Data Science and Database Technology

Politecnico di Torino

Data Mining – Practice 4

Objective

Exploit data mining classification algorithms to analyze a real dataset using the **Scikit-Learn** machine learning library on Python.

To solve the lab, you must use the following Python notebook: Lab4
Once you open the file, click on the top of the page for "Connect Apps" and install Google Colaboratory.
If the previous link doesn't work, you can upload the Python notebook linked to this lab (see the content of the zip on the website) directly to Google Colab.

Extended Google Colab user guide: Colab

Dataset

The Users dataset (Users.xls, placed inside the zip file) collects census data about American users of a given company. Users areclassified as "basic" or "premium" according to the services they commonly use. Each dataset record corresponds to a different user. The dataset collects around 32,000 different users, including some
personal user information (e.g., age, sex, workclass) as well as their corresponding class. The class attribute, which willbe used as a class attribute throughout the practice, is reported as the last record attribute.

The complete list of dataset attributes is reported below.

(1) Age
(2) Workclass
(3) FlnWgt
(4) Education record
(5) Education-num
(6) Marital status
(7) Occupation
(8) Relationship
(9) Race
(10) Sex
(11) Capital Gain
(12) Capital loss
(13) Hours per week
(14) Native country
(15) class (class attribute)

Context

Analysts want to predict the class of new users, according to the already classified user characteristics. To this purpose, analysts exploit three different classification algorithms: a decision tree (Decision Tree), a Bayesian classifier (Naïve Bayes), and a distance-based classifier (K-NN). The Users dataset is used to train classifiers and to validate their performance.

Goal

The aim of this practice is to generate and analyze different classification models and validate their performance on the Users dataset using the Scikit-Learn library on Python. Different. To evaluate classification performance, different configuration settings have to be tested and compared with each other. A 10-fold Stratified Cross-Validation process must be used to validate classifier performance. Results achieved by each algorithm should be analyzed in order to analyze the impact of themain input parameters.

Questions

Answer the following questions:

1. Learn a Decision Tree using the whole dataset as training data and the default configuration setting for the algorithm Decision Tree. (a) Which attribute is deemed to be the most discriminative one for classprediction? (b) What is the height of the generated Decision Tree? (c) Find an example of pure partition in the Decision Tree generated.
2. Analyze the impact of the minimal impurity (using the entropy criterion) and maximal depth parameters on the characteristics on the Decision Tree model learnt from the whole dataset (keep the default configuration for all the other parameters).
3. What happens if we change the class label from “Service class” to “Native Country”? Answer again to question (1) in this new scenario.
4. Considering again the service class as the class attribute and performing a 10-fold Stratified Cross-Validation, what is the impact of the minimal impurity and maximal depth parameters on the average accuracy achieved by the Decision Tree? Compare the confusion matrices achieved using different parameter settings by 1) Keeping the default configuration for all the other parameters, 2) Using at least 4 different values (choose wisely!) for each parameter.
5. Considering the K-Nearest Neighbor (K-NN) classifier and performing a 10-fold Stratified Cross-Validation, what is the impact of parameter K on the classifier performance? Compare the confusion matrices achieved using different K parameter values. Perform a 10-fold Stratified Cross-Validation with the Naïve Bayes classifier. Does K-NN perform on average better or worse than the Naïve Bayes classifier on the analyzed data?
6. Analyze the Correlation Matrix to discover pairwise correlations between data attributes. Considering the results achieved, does the Naïve independence assumption hold for the Users dataset?