Assignment 2

DAIICT Winter 2021

Data Analysis and Visualization
CS306

Questions	Marks
Answer all 5 questions	95 marks
	95 Total

Instructions

- This assignment contributes towards 20% of your total marks for grading. Please submit the answers to the assignment by 18th of April 2021.
- If you took or gave help to others then please do mention in you assignment submission. Both handwritten and typed answers are acceptable, however if you used codes to solve the problems, then do submit your computer codes. Otherwise we will consider you have copied answers from somehwere
- Please do mention your reference materials.
- We have kept 5 marks for presentation and communication. Take care to properly present your answers. Scanning quality is machine and human readable and scan orders are correct.

Least squares

Question 1

Consider the following two dimensional data matrix with 8 samples.

x_i	0	1	3	3	4	5	5	6
y_i	2	3	3	4	5	6	7	7

(a) Carry out ordinary least squares to fit a line $y = \beta_0 + \beta_1 x$ to the data. [3 marks]

(b) Normalise the data to be zero mean and one standard deviation. Does normalisation affect the result of ordinary least square? Why data normalisation is considered a good practice in general? [5 marks]

(c) What is an outlier? Construct an example outlier for the above data. Insert your outlier to the data matrix. [3 marks]

(d) Write a pseudo code of RANdom Sample Consensus (RANSAC) algorithm to fit a line to the above data. [5 marks]

(e) Clearly illustrate how you will identify the data outlier and what will be done after rejecting the outlier. [4 marks]

(f) Carry out ordinary least squares to fit a parabola to the data $y = \beta_0 + \beta_1 x + \beta_2 x^2$ above. You may or may not include the outlier you have introduced. Choose the strategy you are most comfortable with. [5 marks]

[Total for Question 1: 25 marks]

SVD and eigen decomposition

Question 2

Following is a small data matrix of five dimensional features and five samples :

$$\mathbf{X} = \begin{bmatrix} 12 & -4 & 0 & 0 & 0 \\ 4 & 12 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}$$

(a) Assuming that the data matrix has been built under the general format of a data matrix. State the features of the three samples. [3 marks]

(b) Compute the feature covariance matrix $\mathbf{\Sigma}$ for the data matrix $\mathbf{X}$ [4 marks]

(c) Carry out eigen decomposition of the covariance matrix $\mathbf{\Sigma}$. $\mathbf{\Sigma} = \sum_{i=1}^{n} \lambda_i U_i U_i^t$ and state its eigen values and corresponding eigenvetors. Do so in the correct order. [5 marks]

(d) Compute the singular value decomposition (SVD) of the data matrix $\mathbf{X}$. Precisely state the singular values and singular vectors, also state the U matrix. [6 marks]

(e) Comment on the difference between (singular values and eigen values) and (singular vectors and eigen vectors) respectively. [2 marks]

[Total for Question 2: 20 marks]

[Total for Question 2: 20 marks]

Analysis of Variance (ANOVA)

Question 3

Following is a data for number of products manufactured by three different machines : M1, M2, and M3, for one hour of their operation by 6 different instances/operators.

	M1	M2	M3
1.	47	55	54
2.	53	54	50
3.	49	58	51
4.	50	61	51
5.	48	55	50
6.	46	52	49

(a) Compute the variance amongst the three sample means S_x^2 . [2 marks]

(b) What is the residual variance S_p^2 ? . [2 marks]

(c) Compute the F ratio. What are the degree of freedom for F? . [3 marks]

(d) The p-value for this ANOVA test comes out to be less than 0.01. State the null hypothesis (H_0) and interpret the p-value for acceptance or rejection of the null hypothesis. What is the percentage error to be incurred? . [3 marks]

(e) Carry out two factor ANOVA analysis for the above data when the machines have been operated by six different operators. Give interpretations of the numbers from your analysis. [5 marks]

[Total for Question 3: 15 marks]

Contraint optimization and modeling using linear algebra techniques

Question 4

General equation of a circle is $(x - x_0)^2 + (y - y_0)^2 = r^2$, where (x_0, y_0) is the circle centre and r is the radius. General equation of a conic is $ax^2 + bxy + cy^2 + dx + ey + f = 0$. A circle is a conic with certain constraints.

(a) Compute and state the values of a, b, c, d, e, f in terms of the parameters: circle center, and the radius [4 marks]

(b) Outline the Direct Linear Transform (DLT) algorithm for computing least square estimate of a, b, c, d, e, f for a given geometric 2 dimensional point data. [6 marks]

(c) We are given geometric point data to compute the parameters of a circle which will best fit the data. Construct and state the constraint matrix for this circle fitting problem. [8 marks]

[Total for Question 4: 18 marks]

Maximum Likelihood Estimation (MLE)

Question 5

Poisson distribution is a probability density function used to describe and analyse the probability of various events like how many customers go through the drive-through, how many phone calls to be received in a time period at a call center etc. This modelling information can help a managers to plan for events with staffing and scheduling. Possion distribution is of the form $p(X/\theta) = \frac{e^{-\theta}\ \theta^X}{X!}$. Let X be the discrete random variable that represents the number of events observed over a given time period. Let θ be the expected value (average) of X.

(a) For a sample of n observations $(X_1, X_2, ...X_i, ...X_n)$ calculate the MLE of θ. [10 marks]

(b) Clearly state the likelihood function and log-likelihood function in the above derivation. Why is log-likelihood preferred over likelihood? [4 marks]

(c) Compute specific value of θ when the samples are $[15, 8, 13, 11, 7, 16, 25, 30]$ in hour time period. [3 marks]

[Total for Question 5: 17 marks]

End of assignment