

(i) Give an expression for the log-likelihood of this exam data (the data consisting of the answers by all 200 students). Hint: this is an example of a logistic regression model. [5 marks]

(ii) Outline how gradient descent might be used to find the maximum likelihood estimates for the unknown parameters a_i and d_j. [5 marks]

(iii) With reference to Bayes Rule explain what is meant by the likelihood, prior and posterior. [5 marks]

(iv) Explain how the maximum a posteriori (MAP) estimate of a parameter differs from the maximum likelihood estimate. [5 marks]

(v) How could you incorporate knowledge of the prior probability distribution of parameters a_i into the above model to obtain a MAP estimate ? [5 marks]

Model Solution

(i) The likelihood of the observed mark data s_{ij}, i=1,..,200, j=1,…,10 is $P\big(S_{ij} = s_{ij}, i = 1,..200, j = 1,..,10 | a_i, d_j\ i = 1,..200, j = 1,..,10\big) = \prod_{i=1}^{200}\prod_{j=1}^{10} P(S_{ij} = s_{ij} | a_i, d_j)$. The log-likelihood is $L = \mathrm{Log}\prod_{i=1}^{200}\prod_{j=1}^{10} P(S_{ij} = s_{ij} | a_i, d_j) = \sum_{i=1}^{200}\sum_{j=1}^{10} \log P(S_{ij} = s_{ij} | a_i, d_j) = -\sum_{i=1}^{200}\sum_{j=1}^{10} \log(1 + \exp(-s_{ij}(a_i - d_j))$

(ii) The ML estimates select the parameters a_i and d_j to maximise the likelihood L. Starting from an intial estimate, these values can be found by iteratively updating the estimates such that L decreases after each update until the decrease in L becomes small enough. We can find updates that decrease L by local search or by taking a step in the direction of the derivatives of L wrt a_i and d_j.

(iii) For random 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 and P(E|F) the posterior.

(iv) In a MAP estimate the parameter values are selected to maximise ths posterior probability P(parameters|data) rather than the likelihood P(data|parameters).

(v) By Bayes, the posterior is proportional to $P\big(S_{ij} = s_{ij}, i = 1,..200, j = 1,..,10 | a_i, d_j, s_{ij}, i = 1,..200, j = 1,..,10\big)P(a_i, i = 1,..200)$. The MAP estimate of the a_i, i=1,..,200 maximises this.

© Trinity College Dublin, The University of Dublin 2018