UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Data and Web Science Group
Prof. Dr. Christian Bizer
B6 – B1.15
68131 Mannheim

Data Mining

5.1. Parameter optimization

In Exercise 4.2 we have used the German credit data set from the UCI data set library (http://archive.ics.uci.edu/ ml/index.html), which 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. (recap) Go back to the results of exercise 4.2.4, in which you have compared Naïve Bayes, k-NN (k=5) and Decision Tree classifiers. In that exercise you
    a. Used the 10-fold cross-validation approach.
    b. Balanced the training set multiplying the “bad customer” examples.
    c. Evaluated the results, setting up your cost matrix to ((0,100)(1,0)) – thus, you assumed you will lose 1 unit if you refuse a credit to a good customer but you lose 100 units if you give a bad customer a credit.

    Rerun your process to get the performance results. What were the default parameters of the *Decision Tree*?

2. Now try to find a more appropriate configuration for the Decision Tree classifier. Use the *GridSearchCV* from scikit-learn. Try the following parameters of the Decision Tree:
    - criterion: ['gini', 'entropy']
    - 'max_depth': [1, 2, 3, 4, 5, None] (What does None mean?)
    - 'min_samples_split': [2,3,4,5]

    You should come up with 48 (2 x 6 x 4) combinations.

    What is the best configuration for the data set and the classification approach?

3. What is the cost of misclassification for this configuration?
4. How does the optimal decision tree differ from the one you have learned in 4.2.4?

5.2. 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.