CoIL 2000 Submission

Petri Kontkanen
Complex Systems Computation Group (CoSCo)
University of Helsinki, Finland
e-mail: pkontkan@cs.helsinki.fi

PREDICTION TASK

The result was obtained by using a mixture of 20 "pruned" Naive Bayes classifiers, where the attribute set was in each case selected by using a greedy search algorithm with the hitrate in a validation set (separated from the training data) used as the score to be optimized.

DESCRIPTION TASK

The training data was randomly partitioned into two separate data sets 400 times, and each time one of the data sets was used as training data, and the other set was used as validation data. For each split, a greedy feature selection algorithm was run, starting with no attributes, and adding new attributes to a Naive Bayes classifier until no improvement in the validation score (the hit rate of caravan policy owners, measured in the validation set) was observed. Attached is a list of all the attributes, ordered by their "importance score", i.e., by a number expressing how many times (in the 400 runs performed) the attribute in question was selected by the greedy feature selection procedure. Consequently, high importance score means that the attribute plays an important role in the problem of recognizing caravan policy owners. However, please note that the importance scores are not fully independent, and some of the attributes get a low importance score if there exists a highly correlated attribute with a high importance score. An example of this type of a situation is the attribute "Number of car policies" with an importance score of 10.56%: this attribute is obviously important, but as it is strongly dependent with the attribute "Contribution car policies" with importance score 95.33%, it will not be chosen by the greedy feature selection procedure very often (if "Contribution car policies" is chosen, the information contained in "Number of car policies" is already included in the model), which results in a relative low importance score.

From the list we can make the following observations. First of all,it is evident that most of the sociodemographic attributes play a minor role in this prediction task. In particular, the attributes "Private health insurance" and "Home owners" seem to be totally irrelevant. The most important sociodemographic attributes are "Protestant" (with importance score of 14.25%), "High status" (10.81%), "Lower level education" (9.58%) and "Number of houses 1 - 10" (8.84%). Comparing the distribution of the values among the caravan policy owners to the whole population, we can deduce that the probability of caravan policy ownership is increased in areas with a high frequency of protestants, and/or among people with higher (than lower) level education and high status, and possibly with more than 1 house.

Second, it is obvious that the most important attributes are related to other type of policies - top seven of the most important attributes are all policy-related attributes. To summarize the result of analyzing the value distributions among caravan policy owners we can say that the probability of caravan policy ownership is highly increased if the contribution or number of other policies is increased, especially with respect to car, fire, boat, disability and bike policies.

It is also interesting to note that many of the distributions of the most important attributes are highly peaked, meaning that some of the values are quite rare. For example, although there are only 13 boat policy owners in the training set, this attribute turned out to be the third most important attribute in recognizing caravan policy owners.

The importance scores obtained are as follows:

95.33 Contribution car policies
91.15 Contribution fire policies
57.24 Number of boat policies
43.48 Contribution boat policies
23.34 Contribution disability insurance policies
17.93 Contribution bicycle policies
15.23 Number of disability insurance policies
14.25 Protestant
14.25 Contribution third party insurane (agriculture)
10.81 High status
10.56 Number of car policies
10.07 Contribution trailer policies
9.58 Lower level education
9.33 Number of social security insurance policies
9.33 Number of bicycle policies
9.33 Contribution social security insurance policies
8.84 Number of houses 1 - 10
8.59 Customer main type
7.61 Number of property insurance policies
7.12 Contribution third party insurance (firms)
6.87 Other religion
6.87 Farmer
6.63 Number of third party insurane (agriculture)
6.14 Roman catholic
6.14 Income >123.000
5.65 No car
5.40 Contribution tractor policies
5.40 Living together
5.15 No religion
5.15 Contribution life insurances
5.15 1 car
4.91 Number of family accidents insurance policies
4.91 Contribution family accidents insurance policies
4.91 Contribution motorcycle/scooter policies
4.91 Average income
4.66 Contribution property insurance policies
4.42 Purchasing power class
4.17 Contribution private third party insurance
4.17 Avg age
4.17 Social class A
4.17 Customer Subtype
3.93 High level education
3.19 Income 75-122.000
3.19 Avg size household 1 - 6
3.19 Medium level education
2.94 Number of life insurances
2.94 Number of trailer policies
2.94 Contribution lorry policies
2.94 Contribution delivery van policies
2.70 Contribution agricultural machines policies
2.70 Social class D
2.45 Number of private accident insurance policies
2.45 Household with children
2.21 Number of moped policies
2.21 Number of private third party insurance 1 - 12
2.21 Social class B1
2.21 Skilled labourers
1.96 Contribution surfboard policies
1.96 Income < 30.000
1.96 Social class C
1.71 Number of surfboard policies
1.71 Number of tractor policies
1.71 Entrepreneur
1.71 Other relation
1.47 Number of agricultural machines policies
1.47 Number of delivery van policies
1.22 Number of lorry policies
0.98 Contribution moped policies
0.98 2 cars
0.73 Number of motorcycle/scooter policies
0.73 Number of third party insurance (firms)
0.73 Income 45-75.000
0.73 Income 30-45.000
0.73 Middle management
0.49 Number of fire policies
0.49 Rented house
0.49 Unskilled labourers
0.49 Household without children
0.49 Singles
0.49 Married
0.24 Contribution private accident insurance policies
0.24 National Health Service
0.24 Social class B2
0.00 Private health insurance
0.00 Home owners