MVE137 Exam

Probability and Statistical Learning Using Python

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

3 January 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.

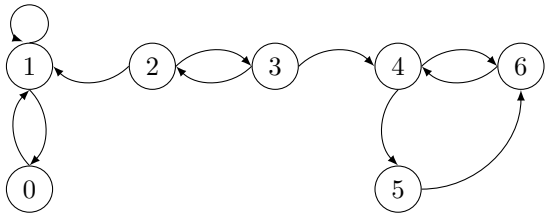
Part I

1. A monkey types on a 26-letter english keyboard that has lowercase letters only. Each letter is chosen independently and uniformly at random from the English alphabet.
   (a) What is the probability that the monkey types the letter "p"? **(1 pt)**
   (b) What is the expected number of letters that the monkey has to type until the letter "p" appears? **(2 pt)**
   (c) What is the expected number of letters that the monkey has to type until the sequence "pr" appears? **(3 pt)**
   (d) And what is the expected number of letters until the word "proof" appears? **(3 pt)**

   Hint: Use the following conditioning trick: if M is the number of letters until the word "proof" appears and X_n is the nth letter in the word "proof", write first $\mathbb{E}[M]$ in terms of the conditional expectation of M given that $X_1 =$ "p". Proceed in a similar way to evaluate $\mathbb{E}[M|X_1 =$ "p"$]$ in terms of the conditional expectation of M given $X_1 =$ "p" and $X_2 =$ "r" and so on...

2. Each person in Sweden is issued with a personal number with the following structure: YYYYMMDD - XXXX, where YYYY is the year you are born, MM is the month, and DD is the day. We will assume that the last four digits XXXX are uniformly and independently generated for each person. You are at a class reunion with your N classmates who are all born the same year as you and you have stolen everyone's bankcard. Some people use their last four numbers as a PIN for their bankcard, and we assume that this is the case for all of your classmates.
   (a) You try to take out 200 SEK from an ATM by using your own last four numbers with each bankcard. How many classmates would you need for the probability to get no money to be less than 0.5? **(3 pt)**
   (b) What is the expected amount of money that you will get with N classmates? **(3 pt)**
   (c) You come back to the reunion with $z > 0$ SEK in cash and quietly give everyone their cards back. You wonder what the probability is that two or more of the now $N + 1$ people in the room have the same personal number as you. You don't know your classmates so well, so you assume uniform and independent birthdays (recall that everyone is born the same year). What is the probability that two or more of the now $N + 1$ people in the room have the same personal number as you given that you were able to get $Z = z > 0$ SEK from the bankcards (note that z must be a multiple of 200 SEK)? **(6 pt)**

3. In this Markov chain, all transitions with nonzero probability are shown.



(a) What are the communicating classes? **(3 pt)**

(b) For each communicating class, identify whether the states are periodic or aperiodic. **(3 pt)**

(c) For each communicating class, identify whether the states are transient or recurrent. **(3 pt)**

4. You would like to perform a simulation that uses exponentially distributed random variables. The programming language you are using has a random number generator that produces independent, uniformly distributed numbers from the real interval (0,1). Give a procedure that transforms a uniform random variable distributed on (0,1) into an exponentially distributed random variable with mean μ. **(6 pt)**

   Hint: Compute the inverse of the CDF of an exponential distribution. What is the domain of the inverse? What happens if you set the input to the inverse of the CDF equal to a random variable that is uniform on (0,1)?

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 of the model? Will the bias be high or low? **(3 pt)**

   (b) What causes the variance of the model? Will the variance be high or low? **(3 pt)**

   (c) What causes the irreducible error of the model? **(2 pt)**

   (d) Where do you see the biggest potential improvement of the model (reducing bias, variance, or irreducible error) and how would you improve it? **(2 pt)**

6. Consider the following simple probabilistic model,

   $$y = \tilde{f}(x) + \epsilon, \qquad \tilde{f}(x) = 1\,,$$

   where ϵ is Gaussian-distributed with mean 0 and variance σ^2. We assume linear regression with only an intercept term, i.e., we assume the model

   $$y = f(x) + \epsilon = \beta_0 + \epsilon$$

   where we learn β_0 using ridge regression with regularization parameter λ. We recall that, in the general case, the solution of ridge regression is

   $$\boldsymbol{\beta}^*_{\mathsf{ridge}} = (\mathbf{X}^\mathsf{T}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\mathsf{T}\boldsymbol{y}\,,$$

   Assume that we have one data sample y_1.

   (a) What is the closed-form solution for the optimal value β_0^* as a function of the training data y_1 and the regularization parameter λ? What is $f(x;\mathcal{D})$ where $\mathcal{D}$ is the training data? **(3 pt)**

   (b) What is the average trained model $\mathbb{E}_\mathcal{D}[f(x;\mathcal{D})]$? The expectation operator $\mathbb{E}_\mathcal{D}$ is an expectation over all random variations in the training data. **(3 pt)**

   (c) What is the squared bias $\mathbb{E}_\mathsf{x}[(\tilde{f}(x) - \mathbb{E}_\mathcal{D}[f(x;\mathcal{D})])^2]$? Here, the expectation operator $\mathbb{E}_\mathsf{x}$ is the expectation over the test input $x \sim p(x)$. **(2 pt)**

   (d) What is the variance $\mathbb{E}_\mathsf{x}[\mathbb{E}_\mathcal{D}[(f(x;\mathcal{D}) - \mathbb{E}_\mathcal{D}[f(x;\mathcal{D})])^2]]$? **(2 pt)**

Formulas

You might find the following formulas helpful.

- **Theorem 1** *For any two RV's Y, Z*
$$\mathbb{E}[\mathbb{E}[Y \mid Z]] = \mathbb{E}[Y]\,. \tag{1}$$

- **Definition 2 (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 $j = 0, 1, \ldots, n$:*
$$\mathbb{P}[X = j] = \binom{n}{j} p^j (1-p)^{n-j}. \tag{2}$$
A Binomial-distributed random variable has mean np and variance $np(1-p)$.

- **Definition 3 (Geometric Random Variable)** *A geometric random variable X with parameter p is specified by the following probability distribution on $n = 1, 2, \ldots$:*
$$\mathbb{P}[X = n] = (1-p)^{n-1} p. \tag{3}$$
A geometric-distributed random variable has mean $1/p$ and variance $(1-p)/p^2$.

- We have the following results concerning a discrete Markov chain
  - **Definition 4 (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 5 (Communicating States)** *States i and j communicate, which we write as $i \leftrightarrow j$, if $i \to j$ and $j \to i$.*
  - **Definition 6 (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 7 (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 8** *Communicating states all have the same period.*
  - **Definition 9 (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 10** *An exponential distribution with parameter θ is given by the following probability distribution*
$$F(x) = \begin{cases} 1 - e^{-\theta x}, & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$
The density function is
$$f(x) = \theta e^{-\theta x}, \quad \text{for } x \geq 0. \tag{5}$$
The mean is $\mathbb{E}[X] = 1/\theta$ and the variance is $\mathbb{V}\text{ar}[X] = 1/\theta^2$.

- We say that a random variable X that assumes values in the interval $[a, b]$ is uniformly distributed if all subintervals of equal lengths have the same probability. The density function of such random variable is
$$f(x) = \begin{cases} 0 & \text{if } x < a \\ \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{if } x > b. \end{cases} \tag{6}$$
The distribution function is given by
$$F(X) = \begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x \leq b \\ 1 & \text{if } x > b. \end{cases} \tag{7}$$
The expectation of X is $\mathbb{E}[X] = (b+a)/2$, and the variance is $\mathbb{V}\text{ar}[X] = (b-a)^2/12$.

Part II

It has been assigned as take-home exam on Monday, 27 December 2021. 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 Wednesday, January 05, 2022.** You should upload the solution file to "Take-Home Exam SP2" in Canvas via "Home–>Exam—>Take-Home Exam SP2". 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 the target variable Sales using the features Price, Urban, and US. **(3 pts)**

   (b) Interpret each coefficient in the model. (Be careful, some of the variables in the model are qualitative!) **(3 pts)**

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

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

   (e) How well do the models in (a) and (d) fit the data?. **(2 pts)**

   (f) Is there evidence of outliers in the model from (d)? **(2 pts)**