CIS 735 Machine Learning for Security
Spring 2021
Syracuse University

Homework

Due: March 29, 2021 by 5:00 pm

1. Given the following

Variables	**Type**	**XYZ File's template**
Size in bytes	Integer	52 K
Type of document	Categorical	An essay
Possibility of change	Measured on a Likert's scale: Un-likely, Likely, Moderate, Highly Likely, Definitely likely	Likely
File Attribute	Read, Write, Archive	Read only

Describe a procedure and compute a distance measure to classify the following instances:

(a) (54K, Report, Likely, Read)

(b) (5K, E-mail, Unlikely, Write)

2. (ii) For the following vectors X and Y, calculate the indicated similarity or distance measures

   X = (0, 1, 0, 1, 2, 1), Y = (5, 0, 1, 0, 2, 2) Cosine, Correlation, Euclidean, Jacckard.

Is Cosine distance a metric? Illustrate with an example.

3. List at least five features to classify computer files that have been altered (for example size in bytes). Do you think any of these features might be correlated? Explain your reasoning.

4. Often multivariable observations may not be detected as outliers when each variable is considered independently. In such situations outlier detection can only be successfully done when multivariate analysis is performed, and the interactions among different variables examined. With the aid of a diagrammatic illustration, given an example of how an outlier in a bivariate distribution may be confused as an inlier if each dimension of the bivariate distribution is treated as a distinct univariate distribution.

5. Consider the points x^1 (2,10), x^2 (2,5), x^3 (8,14), x^4 (5,8), x^5 (5,5), x^6 (6,7), x^7 (1,2), x^8 (4,9), and x^9 (2,9). Compute the three final clusters using the k-means clustering algorithm. Use Euclidean distance as the similarity measure. Assume x^1, x^4 and x^7 as initial clusters for k-means clustering.

Note: You may have to write code or use a software package such as Weka to do the following two questions.

6. From a time series data the features F_1, F_2, F_3, F_4, F_5 get extracted to classify a user (User). Use Correlation Based Feature Subset (CFS) to extract three most relevant features. Show your work.

F_1	F_2	F_3	F_4	F_5	User
2.3	3.0	3.1	7.0	0.1	1
5.8	4.0	1.2	3.0	0.2	1
5.7	4.4	1.2	2.0	0.4	1
5.4	3.5	1.3	1.5	0.4	1
5.1	3.5	1.4	1.9	0.3	1
5.7	3.8	1.7	2.0	0.3	1
7.0	3.2	4.7	2.0	1.4	2
6.4	3.2	4.5	3.0	1.5	2
6.9	3.1	4.9	3.9	1.5	2
5.5	2.3	4.0	2.3	1.3	2
5.8	3.2	2.9	4.8	2.3	3
6.7	3.3	5.7	5.0	2.5	3
6.7	3.0	5.2	3.3	2.3	3

7. Following data are coming from a Biometric system. It has four sensors. To make the system more cost effective, the designer decides to reduce its sensors to two. Can you suggest, which two sensors can be get rid of and why? **Show your work**. ***(Use all of the dataset given to come to an answer.) Depending on the method you choose, you may use the class information or ignore the class information.***

S_1	S_2	S_3	S_4	Class (Optional)
2.0	3.0	5.0	4.0	0
5.0	4.0	1.0	0.0	1
3.0	4.0	1.0	0.0	1
8.0	3.0	5.0	2.0	1
6.0	3.0	5.0	2.0	0
6.0	4.0	3.0	2.0	0
5.0	3.0	1.0	0.0	0
5.0	5.0	1.0	2.0	1
5.0	3.0	1.0	0.0	1
7.0	3.0	4.0	1.0	1
6.0	6.0	3.0	1.0	2
6.0	3.0	4.0	1.0	2
5.0	2.0	4.0	1.0	2