Middle East Technical University Department of Computer Engineering

CENG 414

Introduction to Data Mining

Fall 2019-2020

THE 2

Murat Ozturk
mozturk@ceng.metu.edu.tr

Due date: November 16, 2019, 23:59

1 Overview

In this THE, you are going to use Weka to do some experiments on various datasets. The aim is to make you familiar with certain machine learning algorithms and Weka. Weka is a collection of machine learning algorithms for data mining tasks. In Weka, the algorithms can either be applied directly to a dataset through Weka desktop application or called from your own Java code. You are expected to use the latest stable release of Weka in this THE.

2 Questions

You are expected to answer 5 questions. The datasets you are going to use in these questions will be provided inside THE2_datasets.zip: You can download the zip file from Odtuclass.

2.1 Handling Missing Values (15 Pts.)

In this question, you are given ”labor.arff” file. In this dataset, missing values for various attributes exist. You are expected to replace missing values. For this purpose, you are going to use ReplaceMissingValues function in Weka. You can use this in Weka using the following steps:

- Open Weka-Explorer.
- While Preprocess tab is active, click open file and select labor.arff.
- In the Filter section, choose Filters → Unsupervised → ReplaceMissingValues.
- Run the filter with the default parameters by clicking Apply.

After applying the ReplaceMissingValues function, the statistics of the dataset are expected to change. Answer the following questions, based on the changes:

1. Which method(s) did ReplaceMissingValues use to replace missing values?(2 pts)

2. When you compare the statistics of all attributes after applying the function with the raw statistics, which statistic(s) have changed?(3 pts)

3. How did the dataset be affected after applying ReplaceMissingValues function in terms of the changes in its statistics ? Discuss briefly for the following attributes only: "duration","standby-pay","wage-increase-third-year","wage-increase-first-year". (4 pts)

4. Is ReplaceMissingValues function suitable to replace missing values? Discuss briefly for the following attributes only: "duration","standby-pay","wage-increase-third-year","wage-increase-first-year". If you think it is not suitable for some attbiute(s), briefly discuss why. (6 pts)

2.2 Discretization (15 Pts.)

In this question, you are given the "diabetes.arff" file.In this dataset, you are expected to apply the discretization technique which is either the equal-width binning or equal-depth (or equal frequency) binning over all the attributes on the selected file using Weka. You can do this in Weka using the following steps:

- Open Weka-Explorer.

- While Preprocess tab is active, click open file and select "diabetes.arff" file .

- In the Filter section, choose Filters → Unsupervised → Discretize.

- Run the filter with the default parameters by clicking Apply.

After applying the Discretize function, the dataset is expected to change. Answer the following questions, based on the changes:

1. Which method did Discretize use to discretize attribute values?(2 pts)

2. When you compare the original distribution of "preg" attribute with the distrbiution after applying the function, explain one of the difference you observed. (3 pts)

3. Assume you are given a hypothetical dataset called "health" and assume all the values of "age" attribute from this dataset are given as follows: 24,15,25,28,4,21,8,26,9,21,34,29. You are expected to apply binning methods on this attribute:

   a. Apply equal-width binning method to discretize the values where bin size is 3. Show the resulting bins(2.5 pts)

   b. Apply equal-depth binning method to discretize the values. Show the resulting bins(2.5 pts)

4. What is the difference between equal-depth binning and equal-width binning method? Which one of the methods do you prefer to work with numerical attributes? (5 pts)

2.3 Feature Reduction (20 Pts.)

In this question, you are given "vehicles_silhouettes.arff" file which originates from UCI Machine Learning Library in the following link: http://archive.ics.uci.edu/ml/datasets/Statlog+%28Vehicle+Silhouettes%29.
You are expected to apply a well-known feature reduction technique, Principal Component Analysis (PCA) over the given dataset. You will use Weka for this purpose. You can do this in Weka using the following steps:

- Open Weka-Explorer.
- While Preprocess tab is active, click open file and select vehicles_silhouettes.arff.
- Click Select Attributes tab. Select Principal Components in Attribute Evaluator section.
- If an alert window appears which warns to select Ranker select method, click Yes.
- Run the classifer with the default parameters by clicking Start.
- On the right hand side, in the Atrribute Selection Output Section, the results of the analysis are given.

Answer the following questions according to the results given in the Attribute Selection Output Section:

1. Using the correlation matrix, which attributes have the highest positive correlation?(3 pts)
2. Using the correlation matrix, which attributes have the highest negative correlation?(3 pts)
3. Below, "Eigenvalue", "Proportion" and "Cumulative" titles as well as the corresponding numerical values are given. What do they mean? Briefly explain.(6 pts)
4. Discuss the results of this analysis. Do you think this dataset is suitable for feature reduction? Justify your answer. (8 pts)

2.4 Multilayer Perceptron (25 Pts.)

In this question, you are given the "vehicles_silhouettes.arff" file that you already used in the previous question. You are expected to apply Multilayer Perceptron (MLP) classifier using Weka. You can do this in Weka using the following steps:

- Open Weka-Explorer.
- While Preprocess tab is active, click open file and select vehicles_silhouettes.arff.
- Click Classify tab. Select Classifiers → Functions→ MultilayerPerceptron.
- By clicking above the MultilayerPerceptron under Classifier section, you can view/change the default parameters of MultilayerPerceptron.
- In the test options section, select and activate Percentage Split which is %66 (hence %66 of the dataset is set for training and the rest is set for testing).
- Run the classifer with the default parameters by clicking Start.

- On the right hand side, in the Classifier Output Section, the results of the analysis are given.

Answer the following questions according to the results given in the Classifier Output Section:

1. How many hidden layers and hidden nodes are created?(2 pts)
2. Did Weka normalize the attributes? What is the effect of normalizing the attributes? (3 pts)
3. What is the benefit of splitting the dataset as training set and test set? Why don't we just train our model with whole data? (3 pts)
4. Which halting strategy did MLP use?(2 pts)
5. What is the detailed accuracy table by class of the run? (5 pts)

Now change the configurations of the MLP by clicking on the name of the classifier. Run the classification task with different training times (100, 500, 1000, 5000,) while keeping other variables same. You are expected to carry out/answer the following:

6. Plot the training time-test accuracy plot. (5 pts)
7. Interpret the accuracy plot: What is the relation between accuracy and training time (epoch count)? What may cause this situation? (5 pts)

2.5 Support Vector Machine (25 Pts.)

In this question, you are given the "vehicles_silhouettes.arff" file that you already used in the previous question. You are expected to apply the Support Vector Machine (SVM) classifier using Weka. You can do this in Weka using the following steps:

- Open Weka-Explorer.
- While Preprocess tab is active, click open file and select vehicles_silhouettes.arff.
- Click Classify tab. Select Classifiers → Functions→ SMO.
- By clicking above the SMO under Classifier section, you can view/change the default parameters of SVM.
- In the test options section, select and activate Percentage Split which is %66 (hence %66 of the dataset is set for training and the rest is set for testing).
- Run the classifer with the default parameters by clicking Start.
- On the right hand side, in the Classifier Output Section, the results of the analysis are given.

Answer the following questions according to the results given in the Classifier Output Section:

1. Report summary and detailed accuracy by class. (3 pts)
2. Explain the **C** parameter of SVM. Change this parameter and run the classifier with various values: Plot and report the effects. (10 pts)
3. Explain the terms maximum margin hyperplane and support vector. (6 pts)
4. When we run SVM in Weka, it uses a kernel function in default. What is a kernel function? What is the benefit to use a kernel function? Explain clearly (6pts).

3 Submission and Regulations

1. For each task create a directory named **q1**, **q2**, ..., **q5**. All of your solutions, comments, plots about a task should be inside the correspondent directory. If your directory structure is *messy*, you will get **penalty**.

2. Zip all task directories and name it as <ID> _ <FullNameSurname> and submit it through odtuclass. For example:
e1234567_MuratOzturk.zip

3. Copying from others is strictly forbidden and is subject to discplinary action.

Note: Any extra effort will be rewarded. **Late submissions will not be accepted.**