Data Mining / Machine Learning Homework

April 18, 2014

Part I
Guided example

In this part of the homework, we will guide you through the process of using the Weka Data Mining software to predict attributes from real data using the Naive Bayes algorithm. The instructions below should guarantee that you obtain the results shown at the end. Feel free to ask questions to the TAs if something is not clear.

1. Download the Social Evolution Dataset from http://realitycommons.media.mit.edu/socialevolution.html

2. Download and install WEKA from http://www.cs.waikato.ac.nz/ml/weka/downloading.html

3. Make sure you read the description of the data and the data breakdown to get a general idea of the data.

4. In this guided example, we will attempt to predict whether two of the subjects are friends or not, based on their mobile behavior (i.e., SMS messages, phone calls, and proximity to one another). Because of the way the data is presented, it is important to preprocess the data so we can use WEKA.

5. The first thing that we need to do is to combine the information that we need in one CSV (comma-separated values) file.

6. For this problem you need to read the Subjects.csv file, the RelationshipsFromSurvey.csv file, the Proximity.csv file, the SMS.csv file, and the Calls.csv file.

7. From the Subjects.csv file, we are interested in the subject id (you can ignore the school year and floor information).

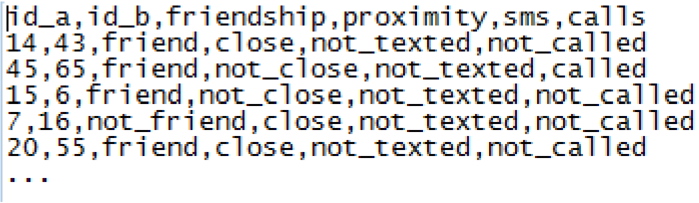
8. From the RelationshipsFromSurvey.csv file, we will consider people as friends if they name their relationship (at least once) as either CloseFriend or SocializeTwicePerWeek.

9. From the Proximity.csv file, we will consider people as having been close to each other if the value of column prob2 is greater than 0 at least once (if the value is missing assume 0)

10. From the SMS.csv file, we will assume they text each other if there was at least one sms message (either sent or received) between themselves.

11. From the Calls.csv file, we will assume they call each other if there is a call longer than 0 minutes.

12. The final csv should look like this:



```
id_a,id_b,friendship,proximity,sms,calls
14,43,friend,close,not_texted,not_called
45,65,friend,not_close,not_texted,called
15,6,friend,not_close,not_texted,not_called
7,16,not_friend,close,not_texted,not_called
20,55,friend,close,not_texted,not_called
...
```

13. Start Weka and open the Explorer view

14. Load the CSV file that you created

15. Select the variables that hold the user ids of the subjects and remove them the dataset.

16. Go to Classify Tab and press the Choose button (under classifier). Select the NaiveBayes under the bayes folder.

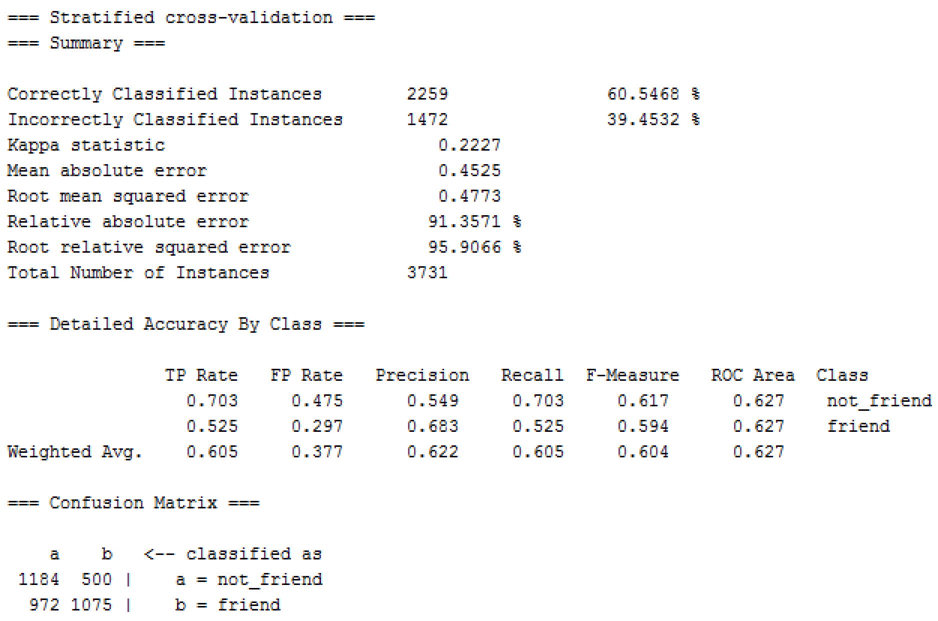
17. Make sure that Crossfold validation is selected (with a value of 10)

18. In the More Options... button select “Output Predictions”

19. Above the Start and Stop Button change the class to friendship.

20. Press Start

21. The result should look like this



```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2259               60.5468 %
Incorrectly Classified Instances      1472               39.4532 %
Kappa statistic                          0.2227
Mean absolute error                      0.4525
Root mean squared error                  0.4773
Relative absolute error                 91.3571 %
Root relative squared error             95.9066 %
Total Number of Instances             3731

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.703     0.475     0.549      0.703     0.617      0.627    not_friend
                 0.525     0.297     0.683      0.525     0.594      0.627    friend
Weighted Avg.    0.605     0.377     0.622      0.605     0.604      0.627

=== Confusion Matrix ===

    a    b   <-- classified as
 1184  500 |    a = not_friend
  972 1075 |    b = friend
```

From the previous screenshot, pay special attention the "Correctly Classified Instances" (this is the accuracy) and to the "Confusion Matrix". This matrix tells you the number of instances that were classified on each class (and how they should have been).

Part II

Guided Problem

For part 2, you will explore the health habits of the participants and explore the different evaluation metrics mentioned in class.

Preprocessing

1. Remove the user_id, current_weight, current_height and surver.month columns from the Health.CSV file.
2. Discretize (i.e., make binary) the salads_per_week column using 4.5 as threshold. This means that if a participant eats less or equal than 4.5 salads a week you will label him as "few_salads" and if he eats more then 4.5, you will label him as "lots_of_salad".
3. Using a similar procedure, discretize the variable veggies_fruits_per_day using a threshold of 3.5, the variable aerobic_per_week using a threshold of 3.5, and the variable sports_per_week using a threhold of 3.5.

4. Combine the values of the variable healthy_diet as follows. If a participant claims to to eat below average, unhealthy, or very unhealthy, label him as “unhealthy”. If a parcitipant claims to eat healthy or very healthy, label him as “healthy”. This will create a three-valued variable with values unhealthy, average, and healthy.

5. Leave the variable current_smoking as is.

Classification

1. Build a Naive Bayes classifier to predict current_smoking. What is the accuracy? Look at the confusion matrix, what does it tell you? Is this a good classifier? Explain your answers and report the accuracy and the confusion matrix.

2. Build a Naive Bayes classifier to predict aerobic_per_week. What is the accuracy? Look at the confusion matrix, what does it tell you? Again, explain your answers and report the accuracy and the confusion matrix.

3. Build a Naive Bayes classifier to predict healthy_diet. What is the accuracy? Look at the confusion matrix, what does it tell?

4. Now build a decision tree for healthy_diet. Under Classifier, press the Choose button and select J48 under the trees folder. Run it and compare the accuracy and confusion matrix with the ones obtained with the Naive Bayes classifier. Which one is better? Report both accuracies and confusion matrices.

5. Now look at the learned decision tree. What does it tell you? Write a short paragraph interpreting the decision tree and its adequacy in learning the healthy_diet variable. Does it make sense? Show the tree.

Part III

Your own design

For this part of the homework, you will choose an interesting variable to predict by yourself.

You are free to choose any variable of the dataset and use any amount of features.

Report all the preprocessing that you do (and why you think it is reasonable) and your results.

To choose your variable try to tell a story, what do your findings tell us?

We will grade your explanation of who is the audience (who cares about predicting the variable that you chose), how and why did you processed the data the way you did, and your final accuracy.