MVE137 Exam

Probability and Statistical Learning Using Python

Total time: 4 h, 14:00-18:00
Total points (part I + part II): 70

25 August 2022

Grade scale: 40 pt: 3, 60 pt: 4, 80 pt: 5. More info on Canvas about requirements. Carl and Charitha will come by around 15:00 and 17:00 if you have any questions. Also available by phone, Carl: 031 772 **16 09**, Charitha: 031 772 **15 73**. Questions for Examiner/Teachers: Giuseppe: 031 772 **18 02**, Alex: 031 772 **17 53**.

Allowed aids: Chalmers approved calculator.

Part I

1. Suppose that we roll a fair k-sided dice twice with the numbers 1 through k on the dice's faces, obtaining the values X_1 and X_2. Let $Y_1 = \max(X_1, X_2)$ and $Y_2 = \min(X_1, X_2)$.

   (a) What is $\mathbb{P}[Y_1 = 1]$? What is $\mathbb{P}[Y_1 = 2]$? More generally, what is $\mathbb{P}[Y_1 = n]$, for $n = 1, 2, \dots, k$? **(1 pt)**

   (b) How are $\mathbb{P}[Y_1 = n]$ and $\mathbb{P}[Y_2 = k + 1 - n]$ related? **(2 pt)**

   (c) Use the result in point (a) to compute $\mathbb{E}[Y_1]$. **(2 pt)**
   The following two formulas may be useful

$$\sum_{i=1}^{k} i = \frac{k(k+1)}{2}$$

$$\sum_{i=1}^{k} (2i^2 - i) = \frac{k(k+1)(4k-1)}{6}.$$

   (d) Verify that $\mathbb{E}[Y_2] = k + 1 - \mathbb{E}[Y_1]$. **(2 pt)**

   (e) Verify that $\mathbb{E}[Y_1] + \mathbb{E}[Y_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$. Why is this result not surprising? **(2 pt)**

2. Let X be the number of heads observed when flipping a fair coin n times, where n is an even number.

   (a) Verify that the following symmetry property holds: for $k < n/2$ **(2 pt)**

$$\mathbb{P}[X \geq n/2 + k] = \mathbb{P}[X \leq n/2 - k]$$

   (b) Use this result to conclude that **(3 pt)**

$$\mathbb{P}\Big[X \geq \frac{n}{2} + k\Big] = \frac{1}{2} - \frac{1}{2}\mathbb{P}\Big[X = \frac{n}{2}\Big] - \mathbb{P}\Big[\frac{n}{2} - k < X < \frac{n}{2}\Big]$$

   (c) Compute now $\mathbb{P}[X \geq 55]$ and $\mathbb{P}[X \geq 550]$ using the expression in (b). **(2 pt)** You might find the following approximations useful to simplify your calculations:

$$\binom{100}{50} \approx 1.01 \cdot 10^{29} \qquad \sum_{k=46}^{49} \binom{100}{k} \approx 3.50 \cdot 10^{29} \qquad 2^{100} \approx 1.27 \cdot 10^{30}$$

$$\binom{1000}{500} \approx 2.70 \cdot 10^{299} \qquad \sum_{k=451}^{499} \binom{1000}{k} \approx 5.21 \cdot 10^{300} \qquad 2^{1000} \approx 1.07 \cdot 10^{301}$$

   (d) Use Chernoff bound to obtain an upper bound on $\mathbb{P}[X \geq 55]$ and $\mathbb{P}[X \geq 550]$. Discuss the tightness of the bound. **(3 pt)**

3. A particular discrete-time finite Markov chain has the property that the set of states can be partitioned into L classes $\{C_0, C_1, \cdots, C_{L-1}\}$ such that for all states $i \in C_\ell$, the transition probability $P_{ij} = 0$ for all $j \notin C_{\ell+1 \bmod L}$.

(a) Set $L = 3$ and draw a Markov chain satisfying this property. **(2 pt)**

(b) Prove that all states have period $d = L$. **(5 pt)**

4. A chi-squared $\chi^2(k)$ distribution, where the integer k denotes the number of degrees of freedom, is a continuous-time distribution that plays an important role in hypothesis testing. Its probability density function is given by

$$f_X(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$$

where $\Gamma(\cdot)$ denotes the Gamma function, and its moment-generating function is

$$M_X(t) = (1-2t)^{-k/2}, \quad \text{for } t < 1/2.$$

Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(0,1)$ variables.

(a) Show that X_1^2 follows a chi squared distribution with 1 degree of freedom $\chi^2(1)$. *Hint:* Recall that $\Gamma(1/2) = \sqrt{\pi}$. **(5 pt)**

(b) More generally, show that $\sum_{i=1}^n X_i^2$ is $\chi^2(n)$ distributed. *Hint:* It may be convenient to work with the moment-generating function to establish this result. **(4 pt)**

5. Suppose you have access to a European database consisting of one million individual trees of various types which include the following entries:

- Tree type (birch, pine, aspen, etc.)
- Age
- Height
- Circumference
- Geographical coordinate of the position of the tree
- Vegetation type (openwoodland, mixedwood, highland, wet coniferous, etc.)

Consider a regression problem where you want to model the age of a tree based on its height and circumference. We use the following linear regression model:

$$y = f(x) + \epsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon\,,$$

where the input variables x_1 and x_2 represent the height and the circumference of the tree, respectively, and the output y is the age.

(a) What causes the bias and variance of the model? **(3 pt)**

(b) Will the bias and the variance be high or low? **(2 pt)**

Consider the same European database as above and now consider a classification problem where the tree class is the output and the geographical coordinates the input. We use a k-nearest neighbor (KNN) model with $k = 1$

(a) What causes the bias and variance of the model?**(3 pt)**

(b) Will the bias and the variance be high or low? **(2 pt)**

6. Suppose that you collect $n = 200$ observations of a single variable x and its single output y. You then go ahead and fit the linear regression model

$$y = f(x) = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

to the first half of your data (the training data), as well as a cubic polynomial

$$y = f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon \tag{2}$$

(a) Consider that the true relationship between x and y is linear. Which of the models, (1) or (2), will be able to fit your training data the best, i.e., which model gives you the smallest E_{train}? Explain. **(5 pt)**

(b) Consider (a) again, but suppose that the true relationship is not linear. Which model will have the smallest E_{train}? Explain. **(5 pt)**

Formulas

You might find the following formulas helpful.

- **Definition 1 (Binomial Random Variable)** *A binomial random variable with parameters n and p, denoted by $B(n, p)$, is defined by the following probability distribution on $j = 0, 1, \dots, n$:*

  $$\mathbb{P}[X = j] = \binom{n}{j} p^j (1-p)^{n-j}.$$

  A Binomial-distributed random variable has mean np and variance $np(1-p)$. The Moment Generating Function for a binomial variable is

  $$M_x(t) = \left(1 - p + pe^t\right)^n$$

- **Theorem 2 (Chernoff bound upper deviations)** *Let $X_i, \dots, X_n$ be independent Bernoulli random variables such that $\mathbb{P}[X_i = 1] = p_i$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X]$. Then the following Chernoff bounds hold:*

  – *For every $\delta > 0$,*

  $$\mathbb{P}[X \geq (1+\delta)\mu] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}.$$

  – *For every $0 < \delta \leq 1$*

  $$\mathbb{P}[X \geq (1+\delta)\mu] \leq e^{-\mu\delta^2/3}.$$

- We have the following results concerning a discrete Markov chain

  – **Definition 3 (Accessibility)** *State j is accessible from state i, which we write as $i \to j$ if $P_{ij}(n) > 0$ for some $n > 0$.*

  – **Definition 4 (Communicating States)** *States i and j communicate, which we write as $i \leftrightarrow j$, if $i \to j$ and $j \to i$.*

  – **Definition 5 (Communicating Class)** *A communicating class is a nonempty subset of states $\mathcal{C}$ such that if $i \in \mathcal{C}$, then $j \in \mathcal{C}$ if and only if $i \leftrightarrow j$.*

  – **Definition 6 (Periodic and Aperiodic States)** *State i has period d if d is the largest integer such that $P_{ii}(n) = 0$ whenever n is not divisible by d. If $d = 1$, then the state i is called aperiodic.*

  – **Theorem 7** *Communicating states all have the same period.*

  – **Definition 8 (Transient and Recurrent States)** *In a finite Markov chain, a state i is transient if there exists a state j such that $i \to j$ but $j \not\to i$. If no such state j exists, then the state i is recurrent.*

- **Definition 9** *The univariate normal distribution is characterized by two parameters μ and σ corresponding to the mean and the standard deviation and is denoted by $\mathcal{N}(\mu, \sigma^2)$. Its density function is*

  $$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}.$$

- **Definition 10 (Moment Generating Functions)** *The moment generating function of a random variable X is defined as*

  $$M_X(t) = \mathbb{E}\left[e^{tX}\right].$$

- **Theorem 11 (Moment Generating Function of Sum of Independent RVs)** *If X and Y are independent random variables then*

  $$M_{X+Y}(t) = M_X(t) \cdot M_Y(t).$$

Part II

It has been assigned as take-home exam on Wednesday, 17 August 2022. It is reported here for completeness.

Take-Home Exam: MVE137 Probability and Statistical Learning Using Python

Formalities

This is the take-home part of the re-exam for the course Probability and Statistical Learning Using Python, 2021. Here, you are asked to carry out the analysis using the tools and techniques from the course and hand in a .pynb file with solutions.

The **deadline is Thursday, August 25, 2022.** You should upload the solution file to “Take-Home Exam SP4” in Canvas via “Home–>Exam—>Take-Home Exam SP4”. Note that this is an individual exam.

We will use the *Seats* data set which is provided in the Canvas page.

1. (a) Fit a multiple regression model to predict Sales using Advertising, Urban, and US. **(3 pts)**

   (b) Is there a statistical significant relationship between Sales and whether the store is in the US or not ? (Be careful, some of the variables in the model can be qualitative!) **(2 pts)**

   (c) Is there a statistically significant relationship between Sales and whether the store is in an urban or rural area? (Be careful, some of the variables in the model can be qualitative!) **(2 pts)**

   (d) Write out the model in equation form, being careful to handle the qualitative variables properly. **(2 pts)**

   (e) Now, fit a smaller model that only uses two predictors for which there is evidence of highest association with the outcome. **(3 pts)**

   (f) How well do the models in (a) and (e) fit the data? **(3 pts)**