Data Science and Database Technology

Politecnico di Torino

Homework 2

Objective

Exploit data mining classification algorithms to analyze a real dataset using Python.

To solve the homework, you must use the Python notebook inside the zip file and upload it on Google Colab.

Google Colab user guide: Colab

Dataset

The Breast dataset (Breast.xls, available on the course website) collects medical data about patients who have contracted breast cancer. Each dataset record corresponds to a different patient and consists of a set of patient, treatment, and disease characteristics (e.g., the patient age, the tumor size). Depending upon the tumor is a recurrent or nonrecurrent event in patient life, each record is also labeled with class label "Recurrence events" or "No recurrence events". Such a data attribute, which will be used as class attribute throughout the homework, is reported as the last record attribute.

The complete list of dataset attributes is reported below.

(1) Age
(2) Menopause
(3) Tumor-size
(4) Inv-nodes
(5) Node-caps
(6) Deg-malig
(7) Breast
(8) Breast-quad
(9) Irradiat
(10) class (class attribute)

Context

Oncologists want to predict the property of recurrence or not of breast tumors according to patient, tumor, and treatment characteristics. To this purpose, they exploit three different classification algorithms: a decision tree (Decision Tree) and a Bayesian classifier (Naïve Bayes), and a distance-based classifier (K-NN). The Breast dataset is used to train classifiers and to validate their performance.

Questions

Answer to the following questions:

1. Learn a Decision Tree from the whole dataset by setting the max depth threshold to 5, while keeping the default configuration for all the other parameters. (a) Which attribute is deemed to be the most discriminative one for class prediction? (b) What is the height of the Decision Tree generated? (b) Find a pure partition in the Decision Tree and report a screenshot that shows the example identified.
2. Analyze the impact of the minimal impurity (using the entropy splitting criterion), minimum number of samples to split, minimum number of samples for each leaf 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). Report at least 5 different screenshots showing Decision Trees (or portions of them) generated with different configuration settings.
3. Performing a 10-fold Stratified Cross-Validation, what is the impact the minimal impurity, minimum number of samples to split, minimum number of samples for each leaf and maximal depth parameters on the average accuracy achieved by Decision Tree? Report at least 5 screenshots showing the confusion matrices achieved using different parameter settings (consider *at least* all the configurations used to answer Question 2). Keep the default configuration for all the other parameters.
4. Considering the K-Nearest Neighbor (K-NN) classifier and performing a 10-fold Stratified CrossValidation, what is the impact of parameter K on the average classifier accuracy? Report at least 5 screenshots showing the confusion matrices achieved using different K parameter values. Perform a 10-fold Stratified Cross-Validation with classifier Naïve Bayes. Does K-NN perform on average better or worse than the Naïve Bayes classifier on the analyzed data? Report a screenshot showing the confusion matrix achieved by Naïve Bayes on the analyzed dataset.
5. Analyze the Correlation Matrix to discover pairwise correlations between data attributes. Report a screenshot showing the correlation matrix achieved. (a) Does the Naïve independence assumption actually hold for the Breast dataset? (b) Which is the pair of most correlated attributes?

Assignment

Write a 4-5 page report containing the answers to the above questions.
Analyze the performance and the behavior of the models for the different settings, underlying edge cases.