

Midterm Exam
Introduction to ML
2024-2025



Duration: 60 minutes
Dr. Abbas Rammal

Exercise 1: Logistic & Linear regression

You are given a dataset with information on individuals' age groups and whether they have hypertension (Yes or No). The goal is to predict the likelihood of having hypertension based on age. The dataset is as follows:

Age Range	No hypertension	Yes hypertension	Total
20-29	60	5	65
30-39	50	15	65
40-49	35	25	60
50-59	20	30	50
60-69	10	25	35
70-79	5	20	25
80-89	2	10	12
Total	182	130	312

1. For logistic regression, we need to transform the categorical Age ranges into a continuous variable. One common approach is to use the midpoint of each BMI range as a proxy for the BMI value. Convert BMI Range to a continuous variable.
 2. Fit the model using the age midpoints as the predictor and the proportion of hypertension cases as the outcome.
 3. Predict Probability for an Individual of Age 55.
 4. Given that the sum of squared errors (SSE) = 0.03. Estimate the standard errors of β_0 and β_1 .
 5. Calculate the 95% confidence intervals for the coefficients of the regression model (T-distribution value $t_{6, 0.025} = 2.52$).
 6. Test $H_0: \beta_1 = 0$ versus $H_1: \beta_1 \neq 0$ using the analysis of variance procedure with $\alpha = 0.05$ and the critical value $F_{0.025, 1, 6} = 9.78$.
 7. Calculate the coefficient of determination.
-

Exercise 2: Clustering

Consider the following 6 data points in a 2-dimensional feature space:

$$x_1 = (0, 0); x_2 = (0, 1); x_3 = (-1, 2); x_4 = (2, 0); x_5 = (3, 0); x_6 = (4, -1)$$

Let's calculate the squared distances between each pair of points using the Euclidean distance formula:

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	1	5	4	9	17

x_2	1	0	2	5	10	20
x_3	5	2	0	12	20	34
x_4	4	5	12	0	1	5
x_5	9	10	20	1	0	2
x_6	17	20	34	5	2	0

- ✓ a) Perform K-means clustering on this dataset. Use the first and last data points as initial centers ($K = 2$). Given the final parameters, which cluster would $x^* = (1, 1)$ belong to?
 - ✓ b) Perform agglomerative Hierarchical Clustering using single linkage as the cluster distance measure. Draw the associated tree.
 - ✓ c) Perform agglomerative Hierarchical Clustering using complete linkage as the cluster distance measure. Draw the associated tree.
 - d) Use the Nearest Neighbor clustering algorithm and Euclidean distance to cluster this dataset. Suppose that the threshold t is 4.
-