Department of Computer Science and Engineering
University of California at San Diego

CSE 255
Winter 1998

Project 1

DUE AT THE START OF CLASS, TUESDAY JANUARY 20.

The assignment here is to to do useful probabilistic data mining with credit-scoring data. You must implement the naive Bayes learning algorithm discussed in class. All teams will use the same training set of 800 examples Your score will be based partly on performance on a test set of 200 examples that will be revealed later. Performance will be measured using a cost matrix as discussed in class.

You should document your work in a well-written report of eight to ten pages. The report should have a table of contents and a list of references, with references cited as appropriate in the text of the report. The contents of this report will be discussed in more detail in class.

As an appendix, you should submit a printout of your software, with comments and documentation of professional quality. This documentation should be sufficient for another software engineer to maintain the program. Remember that good documentation is necessary but not sufficient: it cannot alleviate bad engineering.

You should also attach a printout showing how a user interacts with your software and performs the data mining task. No fancy user interface is necessary, but you should devote thought to making the interaction convenient. Consider carefully exactly what inputs should be requested from the user, and exactly what outputs does s/he want to see?

Description of the credit-scoring data

```
Source Information:

Professor Dr. Hans Hofmann
Institut f"ur Statistik und "Okonometrie
Universit"at Hamburg
FB Wirtschaftswissenschaften
Von-Melle-Park 5
2000 Hamburg 13

Number of Attributes: 20 (7 numerical, 13 qualitative)

Description of Attributes:

Attribute 1:  (qualitative)
       Status of existing checking account
               A11 :      ... <    0 DM
       A12 : 0 <= ... <  200 DM
       A13 :      ... >= 200 DM /
     salary deposits for at least 1 year
               A14 : no checking account

Attribute 2:  (numerical)
      Duration in months of the loan
```

Attribute 3: (qualitative)
Credit history
A30 : no loans taken/
all loans paid back on time
A31 : all loans at this bank paid back on time
A32 : existing loans paid back on time till now
A33 : delay in paying off in the past
A34 : overdue account/
other loans existing (not at this bank)

Attribute 4: (qualitative)
Purpose
A40 : car (new)
A41 : car (used)
A42 : furniture/equipment
A43 : radio/television
A44 : domestic appliances
A45 : repairs
A46 : education
A47 : (vacation - does not exist?)
A48 : retraining
A49 : business
A410 : others

Attribute 5: (numerical)
Loan amount

Attribute 6: (qualitative)
Savings account/bonds
A61 : ... < 100 DM
A62 : 100 <= ... < 500 DM
A63 : 500 <= ... < 1000 DM
A64 : .. >= 1000 DM
A65 : unknown/ no savings account

Attribute 7: (qualitative)
Length of time in present job
A71 : unemployed
A72 : ... < 1 year
A73 : 1 <= ... < 4 years
A74 : 4 <= ... < 7 years
A75 : .. >= 7 years

Attribute 8: (numerical)
Repayment as a percentage of disposable income

Attribute 9: (qualitative)
Marital status and sex
A91 : male : divorced/separated
A92 : female : divorced/separated/married
A93 : male : single
A94 : male : married/widowed
A95 : female : single

Attribute 10: (qualitative)
Other debtors / guarantors

A101 : none
A102 : co-applicant
A103 : guarantor

Attribute 11: (numerical)
Present residence since

Attribute 12: (qualitative)
Assets
A121 : real estate
A122 : if not A121 : building society savings agreement/ life insurance
A123 : if not A121/A122 : car or other, not in attribute 6
A124 : unknown / no assets

Attribute 13: (numerical)
Age in years

Attribute 14: (qualitative)
Other installment plans
A141 : bank
A142 : stores
A143 : none

Attribute 15: (qualitative)
Housing
A151 : rent
A152 : own
A153 : for free

Attribute 16: (numerical)
Number of existing loans at this bank

Attribute 17: (qualitative)
Job
A171 : unemployed/ unskilled - non-resident foreigner
A172 : unskilled - resident
A173 : skilled employee / official
A174 : management/ self-employed/ highly qualified employee/ officer

Attribute 18: (numerical)
Number of people liable to provide financial support for

Attribute 19: (qualitative)
Telephone
A191 : none
A192 : yes, registered under the customer’s name

Attribute 20: (qualitative)
foreign worker
A201 : yes
A202 : no