Midterm Exam
Introduction to ML
2024-2025

Duration: 75 minutes
Dr. Abbas Rammal

Exercise 1: Clustering

Consider the following 10 data points in a 2-dimensional feature space:

$$a(1,7);\ b(2,7);\ c(6,6);\ d(3,5);\ e(4,5);\ f(3,4);\ g(7,3);\ h(1,2);\ i(6,2);\ j(3,1)$$

Your task is the execution of the K-Means algorithm with k = 3 based on these data points.
Let's calculate the distances between each pair of points using the Euclidean distance formula:

	a	b	c	d	e	f	g	h	i	j
a	0	1	6.08	2.23	3.16	3.61	7.07	5	5.83	6.08
b	1	0	5.1	1.41	2.24	2.83	6.71	5	5.66	6.08
c	6.08	5.1	0	3.16	2.24	2.83	1.41	4.12	1	3.16
d	2.23	1.41	3.16	0	1	1.41	5	3.16	3.61	4.47
e	3.16	2.24	2.24	1	0	1	4.47	3.61	4.47	5
f	3.61	2.83	2.83	1.41	1	0	4.12	3	4.24	5
g	7.07	6.71	1.41	5	4.47	4.12	0	4.47	1	2.83
h	5	5	4.12	3.16	3.61	3	4.47	0	5	5.83
i	5.83	5.66	1	3.61	4.47	4.24	1	5	0	1.41
j	6.08	6.08	3.16	4.47	5	5	2.83	5.83	1.41	0

This matrix represents the pairwise distances between each pair of points in the given set, calculated using the Euclidean distance formula.

1. Suppose you are initializing K-means method, that is, you initialize the cluster centers to K randomly chosen data points. Let's assume that points $a(1,7)$, c(6,6) and g(7,3) were chosen. Perform one iteration of the K-means algorithm and report the coordinates of the resulting centroids.

At the end of this epoch show:

a) The new clusters (i.e. the examples belonging to each cluster)
b) The centers of the new clusters
c) How many more iterations are needed to converge?

2. Use single and complete link agglomerative clustering to group the data described by the previous distance matrix. Show the dendrograms.

Exercise 2: Naïve Bayes Classifier

Consider the following dataset, where we aim to predict whether someone will purchase a product based on two features: Age Group (Young, Middle-Aged, Old) and Income Level (Low, Medium, High).

Instance	Age Group	Income Level	Purchase
1	Young	Low	No
2	Young	Low	No
3	Young	Medium	Yes
4	Middle-Aged	High	Yes
5	Middle-Aged	High	Yes
6	Old	Low	No
7	Old	Medium	Yes
8	Middle-Aged	Medium	Yes

1. Using the Naive Bayes classification method, predict whether someone will purchase the product (Yes or No) based on the following samples:
   - S1: (Young, Medium)
   - S2: (Old, Medium)
   - S3: (Middle-Aged, Medium)
   - S4: (Old, Low)
   - S5: (Young, Low)
   - S6: (Old, High)

   Show all computation steps.
2. The actual class labels for the unknown samples are as follows:
   - S1: Yes
   - S2: Yes
   - S3: No
   - S4: No
   - S5: No
   - S6: No

   a) Construct the confusion matrix to compare the predicted class labels with the actual class labels for all samples.

   b) Calculate various evaluation metrics such as accuracy, precision, recall, and F1-score to assess the performance of the classifier.
3. To construct a Receiver Operating Characteristic (ROC) curve, we need to calculate the True Positive Rate (TPR) and False Positive Rate (FPR) at different cutoff points.

The cutoff point represents the threshold used to classify samples as positive or negative. Construct the ROC curve on cutoff point = 0.3 and 0.7.

Instance	Actual	Probability Yes	Probability No
S1	Yes	0.80	0.20
S2	No	0.25	0.75
S3	Yes	0.40	0.60
S4	No	0.65	0.35
S5	Yes	0.70	0.30
S6	No	0.50	0.50