Classification Methods: Implementations in R

Sourav Adhikari, Verena Koeck

13 May 2022

(Statistical) Classification: What is it anyway?

-The problem of identifying which of a set of categories an observation belongs to.

-E.g. assigning an incoming email to “spam” or “inbox” mailbox.

-Classification can be thought of as two separate problems – binary classification and multiclass classification.

Methods of Classification

Dataset

-The Iris Dataset contains four features (length and width of sepals and petals) of 50 samples of three species of Iris (Iris setosa, Iris virginica and Iris versicolor).

-The dataset is often used in data mining, classification and clustering examples and to test algorithms.

Figure 1: Three species of Iris

Naive Bayes

-Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

-The assumption is that the features are independent, i.e presence of one particular feature does not affect the other. Hence the adjective "naive".

-Requires a small number of training data to estimate the parameters necessary for classification.

-Bayes theorem expressed as:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

-y is the category variable, and X represent the parameters/features.

k-Nearest Neighbours

-Implementation

Neural Networks

-Implementation

Summary

-Summary

References

-References