MVE137 Exam

Probability and Statistical Learning Using Python

Total time: 4 h, 08:30 – 12:30
Total points (part I + part II): 70

29 October 2021

Grade scale: 40 pt: 3, 60 pt: 4, 80 pt: 5. More info on Canvas about requirements. Carl and Charitha will come by at 09:30 and 11:30 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. We roll a fair dice over and over. Let N be the number of rolls until the first pair of consecutive sixes appear. What is $\mathbb{E}[N]$? (*Hint*: Let X_i be the outcome of the ith roll and $p = \mathbb{P}[X_i = 6]$. Argue, by conditioning on the value of X_1, that $\mathbb{E}[N] = (1-p)\,\mathbb{E}[N] + (1-p) + p\,\mathbb{E}[N \mid X_1 = 6]$. Evaluate $\mathbb{E}[N \mid X_1 = 6]$ in a similar way by conditioning on X_2) **(5 pt)**

2. The weak law of large numbers states that, if $X_1, X_2, X_3, ...$ are independent and identically distributed random variables with mean μ and variance σ^2 , then for any constant $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{X_1 + X_2 + ... + X_n}{n} - \mu\right| > \varepsilon\right] = 0$$

   (a) Use Chebyshev's inequality to prove the weak law of large numbers. **(5 pt)**

   (b) Show that the weak law of large numbers also applies with the relaxed conditions $\mathbb{E}[X_i] = \mu \,\forall\, i$, $\sigma_i^2 = \mathrm{Var}[X_i] \,\forall\, i$, where all σ_i are not necessarily equal but bounded. **(3 pt)**

3. A checkout counter has a "Take-a-Penny-Leave-a-Penny" bin. The kth customer's charge (in pennies) is a random variable X_k such that $Y_k = X_k \bmod 5$ is a discrete uniform (0, 4) random variable, independent of any other transaction. When customer k checks out, the penny bin goes unused if $Y_k \in \{0, 2, 3\}$. If $Y_k = 1$ and the penny bin has at least one penny, the customer takes a penny from the bin to pay the clerk. If $Y_k = 4$, the customer receives a penny in change from the clerk and deposits this in the penny bin. If this is the fourth penny in the bin, the clerk puts the four pennies in his pocket immediately after the customer departs, leaving the penny bin empty. (The clerk views these four pennies as a reward for providing the penny bin.)

   (a) Let N_k denote the number of pennies in the bin when the kth customer is served. Construct a discrete time Markov chain for N_k and find the stationary probabilities $\pi_n = \lim_{k\to\infty} \mathbb{P}[N_k = n]$. **(6 pt)**

   (b) Let R_k denote the reward received by the clerk after serving the kth customer. What is $\lim_{n\to\infty} \mathbb{E}[R_k]$? In other words, on average how many pennies can the clerk expect per customer? **(6 pt)**

4. The Chernoff bound states that for a random variable X,

$$\mathbb{P}[X \geq a] \leq \min_{t>0} \frac{M_X(t)}{e^{ta}}.$$

   Here, $M_X(t) = \mathbb{E}\left[e^{tX}\right]$ is the moment-generating function (MGF) of the random variable X. Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Show that the MGF for a Gaussian RV is **(3 pt)**

$$M_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right).$$

   Derive the best Chernoff bound by optimizing over t. In particular, show that for $\gamma \geq 0$ **(7 pt)**

$$\mathbb{P}[X \geq \mu + \gamma] \leq \exp\left(-\frac{\gamma^2}{2\sigma^2}\right).$$

5. We consider the regression problem of predicting $y \in \mathbb{R}$ as a function of the input $x \in \mathbb{R}$. We will consider the following regression models:

   i Linear regression

   ii Polynomial regression with degree-2 polynomial

   iii Polynomial regression with degree-10 polynomial

   iv Ridge regression

   v k-nearest neighbor regression

   We consider the following regression problems.

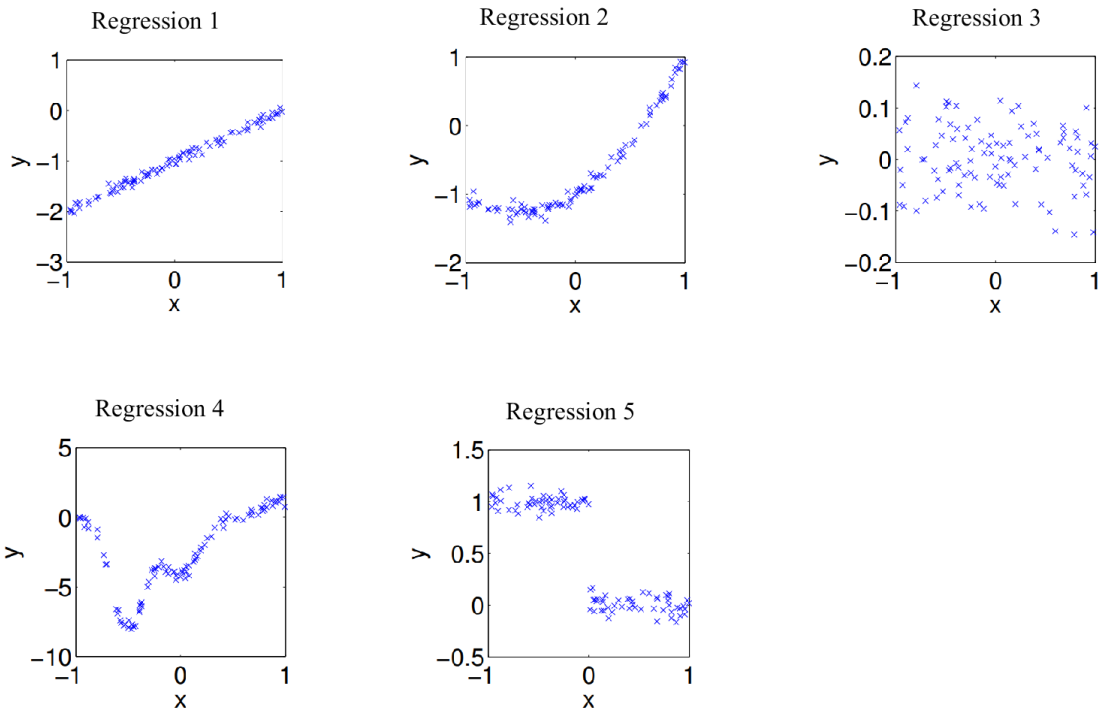

Figure 1: Regression problems

   (a) For the generic case where $y \in \mathbb{R}$ is the output and $\boldsymbol{x} \in \mathbb{R}^n$ is the input vector, what is the expression of the estimator for linear regression? **(2 pt)**

   (b) What are the hyperparameters of ridge regression and k-nearest neighbors? **(2 pt)**

   (c) For each regression problem above, what would be a good model(s) to choose? Discuss your answers. **(4 pt)**

   (d) Which models would lead to overfitting in the "Regression 1" problem? **(1 pt)**

   (e) Explain what can be done to reduce overfitting. **(1 pt)**

6. We consider the classification problem of assigning a new input $\boldsymbol{x} \in \mathbb{R}^2$ to a class $y \in \{0, 1\}$.

   We consider the following models:

   i Logistic regression, where $g(p(y = 0|\mathbf{x} = \boldsymbol{x})) = \beta_0 + \boldsymbol{\beta}^\mathsf{T}\boldsymbol{x}$

   ii Linear discriminant analysis

   iii Logistic regression with 2-nd order polynomials, i.e., we consider and expanded variable set including quadratic terms and then apply logistic regression

   iv k-nearest neighbor classification

where circles correspond to class 0 and crosses to class 1.

We consider the following classification problems.

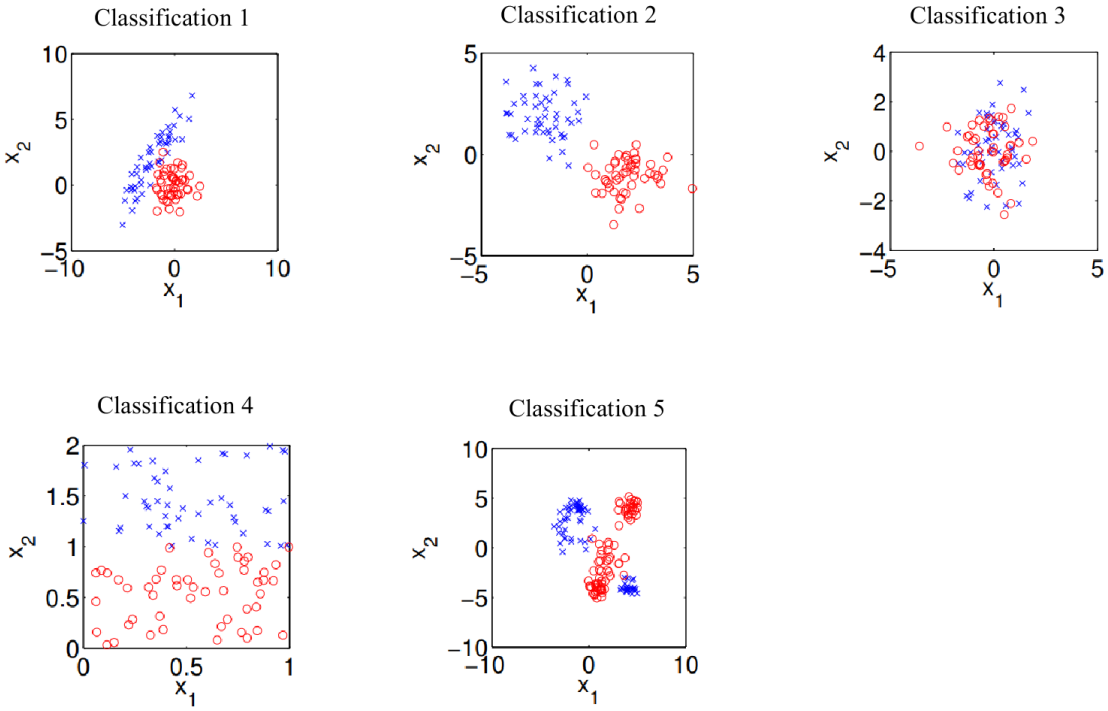

Figure 2: Classification problems

(a) Write the optimization problem that logistic regression solves. How is it solved? **(1 pt)**

(b) What is the main assumption on the data distribution made by linear discriminant analysis? **(1 pt)**

(c) For each of the five classification problems above, what would be the good model(s) to choose? **(2 pt)**

7. Let us consider the case of learning a linear function from data that are actually generated by a quadratic model. Assume that the joint distribution $p(x, y)$ is (implicitly) defined by

$$y = \tilde{f}(x), \qquad \tilde{f}(x) = x^2, \qquad x \sim U[-1, 1]\,,$$

from which one randomly observes $N = 2$ independent data points. Those two data point become our training data $\mathcal{D} = \{(x_1, y_1), (x_2, y_2)\}$. (The notation $x \sim U[-1, 1]$ means x is uniformly distributed between -1 and 1).

To simplify calculations, consider a problem with no noise; the only randomness in the problem is from which training points x_1 and x_2 we happen to learn about $\tilde{f}(x)$ by observing y_1 and y_2. From the training data with two training data points, learn a linear regression model,

$$y = f(x) = \beta_0 + \beta_1 x$$

(a) What is the closed-form solution for $f(x; \mathcal{D})$, as a function of the inputs in the training data x_1, x_2? **(2 pt)**

(b) Show that the average trained model $\mathbb{E}_{\mathcal{D}}[f(x; \mathcal{D})]$ is equal to 0. **(2 pt)** Hint:

$$\mathbb{E}_{\mathcal{D}}[f(x; \mathcal{D})] = \iint f(x; x_1, x_2) p(x_1, x_2) dx_1 dx_2$$

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

Formulas

You might find the following formulas helpful.

- **Theorem 1 (Law of Total Probability)** *Let $\mathcal{E}_1, \ldots, \mathcal{E}_n$ be a **partition** of Ω, i.e., a collection of sets satisfying $\mathcal{E}_i \cap \mathcal{E}_j = \emptyset$ when $i \neq j$, and $\bigcup_{i=1}^{n} \mathcal{E}_i = \Omega$. Assume also that $\mathbb{P}[\mathcal{E}_i] > 0$ for all i. Then the probability of an event $\mathcal{A}$ can be written as*

$$\mathbb{P}[\mathcal{A}] = \sum_{i=1}^{n} \mathbb{P}[\mathcal{A} \,|\, \mathcal{E}_i]\, \mathbb{P}[\mathcal{E}_i]. \tag{1}$$

  In the case of a continuous RV X that admits a density function f_X, the law of total probability states

$$\mathbb{P}[\mathcal{A}] = \int_{-\infty}^{\infty} \mathbb{P}[\mathcal{A} \,|\, X = x] f_X(x)\, dx \tag{2}$$

- **Theorem 2** *For any two RV's Y, Z*

$$\mathbb{E}[\mathbb{E}[Y \,|\, Z]] = \mathbb{E}[Y]\,. \tag{3}$$

- **Theorem 3 (Chebyshev's Inequality)** *For every $a > 0$*

$$\mathbb{P}[|X - \mathbb{E}[X]| \geq a] \leq \frac{\mathbb{V}\text{ar}[X]}{a^2} Fort \tag{4}$$

- **Theorem 4** *Consider an irreducible, aperiodic finite Markov chain with transition probabilities P_{ij} and stationary probabilities π_i. Then for every partition of the state space into mutually exclusive subsets $\mathcal{S}$ and $\mathcal{S}'$, we have that*

$$\sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}'} \pi_i P_{ij} = \sum_{j \in \mathcal{S}'} \sum_{i \in \mathcal{S}} \pi_j P_{ji}. \tag{5}$$

- 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

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}. \tag{6}$$

Part II

It has been assigned as take-home exam on Monday, October 25. 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 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. The solution should be handed in as a .pynb file.

The **deadline for handing in the solution to the assignment is Friday, November 05, 2021.** You hand it in by uploading the solution file to "Take-Home Exam" in Canvas via "Home–>Exam—>Take-Home Exam". This is an individual exam. To help and guide you in the analysis, the exam is divided in to 2 questions, described in details below. Each contains several subparts that are graded individually.

1. In this question, use the *Stock* data set, which contains daily percentage returns for the stock index from 1990 until 2010. The variable $Year$ represents the year that the observation was recorded, $Volume$ represents the Volume of shares traded (number of daily shares traded in billions), $Today$ denotes the Percentage return for today and *Direction* represents a factor with levels Down and Up indicating whether the market had a positive or negative return on a given day. Moreover, $Lagx$ denotes the percentage return for x days before, if x is the number of days. For example $Lag1$ denotes percentage return of the previous day and $Lag2$ denotes percentage return for 2 days previous.

   (a) Observe numerical and graphical summaries of the *Stock* data set. You will need to test the descriptive statistics, pairwise correlation and pairwise scatter plots. Are there any patterns? Comment on the results. **(1.5 pts)**

   (b) Perform logistic regression using the five *Lag* variables plus $Volume$ as predictors and *Direction* as the response. Are any of these predictors statistically significant? If yes, which ones? Hint: You may use *summary*() function to get the results. **(1 pt)**

   (c) State the overall fraction of correct predictions. **(0.5 pts)**

   (d) Instead of using the full data set, now use a training data period from 1990 till 2008. Thereby, for this new training data period, use only $Lag2$ as the predictor and fit the logistic regression model. Calculate the overall fraction of correct predictions for the data from 2009 till 2010. **(2 pts)**

   (e) Use Linear Discriminant Analysis and repeat (d). **(0.5 pts)**

   (f) Use Quadratic Discriminant Analysis and repeat (d). **(0.5 pts)**

   (g) Use KNN with $K = 1$ and repeat (d). **(0.5 pts)**

   (h) Which of the above methods give the best results on this data? Comment on the results. **(0.5 pts)**

   (i) Write a function to determine the overall fraction of correct predictions for a range of values of K in the KNN classifier. **(1 pt)**

2. In this question, let us use the Boston data set and try to predict the per capita crime rate.

```
from sklearn.datasets import load_boston
boston = load_boston()
```

   (a) Implement the lasso, ridge regression and best subset selection. Present and discuss the results. **(3 pts)**

   (b) Suggest model(s) that performs well on the dataset, with justification. Evaluate the model performance using crossvalidation, as opposed to using training error. **(3 pts)**

   (c) Does the model you have chosen involve all of the features in the dataset? Justify why/why not. **(1 pt)**