UNIVERSITY
OF MANNHEIM
School of Business Informatics
and Mathematics

Data Mining

4.1. Who should get a bank credit?

The German credit data set from the UCI data set library (http://archive.ics.uci.edu/ ml/index.html) describes the customers of a bank with respect to whether they should get a bank credit or not. The data set is provided as *credit-g.arff* file in ILIAS.

1. Plot ROC curves for k-NN (different k values), Logistic Regression and Naïve Bayes classification (you can use the given avg_roc function). Which classification approach looks most promising to you?
2. For the two most promising classification approaches, compute the accuracy and confusion matrix in a 10-fold cross-validation setup (use cross_val_predict function). Which level of accuracy do you reach?
3. What do the precision and recall values for the class “bad” customer tell you? Try to improve the situation by increasing the number of “bad” customers in the training set (in the cross-validation!). How do precision and recall change if you apply this procedure?
4. To model a use-case specific evaluation, as observed in the previous example, compute the cost of all misclassifications. Set up your cost matrix by assuming that you will lose 1 unit if you refuse a credit to a good customer, but that you lose 100 units if you give a bad customer a credit. Re-run the experiments from 4.1.2 and evaluate the results.
5. As the creation of training data is mostly a manual task and humans tend to be fallible, training data might include noise. Simulate this behavior by using the *Add Noise* function and change the parameter “percentage” from 0% over 10% to 20%. Is your preferred classification approach still feasible for this situation? How does the performance of the other classifiers evolve?

4.2. Open Competition: Finding rich Americans

The Adult data set from the UCI data set library (http://archive.ics.uci.edu/ml/datasets/Adult) describes 48,842 persons from the 1994 US Census. The data set is provided as *adult.arff* file on the website of this course.

Your task is to find a good classifier for determining whether a person earns over $50.000 a year. Besides of being accurate, your classifier should also have balanced precision and recall.

To evaluate your classifiers, use *train_test_split* validation (test_size=0.2, random_state=42).

In order to find the best classifier, you may experiment with:

1. different algorithms
2. different parameter settings
3. the balance of the two classes in the data set
4. the set of attributes that are used or not used
5. other preprocessing techniques

People are described by the following 14 attributes:

age	continuous
workclass	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt	continuous
education	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
education-num	continuous
marital-status	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
occupation	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
relationship	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
race	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
sex	Female, Male.
capital-gain	continuous
capital-loss	continuous
hours-per-week	continuous
native-country	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

In order to increase your understanding of the data set, you might want to visualize different attributes or attribute combinations.