Data Science Essentials

Getting Started with Machine Learning

Classification and Regression

Classification and regression use data with known values to train a machine learning model so that it can identify unknown values for other data entities with similar attributes.

Classification is used to identify distinct values. Regression is used to identify real numeric values. So a question like "In this a chair?" is a two-class classification problem (with possible values True and False), while "How much does this person earn?" is a regression problem.

Both classification and regression are examples of *supervised learning*, in which a machine learning model is trained using a set of existing, known data values. The basic principle is as follows:

Define data entities based on a collection (or *vector*) of numeric variables (which we call *features*) and a single predictable value (which we call a *label*). In classification, the label has a value of -**1** for False and **+1** for True.

Assemble a training set of data that contains many entities with known feature and label values - we call the set of feature values ***x*** and the label value ***y***.

Use the training set and a classification or regression algorithm to train a machine learning model to determine a function (which we call ***f***) that operates on ***x*** to produce ***y***.

The trained model can now use the function ***f(x)=y*** to calculate the label (***y***) for new entities based on their feature values (***x***). In classification, if *y* is positive, the label is True; if ***y*** is negative, the label is False. The function can be visualized as a line on a chart, showing the predicted ***y*** value for a given ***x*** value. The predicted values should be close to the actual known values for the training set, as shown in figure 1 below.

You can evaluate the model by applying it to a set of test data with known label (*y*) values. The accuracy of the model can be validated by comparing the value calculated by ***f(x)*** with the known value for ***y***, as shown in figure 2.

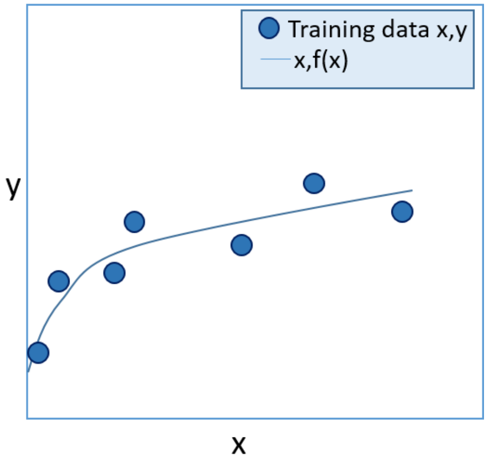

Figure 1: A trained model

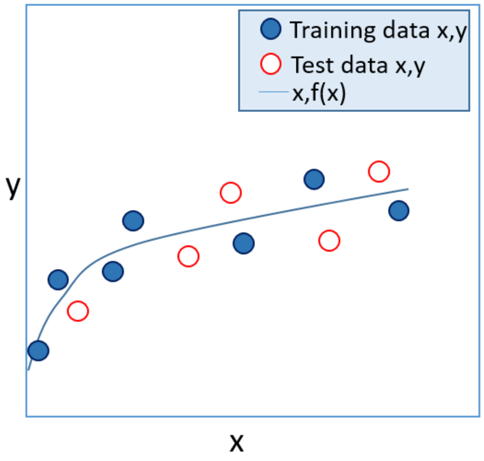

Figure 2: A validated model

In a supervised learning model, the goal is to produce a function that accurately calculates the known label values for the training set, but which is also generalized enough to predict accurately for known values in a test set (and for unknown values in production data). Functions that only work accurately with the training data are referred to as "over-fitted", and functions that are too general and don't match the training data closely enough are referred to as "under-fitted". In general, the best functions are ones that are complex enough to accurately reflect the overall trend of the training data, but which are simple enough to calculate accurate values for unknown labels.

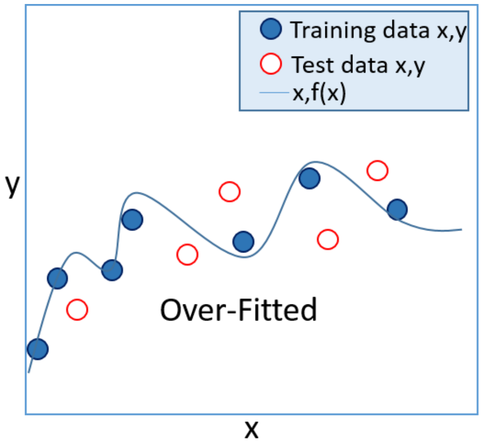

Figure 3: An over-fitted model

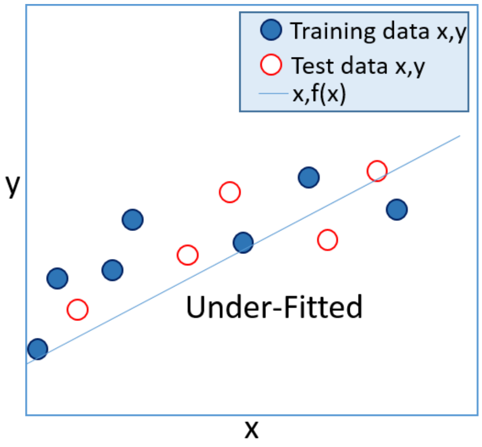

Figure 4: An under-fitted model

Clustering

Clustering is an *unsupervised learning* technique in which machine learning is used to group (or *cluster*) data entities based on similar features.

It is difficult to evaluate clustering models, because there is no training set with known values that you can use to train the model and compare its results.

A common algorithm for clustering is K-Means, in which data values are iteratively divided into *K* clusters based on distance from a centroid point.