Instructor Yusuf Yaslan
Assistant Kıymet Kaya

- **Please do your homework on your own.** You are encouraged to discuss the questions with your class mates, but the code and the homework you submitted must be your own work. Cheating is highly discouraged for it could mean a zero or negative grade from the homework.
- **Late submissions will not be accepted**. Please do not email us your late submissions.
- Unless we indicate otherwise, **do not** use libraries for machine learning methods. When in doubt, or if you have a question related to the homework, reach us via email.
- Submissions are expected to include a pdf file prepared with **Latex** that contains the solutions and a Jupyter Notebook file with Python for the coding-related questions.

Please read the instructions below for coding-related questions

- Install a Conda environment if you do not have it already.
- Install Jupyter Notebook.
- In this homework, you are expected to use matplotlib and numpy.
- Questions or sections of a question that are marked with red color (e.g. b), 5)) should be done in Jupyter Notebook.
- Do not forget to format your code and leave comments for non-trivial sections.

Make sure that you read chapter 1-3 and Appendix A from the textbook.

Question 1

Given the table of joint probabilities between three discrete random variables X, Y, and Z.

P(X,Y,Z)	X	0			1		
	Y	-1	0	1	-1	0	1
Z	1	0.08	0.08	0.04	0.01	0.01	0.08
	2	0.12	0.12	0.06	0.12	0.12	0.16

a) Evaluate the conditional probability mass function $P(X|Z{=}1)$.

b) Are random variables X and Z independent? Show your solution.

c) Given that Y is known, are X and Z conditionally independent? Show your solution.

d) Evaluate the conditional expectation of the following function $g(X, Y|Z) = X \times Y + 1|Z$. (**Note that,** you need to evaluate the expectation for both values of Z.)

Question 2

In an online shopping platform, customers purchase a particular product with a probability of P which is a random variable with the PDF as shown below

$$f_P(p) = \begin{cases} 0.4e^x, & p \in [0, \log(7/2)] \\ 0, & \text{otherwise} \end{cases}$$

Given that, multiple customers visited the product page successively and purchased the product according to the PDF independently.

(a) Find the probability that a page visit results in product purchase.

(b) Given that a page visit resulted in product purchase, find the conditional PDF of P.

(c) Given that the page visit resulted in product purchase, find the conditional probability of product purchase in the next page visit from the same customer.

Question 3

Prove the followings where X and Y are random variables with joint distribution $p(x, y)$.

(a)

$$\mathbb{E}[X] = E_Y\left[\mathbb{E}_X[X \mid Y]\right]$$

(b)

$$\text{var}[X] = \mathbb{E}_Y\left[\text{var}_X[X \mid Y]\right] + \text{var}_Y\left[E_X[X \mid Y]\right]$$

Note that the symbol $\mathbb{E}_X[X \mid Y]$ denotes the expectation of X under the density function $p(x \mid y)$, same is also true for the conditional variance. Show your solutions clearly.

Question 4

A recent survey on a sample from the population shows that the temperature of the healthy subjects distributed according to a normal distribution $\mathcal{N}(36.5, 1)$ while the temperature of the subjects with a particular virus is distributed according to $\mathcal{N}(39.2, 1)$. In the sample, one-tenth of the subjects have the virus. Given the prior information;

a) Write down the two class discrimination function $g(t)$ where t denotes temperature.

b) What is the decision threshold (in terms of temperature) in a two-class, two-action decision problem if the cost of misclassifying the sick subjects is:

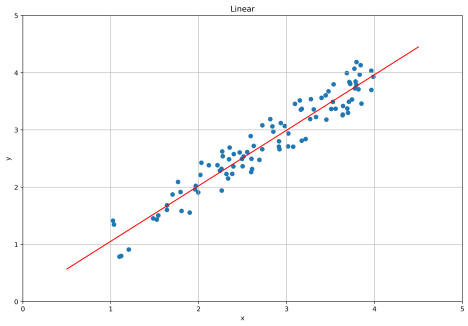
- the same cost as the cost of misclassifying the healthy ones.
- three times the cost of misclassifying the healthy ones.
- 18 times the cost of misclassifying the healthy ones.

Explain the trend in the threshold as we increase the cost of misclassifying the sick subjects.

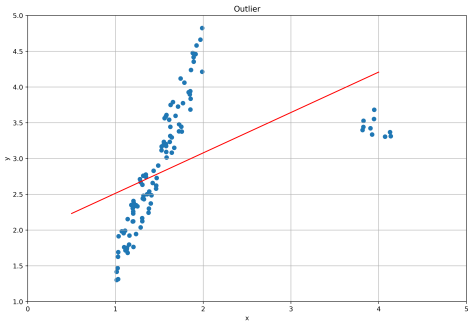
Risk	Truth	
Actions	Sick	Healthy
α_1: Sick	0	$\lambda_{1,2}$
α_2: Healthy	$\lambda_{2,1}$	0

c) Assume that the cost of misclassifying the sick subjects is 7 times higher than misclassifying the healthy subjects. What if we introduce another action, being indecisive, with the cost equal to the one-third of the cost of misclassifying healthy subjects. Give the temperature range for all three actions. (Example: $\alpha_i \rightarrow (-\infty, 37.96)$)

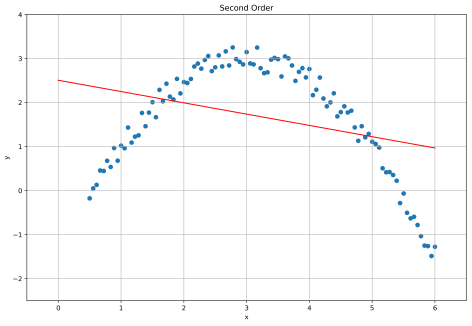
Risk	Truth	
Actions	Sick	Healthy
α_1: Sick	0	$\lambda_{1,2}$
α_2: Healthy	$\lambda_{2,1}$	0
α_3: Indecisive	λ_3	λ_3



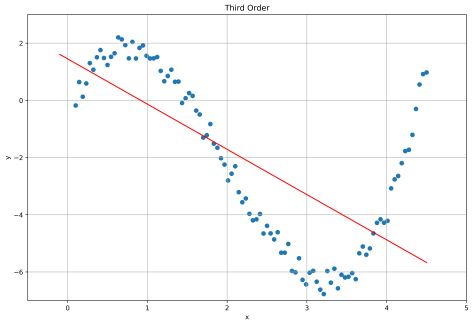
(a) Linear data (linear.csv)



(b) Linear data with outlier (outlier.csv)



(c) Second degree polynomial data (second_order.csv)



(d) Third degree polynomial data (third_order.csv)

Figure 1: Linear regression on the data given within the homework folder

Question 5

In the homework files, you can find four csv files that include example data for the parts b and c. Provide the result of part a and c, and the visualizations in b in the submitted PDF. You can find the visualization of the data in the next page in figure 1.

a) Derive the expected square error in terms of bias and variance. Estimate expected square error, bias, and variance for 1st and 2nd degree polynomial models.

b) Given the following data points (you can find the csv files in the homework directory), compute the least-squares regression line and plot it with a red color on top of the data points for all the cases. (Similar to figure 1 above)

c) Compute the least-squares model for a polynomial of degree up to (including) 4 and compare the expected square error, bias and variance for all these cases.