DS M 3 part-2

techworldthink • February 20, 2022

Understanding classification rules

Classification rules represent knowledge in the form of logical if-else statements that assign a class to unlabeled examples.

They are specified in terms of an **antecedent and a consequent**; these form a hypothesis stating that "if this happens, then that happens."

The antecedent comprises certain combinations of feature values, while the consequent specifies the class value to assign if the rule's conditions are met

Rule learners are often used in a manner similar to decision tree learners. Like decision trees, they can be used for applications that generate knowledge for future action, such as:

• Identifying conditions that lead to a hardware failure in mechanical devices

• Describing the defining characteristics of groups of people for customer segmentation

• Finding conditions that precede large drops or increases in the prices of shares on the stock market

On the other hand, rule learners offer some distinct advantages over trees for some tasks. Unlike a tree, which must be applied from top-to-bottom, rules are facts that stand alone. The result of a rule learner is often more parsimonious, direct, and easier to understand than a decision tree built on the same data

rules can be generated using decision trees.

Rule learners are generally applied to problems where the features are primarily or entirely nominal. They do well at identifying rare events, even if the rare event occurs only for a very specific interaction among features.

Separate and conquer

Classification rule learning algorithms utilize a heuristic known as separate and conquer.

The process involves identifying a rule that covers a subset of examples in the training data, and then separating this partition from the remaining data.

As rules are added, additional subsets of data are separated until the entire dataset has been covered and no more examples remain.

> *Classification rule learning algorithms utilize a heuristic known as separate and conquer. The process involves identifying a rule that covers a subset of examples in the training data, and then separating this partition from the remaining data. As rules are added, additional subsets of data are separated until the entire dataset has been covered and no more examples remain.*

Divide-and-conquer and separate-and-conquer algorithms are known as greedy learners because data is used on a first-come, first-served basis.

Greedy algorithms are generally more efficient, but are not guaranteed to generate the best rules or minimum number of rules for a particular dataset.

As the rules seem to cover portions of the data, separate-and-conquer algorithms are also known as covering algorithms, and the rules are called covering rules.

The One Rule algorithm

OneR, short for "One Rule", is a simple, yet accurate, classification algorithm that generates one rule for each predictor in the data, then selects the rule with the smallest total error as its "one rule".

To create a rule for a predictor, we construct a frequency table for each predictor (feature) against the target. It has been shown that OneR produces rules only slightly less accurate than state-of-the-art classification algorithms while producing rules that are simple for humans to interpret.

The One Rule algorithm (1R or OneR), improves over ZeroR by selecting a single rule.

The way this algorithm works is simple. For each feature, 1R divides the data into groups based on similar values of the feature. Then, for each segment, the algorithm predicts the majority class. The error rate for the rule based on each feature is calculated, and the rule with the fewest errors is chosen as the one rule

Strengths	Weaknesses
• Generates a single, easy-to-understand, human-readable rule-of-thumb • Often performs surprisingly well • Can serve as a benchmark for more complex algorithms	• Uses only a single feature • Probably overly simplistic

Rules from decision trees

Classification rules can also be obtained directly from decision trees. Beginning at a leaf node and following the branches back to the root, you will have obtained a series of decisions. These can be combined into a single rule

The chief downside to using a decision tree to generate rules is that the resulting rules are often more complex than those learned by a rule-learning algorithm. The divide-and-conquer strategy employed by decision trees biases the results differently than that of a rule learner. On the other hand, it is sometimes more computationally efficient to generate rules from trees.

> *The C5.0() function will generate a model using classification rules if you specify rules = TRUE when training the model.*

Understanding regression

Regression is concerned with specifying the relationship between a single numeric dependent variable (the value to be predicted) and one or more numeric independent variables (the predictors). We'll begin by assuming that the relationship between the independent and dependent variables follows a straight line.

You might recall from algebra that lines can be defined in a slope-intercept form similar to y = a + bx, where y is the dependent variable and x is the independent variable. In this formula, the slope b indicates how much the line rises for each increase in x. The variable a indicates the value of y when x = 0. It is known as the intercept because it specifies where the line crosses the vertical axis.

Regression equations model data using a similar slope-intercept format. The machine's job is to identify values of a and b such that the specified line is best able to relate the supplied x values to the values of y. It might not be a perfect match, so the machine should also have some way to quantify the margin of error.

Regression methods are also used for hypothesis testing, which involves determining whether data indicate that a presupposition is more likely to be true or false. The regression model's estimates of the strength and consistency of a relationship provide information that can be used to assess whether the findings are due to chance alone.

regression analysis is not synonymous with a single algorithm. Rather, it is an umbrella for a large number of methods that can be adapted to nearly any machine learning task.

basic regression models—those that use straight lines. This is called **linear regression**. If there is only a single independent variable, this is known as **simple linear regression**, otherwise it is known as **multiple regression**. Both of these models assume that the dependent variable is continuous.

It is possible to use regression for other types of dependent variables and even for classification tasks. For instance, logistic regression can be used to model a binary categorical outcome, while Poisson regression—named after the French mathematician Siméon Poisson—models integer count data. The same basic principles apply to all regression methods.

> *Linear regression, logistic regression, Poisson regression, and many others fall in a class of models known as **generalized linear models (GLM)**, which allow regression to be applied to many types of data. Linear models are generalized via the use of a link function, which specifies the mathematical relationship between x and y.*

> *Despite the name, simple linear regression is not too simple to solve complex problems.*

Simple linear regression

Simple linear regression defines the relationship between a dependent variable and a single independent predictor variable using a line denoted by an equation in the following form:

$$y\ x = + \alpha\ \beta$$

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

Simple linear regression is used to estimate the relationship between two quantitative variables. You can use simple linear regression when you want to know:

1. How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).
2. The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

Assumptions of simple linear regression

Simple linear regression is a parametric test, meaning that it makes certain assumptions about the data. These assumptions are:

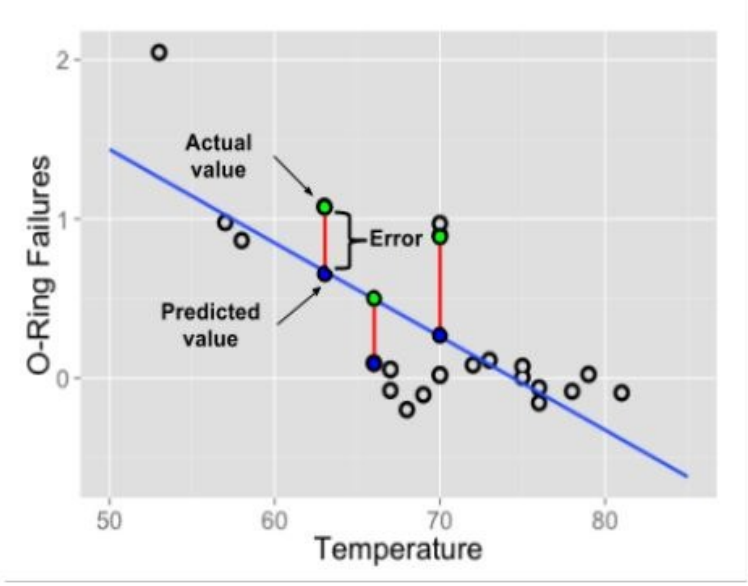
1. Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn't change significantly across the values of the independent variable.
2. Independence of observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.
3. Normality: The data follows a normal distribution.

Linear regression makes one additional assumption:

1. The relationship between the independent and dependent variable is linear: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).

Ordinary least squares estimation

In order to determine the optimal estimates of α and β, an estimation method known as ordinary least squares (OLS) was used. In OLS regression, the slope and intercept are chosen such that they minimize the sum of the squared errors, that is, the vertical distance between the predicted y value and the actual y value. These errors are known as residuals,



In mathematical terms, the goal of OLS regression can be expressed as the task of minimizing the following equation:

$$\sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$

In plain language, this equation defines e (the error) as the difference between the actual y value and the predicted y value. The error values are squared and summed across all points in the data.

The method of **least squares** is a standard approach in regression analysis to approximate the solution of overdetermined systems (sets of equations in which there are more equations than unknowns) by minimizing the sum of the squares of the

residuals (a residual being: the difference between an observed value, and the fitted value provided by a model) made in the results of each individual equation.