Illinois Institute of Technology

CS422: Introduction to Data Mining
Spring 2008
Faculty: Dr. Nazli Goharian
Teaching Assistant: Saket Mengle

Assignment Name: Pre-processing & Naïve Bayes

Date: Jan 29th
Due Date & Time: Feb 19th before the start of the class: Must be submitted via Blackboard , Digital Drop Box, no later than 3:00 pm (Live section and On-Campus Internet students must ALSO to submit a hard copy at the start of the class or in my mailbox).

Grading: This assignment is 8 points out of the total 35 points allocated for all assignments in the semester. It will be graded on the scale of 100.

General Statement about all assignments in this semester:

You are encouraged to use SimpleDM for this assignment and the remaining assignments in this course. However, if for some reason you prefer not to use SimpleDM, it is fine too. In that case, you need to submit also your reason why you decided not to use SimpleDM. This implies that you are free to choose the programming language for the implementation of your assignments. However, your choice of programming language and using/not using SimpleDM does not effect the due date. Thus, if you are considering not using SimpleDM, you need to be very comfortable in programming, as you need to program the functionalities that are already built into SimpleDM and still being able to submit your work on time. Each student can be asked by the instructor or the TA to show up for a demo for each of the assignments during the semester. The demo cannot be requested by the student but only by the instructor/ TA. If you are not able to answer the questions asked in your demo, you will receive a zero for that project.

Assignment Description:

By now you have downloaded SimpleDM, have gone over the code and all documentations. Now you are ready to do your first assignment that includes some data pre-processing and building Naïve Bayes classifier. As it is a supervised algorithm, you need to use a training data set for building your model. Use the provided data set, *adult.arff* for this assignment. As documented on the Weka web page, an ARFF (Attribute-Relation File Format) file is simply an ASCII text file that describes a sample set. More detailed information can be found at http://www.cs.waikato.ac.nz/~ml/weka/arff.html. This file contains all of the data samples that you will train and test on using 10-fold cross-validation. If you look at the beginning of the file you will see a list of attributes and their types. Your classifier should classify the records based on *"Income"* attribute as the class label for your data set. The distinct values of attribute *Income* are >50K, and <=50K. Thus, a sample X (age, gender,…..) should be classified as >50K income or <=50K income.

The assumption is that the whole data does not fit in the memory. Thus, you need to read 1000 records at a time, perform computations, and release the memory that was used for

those 1000 records. Then read the next 1000 records and so on until the entire training set is processed. This is both for the data processing phase and classification phase.

1. You need to perform some data pre-processing for this assignment. Some of these may have been already implemented. For those you need to write a summary in your design document as to how it was implemented. For the ones that are not implemented you need to implement and write the detail in your design document. For this project, following are the data pre-processing items you need to handle in the same order given below:

   a.Replace missing values: take care of the missing numeric values by replacing each missing value by the mean attribute value of the class that the sample with missing value belongs to. For categorical values, use the mode value (most common value) in each class.

   b.Discretize continuous values (change to categorical): Use Entropy-based discretization. For the termination criterion pick any of the choices discussed in the lecture/book. Look at the data and see if you would think the results of discretization makes sense. You may change your termination criterion based on that. Make sure you explain this in your design document..

2. After you are done with the data pre-processing, then use the naïve Bayes algorithm described in the class and in your textbook for classification. Use 10-fold-cross validation for your experimentation. See the deliverables for the details.

Table 1: Evaluation Results

Models	Micro precision	Micro recall	Micro F1	Macro precision	Macro recall	Macro F1	Accuracy
Model- 1							
Model –2							
Model –3							
Model –4							
Model –5							
Model –6							
Model –7							
Model –8							
Model –9							
Model – 10							
Average Accuracy over all 10 models							

Deliverables:

Summary (2 pts): Summary should contain the following in the exact order as specified:

a. Status of this assignment: Complete or Incomplete. If incomplete state clearly what is incomplete. Failing to report the exact status or giving a false or misleading statement will result a zero for your entire assignment.
b. Time spent on this assignment. Number of hours.
c. Problems encountered. List at least 3 to 4 biggest problems that you encountered while you were working on this assignment and how you solved them.
d. Things you wish you had been told prior to being given the assignment.

Design (8 pts): Specific on the design of your mining engine for this homework.

Coding (20 pts): Include a script or batch file that will automatically execute your program with the default data set and output the requested information (if you are not using SimpleDM). Your code must be very clear, clean, and well commented. Also if you are not using SimpleDM your program must be able to accept any ARFF file as input. Do not hard code anything that will prevent other data files from being used with your system.

Results (60 pts): Your result is based on 10-fold cross validation. This means for each of the 10 runs, you have one result. Include a text output file named *adult.out* that contains all of the runtime information for each fold, this is **mandatory. For each of the 10 runs, provide the specified information in the provided table.**

Analysis (10 pts): You are required to submit your detailed analysis of the results.

How to submit:

For this assignment and all the future assignments, the file you submit to blackboard will be a compressed, archived file. WinZip and variants are fine for Windows users while a tarball is acceptable for those using a Unix compatible system. When uncompressed, the resulting file should be a directory. The name of the directory will be your blackboard user name. The contents of the directory depend on the assignment. For this assignment, include directories for the source code, output, and test data files. If you use SimpleDM then this is already taken care of.

2.When using Blackboard (http://blackboard.iit.edu) to upload your assignments, you must click on the 'Add File' button **and** click the 'Send File' button. The file will not be sent if you just add it.