Final Exam
Introduction to ML

Duration; 120 minutes
Dr. Abbas Rammal

Problem 1: Evaluation of classifiers

1. Given a data set $D = \{o_1, \dots, o_n\}$ with known class labels $C(o_i) \in \mathrm{C} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$ of the objects. In order to evaluate the quality of a classifier K, each object $o_i \in \mathrm{D}$ is additionally classified using K, yielding class label $K(o_i)$. The results are given in the table below.

i	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$C(o_i)$	A	B	A	C	C	B	A	A	A	B	B	C	C	C	B
$K(o_i)$	A	A	C	C	B	B	A	A	A	C	A	A	C	C	B

a) Setup the confusion matrix.
b) Compute the accuracy / classification error.
c) For each class $i \in \mathrm{C}$ compute precision and recall.
d) Compute the F1-measure for all classes.

2. **Explain how Leave One Out Cross-Validation is implemented.**
3. **What are the advantages of K-Fold cross validation relative to the validation set approach?**

Problem 2: Find the class with the K Nearest Neighbors

Suppose that we have a classification problem which consists in determining the class of membership of new instances X_i. The range of possible class values is 1, 2, 3.

According to the following knowledge base, determine by hand the class of instance X6, whose values for numeric attributes A1 to A5 are < 3, 12, 4, 7, 8 >, using the algorithm k-nearest neighbors (K-NN) with K = 1 then K = 3. Show all calculations.

Instances	A1	A2	A3	A4	A5	Classes
X1	3	5	4	6	1	1
X2	4	6	10	3	2	2
X3	8	3	4	2	6	3
X4	2	1	4	3	6	3
X5	2	5	1	4	8	2

Problem 3: **Simple Linear Regression**

The data below show the sugar content of a fruit (SUGAR) for different numbers of days after picking (DAYS). Let x be DAYS and y be SUGAR.

Days	0	1	3	4	5	6	7	8
Sugar	7.9	12.0	9.5	11.3	11.8	11.3	4.2	0.4

The average, standard deviation and sum of squared of x and y are

$\bar{x} = 4.25 \quad \bar{y} = 8.55 \quad S_{xy} = -45.6 \quad S_{xx} = 55.5 \quad S_{yy} = SST = 124.26$

a) Draw a scatter diagram of the data. Does a simple linear regression model seem appropriate here?
b) Fit the simple linear regression model using the method of least squares.
c) Estimate the standard errors of β_0 and β_1.
d) Test H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ using the analysis of variance procedure with $\alpha = 0.05$ and $t_{6,\,0.025} = 2.4469$.
e) Find a 95% prediction interval on sugar content when the numbers of days equal two.

Problem 4: K-means clustering

You are to cluster eight points: $x_1 = (2, 10)$, $x_2 = (2, 5)$, $x_3 = (8, 4)$, $x_4 = (5, 8)$, $x_5 = (7, 5)$, $x_6 = (6, 4)$, $x_7 = (1, 2)$ and $x_8 = (4, 9)$. Suppose, you assigned x_1, x_4 and x_7 as initial cluster centers for K-means clustering (k = 3). The distance matrix based on the Manhattan distance is given in Table 1.

Table 1: Distance matrix for training dataset

Instances	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
x_1	0	5	12	5	10	10	9	3
x_2		0	7	6	5	5	4	6
x_3			0	7	2	2	9	9
x_4				0	5	5	10	2
x_5					0	2	9	7
x_6						0	7	7
x_7							0	10
x_8								0

a) Show the three clusters at the end of this epoch.
b) Compute the centers of the new clusters.
c) Draw a 10 by 10 space with all the 8 points and show the clusters after the first epoch and the new centroids.
d) How many more iterations are needed to converge? Draw the result for each epoch.