Introduction to Machine Learning
Work 2
Case-based reasoning exercise

Contents

1 Case-based reasoning exercise

1.1 Introduction

In this exercise you will learn about instance-based learning, case-based reasoning and feature selection and you will apply this techniques to classification. You will also learn how to use cross-validation for parameter and feature selection. It is assumed that you are familiar with the basic concept of cross-validation. If not, you can read this paper:

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the International Joint Conferences on Articial Intelligence IJCAI-95. 1995.

1.2 Methodology of the analysis

As in the previous work assignment, you will analyze the behavior of the different algorithms in well-known data sets from the UCI repository. In that case, you will also use the class as we are testing supervised learning algorithms. In particular, in this exercise, you will receive the data sets defined in .arff format but divided in ten training and test sets (they are the 10-fold cross-validation sets).

This work is divided in several steps:

1. Improve the parser developed in Work 1 in order to use the class attribute, too. You should write a function in MatLab called `parserArff('filename.arff')` that saves the information from a training or testing file in a `TrainMatrix` or `TestMatrix`.
2. Write a MatLab function that automatically repeats the process described in previous step for the 10-fold cross-validation files.
3. Write a MatLab function `kNN(TrainMatrix, current_instance, K)` that returns the `K` most similar instances/cases from the `TrainMatrix` to the `current_instance`. Use the Minkowski's metric to compute similarity. Remember that this metric contains a parameter `r` whose value can be 1, 2 or 3. When the r= 2, this metric is an Euclidean distance.
4. Write a MatLab function for classifying each instance/case from the `TestMatrix` using the `TrainMatrix` to a classifier called `cbr(…)`. You decide the parameters for this classifier. In this cbr classifier you will use the `kNN(TrainMatrix, current_instance, K)` as a retrieval phase. Write a function for each one of the phases of the CBR. For retaining, you will use the class information to decide if a case is stored or not. When the case has not been solved correctly, the new case will be stored in the case base.
5. Modify the `kNN(TrainMatrix, current_instance, K)` function. In that case, implement a weighted Euclidean function called `weightedKNN(TrainMatrix, current_instance, K)`. The weights should be extracted using a PCA algorithm. In this work, you can use the one that you implemented in Work 1 or you can use the one included in MatLab.

6. Modify the `weightedKNN (TrainMatrix, current_instance, K)` function. In that case, make a selection of those attributes with the highest eigenvalues and remove the remaining ones. The new function will be `selectedKNN(TrainMatrix, current_instance, K)`.
7. Extract conclusions by analyzing two data sets.

The schedule for these steps is as follows:

- Week 1. Steps 1, 2 and 3
- Week 2. Steps 4, 5
- Week 3. Steps 6, 7

1.3 Work to deliver

In this work, you will implement a case-based reasoning and a feature selection algorithm. You may select 2 data sets (a small and a large one) 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 accuracy results. The accuracy measure is the average of correctly classified instances/cases. That is the number of correctly classified instances divided by the total of instances in the test file.

From the accuracy results, you will extract conclusions showing graphs of such evaluation and reasoning about the results obtained.

In your analysis, you will include several considerations.

1. You will analyze the CBR (with no weighting or selection). You will analyze which are the most suitable K parameter for the KNN function and the r value for computing the Minkowski's metric that results in the highest accuracy.
2. Once you have setup the K and the r parameters. You will analyze several configurations of the CBR: the standard CBR, the weighted CBR and the feature selection CBR.

For example, some of questions that it is expected you may answer with your analysis:

- Which is the best value of K for the retrieval phase?
- Which is the best Minkowski's metric configuration?
- Did you find differences among the standard CBR, weighted CBR and feature selection CBR?
- According to the data sets chosen, which algorithm gives you more advice for knowing the underlying information in the data set?
- In the case of the feature selection CBR, how many features where removed?
- Which criterion have you used to decide the features that should be removed?

  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 Racó at UPC by November, 18th, 2012.

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