Machine Learning Reference for R

Contents

Data Preparation

Normalization

Features sometimes need to be scaled so they fit into a standard range. This involves transforming variables into a narrower or wider range than they are found in the observed data.

Two common methods for scaling features are **min-max normalization** and **z-score normalization**:

$$X_{new} = \frac{X - min(X)}{max(X) - min(X)}$$

$$X_{new} = \frac{X - \mu}{\sigma} = \frac{X - mean(X)}{StdDev(X)}$$

Algorithms

Classification Algorithms

Naive Bayes

The **Naive Bayes classifier** is a probabilistic machine learning algorithm that predicts class labels for a factor by using a probability found from the training data. The classifier assumes that all features contribute equally and are independent of each other. This classifier relies on **conditional probability**, or the probability of an event A occurring, given that an event B has occurred:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(B)}{P(A)}$$

In the Naive Bayes setting, the probability of level L for class C (denoted C_L), given feature F, is:

$$P(C_L|F) = \frac{P(F|C_L)P(F)}{P(C_L)}$$

This is generalizable to:

$$P(C_L|F_1, F_2, ..., F_n) = \frac{P(F_1, F_2, ..., F_n|C_L)P(F_1, F_2, ..., F_n)}{P(C_L)} = P(C_L) \prod_{i=1}^{n} P(F_i|C_L)$$