
Introduction to Machine Learning

Work 1

Clustering and factor analysis exercise

Contents

1	Clustering and factor analysis exercise	2
1.1	Introduction.....	2
1.2	Methodology of the analysis	2
1.3	Work to deliver.....	2

1 Clustering and factor analysis exercise

1.1 Introduction

The aim of the exercise is to analyze different clustering algorithms using several data sets from the UCI repository. To this end, first of all you will implement the algorithms using MatLab.

1.2 Methodology of the analysis

You will analyze the behavior of different clustering algorithms in well-known data sets from the UCI repository. These data sets are defined in .arff format. So, you will be able to analyze them with the Weka environment, too. A guide can be found at <http://w3.msi.vxu.se/users/dna/755/wekaTutorial.pdf>. The Weka is used to analyze if your code in MatLab is correct or not.

This work is divided in four tasks:

1. Make a parser to read the .arff file in MatLab and save the information in a matrix.
2. Implement K-Means algorithm and apply it to the data of the file
3. Implement PCA algorithm. There is a function in MatLab that let you extract the eigenvalues.
4. Extract conclusions by analyzing several data sets.

1.3 Work to deliver

In this work, you will implement and analyze K-Means and PCA algorithms. You may select 4 data sets for your analysis. At the end, you will find a list of the data sets available.

You will use your code in MatLab to extract the results. Additionally, from these results, you will extract conclusions showing graphs of such evaluation and reasoning about the results obtained. For example, some questions that may help you to comment your results:

- Which information can be obtained for each data set using K-Means and PCA?
- Did you find differences among both algorithms? According to the data sets chosen, which algorithm gives you more advice for knowing the underlying information in the data set?
- Can you explain the setup that you have used for each algorithm?
- In the case of the K-Means, which has been the best K value?
- Which are the most informative features for each data set according to the PCA algorithm?

Reason each one of these questions in your evaluation. Additionally, you should explain how to execute your code.

You should deliver a word or pdf document as well as the code in MatLab in campus virtual by 28th October 2012.

Domain	#Cases	#Num.	#Nom.	#Cla.	Dev.Cla.	Maj.Cla.	Min.Cla.	MV
<i>Adult</i>	48,842	6	8	2	26.07%	76.07%	23.93%	0.95%
<i>Audiology</i>	226	-	69	24	6.43%	25.22%	0.44%	2.00%
<i>Autos</i>	205	15	10	6	10.25%	32.68%	1.46%	1.15%
* <i>Balance scale</i>	625	4	-	3	18.03%	46.08%	7.84%	-
* <i>Breast cancer Wisconsin</i>	699	9	-	2	20.28%	70.28%	29.72%	0.25%
* <i>Bupa</i>	345	6	-	2	7.97%	57.97%	42.03%	-
* <i>cmc</i>	1,473	2	7	3	8.26%	42.70%	22.61%	-
<i>Horse-Colic</i>	368	7	15	2	13.04%	63.04%	36.96%	23.80%
* <i>Connect-4</i>	67,557	-	42	3	23.79%	65.83%	9.55%	-
<i>Credit-A</i>	690	6	9	2	5.51%	55.51%	44.49%	0.65%
* <i>Glass</i>	214	9	-	2	12.69%	35.51%	4.21%	-
* <i>TA O-Grid</i>	1,888	2	-	2	0.00%	50.00%	50.00%	-
<i>Heart-C</i>	303	6	7	5	4.46%	54.46%	45.54%	0.17%
<i>Heart-H</i>	294	6	7	5	13.95%	63.95%	36.05%	20.46%
* <i>Heart-Statlog</i>	270	13	-	2	5.56%	55.56%	44.44%	-
<i>Hepatitis</i>	155	6	13	2	29.35%	79.35%	20.65%	6.01%
<i>Hypothyroid</i>	3,772	7	22	4	38.89%	92.29%	0.05%	5.54%
* <i>Ionosphere</i>	351	34	-	2	14.10%	64.10%	35.90%	-
* <i>Iris</i>	150	4	-	3	-	33.33%	33.33%	-
* <i>Kropt</i>	28,056	-	6	18	5.21%	16.23%	0.10%	-
* <i>Kr-vs-kp</i>	3,196	-	36	2	2.22%	52.22%	47.78%	-
<i>Labor</i>	57	8	8	2	14.91%	64.91%	35.09%	55.48%
* <i>Lymph</i>	148	3	15	4	23.47%	54.73%	1.35%	-
<i>Mushroom</i>	8,124	-	22	2	1.80%	51.80%	48.20%	1.38%
* <i>Mx</i>	2,048	-	11	2	0.00%	50.00%	50.00%	-
* <i>Nursery</i>	12,960	-	8	5	15.33%	33.33%	0.02%	-
* <i>Pen-based</i>	10,992	16	-	10	0.40%	10.41%	9.60%	-
* <i>Pima-Diabetes</i>	768	8	-	2	15.10%	65.10%	34.90%	-
* <i>SatImage</i>	6,435	36	-	6	6.19%	23.82%	9.73%	-
* <i>Segment</i>	2,310	19	-	7	0.00%	14.29%	14.29%	-
<i>Sick</i>	3,772	7	22	2	43.88%	93.88%	6.12%	5.54%
* <i>Sonar</i>	208	60	-	2	3.37%	53.37%	46.63%	-
<i>Soybean</i>	683	-	35	19	4.31%	13.47%	1.17%	9.78%
* <i>Splice</i>	3,190	-	60	3	13.12%	51.88%	24.04%	-
* <i>Vehicle</i>	946	18	-	4	0.89%	25.77%	23.52%	-
<i>Vote</i>	435	-	16	2	11.38%	61.38%	38.62%	5.63%
* <i>Vowel</i>	990	10	3	11	0.00%	9.09%	9.09%	-
* <i>Waveform</i>	5,000	40	-	3	0.36%	33.84%	33.06%	-
* <i>Wine</i>	178	13	-	3	5.28%	39.89%	26.97%	-
* <i>Zoo</i>	101	1	16	7	11.82%	40.59%	3.96%	-