Lebanese University
Faculty of Information
Department of Data Science

الجامعة اللبنانية
UNIVERSITE LIBANAISE

Year: 6
Duration: 90 minutes

Final Exam
Machine Learning

Documents: not allowed
Date: July 07, 2022

Instructors: Abbas Rammal

Exercise 1: K-means clustering

Use the k-means algorithm and Euclidean distance to cluster the following 8 examples into 3 clusters: A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9).
The distance matrix based on the Euclidean distance is given below:

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

Suppose that the initial seeds (centers of each cluster) are A1, A4 and A7. Run the k-means algorithm for 1 epoch only. 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) 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.

Exercise 2: Error-based Learning

You have been hired by the European Space Agency to build a model that predicts the amount of oxygen that an astronaut consumes when performing five minutes of intense physical work. The descriptive features for the model will be the age of the astronaut and their average heart rate throughout the work. The regression model is

$$\text{OXYCON} = w[0] + w[1] \times \text{AGE} + w[2] \times \text{HEARTRATE}$$

The table below shows a historical dataset that has been collected for this task.

ID	OXYCON	AGE	HEART RATE	ID	OXYCON	AGE	HEART RATE
1	37.99	41	138	7	44.72	43	158
2	47.34	42	153	8	36.42	46	143
3	44.38	37	151	9	31.21	37	138
4	28.17	46	133	10	54.85	38	158
5	27.07	48	126	11	39.84	43	143
6	37.85	44	145	12	30.83	43	138

1010.01

1232.6

1799.57

a. Assuming that the current weights in a multivariate linear regression model are $w[0] = -59.50$, $w[1] = -0.15$, and $w[2] = 0.60$, make a prediction for each training instance using this model.

b. Calculate the sum of squared errors for the set of predictions generated in part (a).

c. Assuming a learning rate of 0.000002, calculate the weights at the next iteration of the gradient descent algorithm.

d. Calculate the sum of squared errors for a set of predictions generated using the new set of weights calculated in part (c).

Exercise 3: Multiple Choice Questions

1. **Discriminant functions are linear combinations of the predictor or independent variables, which will best discriminate between the categories of the criterion or dependent variable (groups).**

   (a) True

   b) False

2. **An examination of differences across groups lies at the heart of the basic concept of _____.**

   (a) Regression analysis

   b) Discriminant analysis

   c) Conjoint analysis

3. **What is true about Machine Learning?**

   a) Machine Learning (ML) is the field of computer science

(b) ML is a type of artificial intelligence that extract patterns out of raw data by using an algorithm or method

c) The focus of ML is to allow computer systems learn from experience without being explicitly programmed or human intervention

4. **A residual is defined as**

a) The difference between the actual Y values and the mean of Y.

(b) The difference between the actual Y values and the predicted Y values.

c) The predicted value of Y for the average X value.

d) The square root of the slope

5. **In regression analysis, the variable that is being predicted is:**

a) the independent variable

(b) the dependent variable

c) usually denoted by x

6. **What is Machine learning?**

(a) The autonomous acquisition of knowledge through the use of computer programs

b) The autonomous acquisition of knowledge through the use of manual programs

c) The selective acquisition of knowledge through the use of computer programs

7. **Machine learning algorithms build a model based on sample data, known as _____**

a) Testing Data

b) Dummy Data

(c) Training Data

8. **The correlation coefficient is used to determine:**

a) A specific value of the y-variable given a specific value of the x-variable

b) A specific value of the x-variable given a specific value of the y-variable

(c) The strength of the relationship between the x and y variables