

Machine Learning Theory

Wesley Matheus

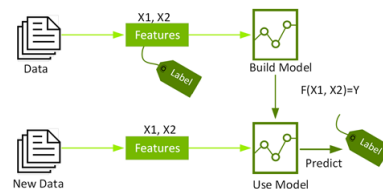
Contents

Supervised and Unsupervised Learning	2
Supervised Learning	3
Classification	3
Regression	4
Unsupervised learning	5
Clustering	5
Association learning	5

Supervised and Unsupervised Learning

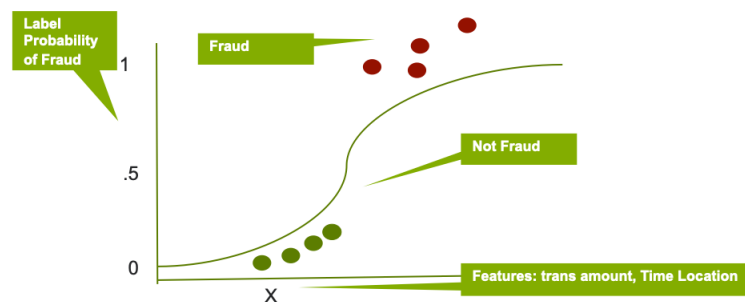
Supervised Learning

Supervised machine learning, also called predictive analytics, uses algorithms to train a model to find patterns in a dataset with labels and features. It then uses the trained model to predict the labels on a new dataset's features. Supervised learning can be further categorized into classification and regression. Supervised learning algorithms include: Logistic regression, Linear regression, Support vector machines, Decision trees, Random forest, Gradient boosting decision trees.



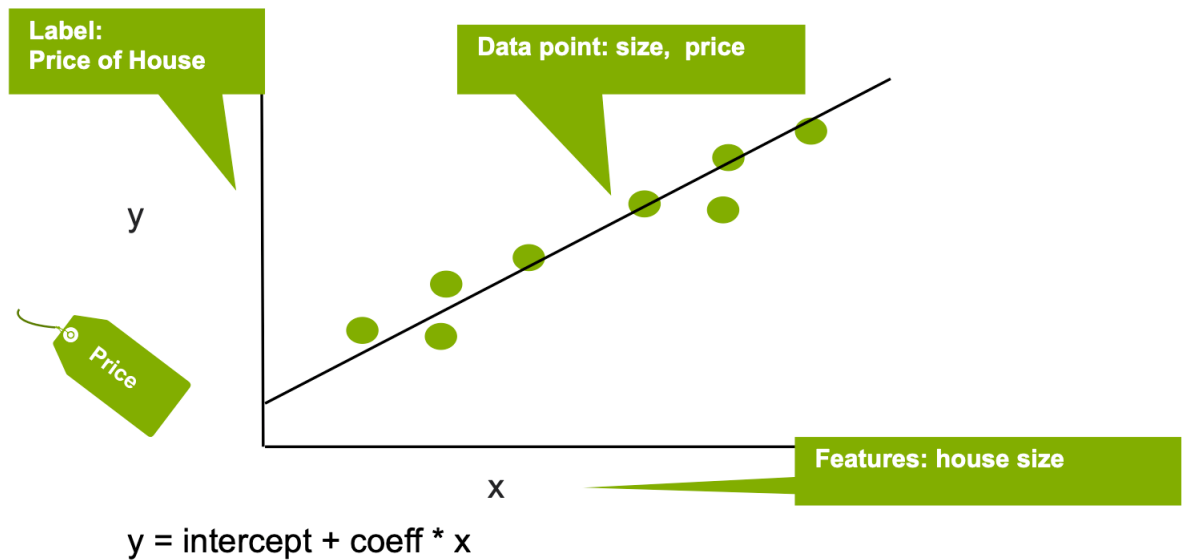
Classification

Classification identifies which category an item belongs to based on labeled examples of known items. In the simple example below, logistic regression is used to estimate the probability of whether a credit card transaction is fraudulent or not (the label) based on features of transactions (transaction amount, time, and location of last transaction) known to be fraudulent or not.



Regression

Regression estimates the relationship between a target outcome label and one or more feature variables to predict a continuous numeric value. In the simple example below linear regression is used to estimate the house price (the label) based on the house size (the feature).

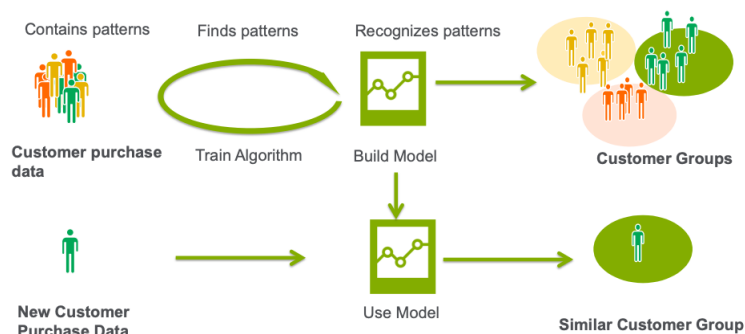


Unsupervised learning

Unsupervised learning, also called descriptive analytics, doesn't have labeled data provided in advance, and can aid data scientists in finding previously unknown patterns in data. These algorithms attempt to 'learn' the inherent structure in the input data, discovering similarities, or regularities. Common unsupervised tasks include clustering and association.

Clustering

In clustering, an algorithm classifies inputs into categories by analyzing similarities between input examples. An example of clustering is a company that wants to segment its customers in order to better tailor products and offerings. Customers could be grouped on features such as demographics and purchase histories. Clustering with unsupervised learning is often combined with supervised learning in order to get more valuable results. Other clustering examples include: Search results grouping, Grouping similar entities, Text categorization, Anomaly detection (finding what is not similar, meaning the outliers from clusters).



Association learning

Association or frequent pattern mining finds frequent co-occurring associations (relationships, dependencies) in large sets of data items. An example of co-occurring associations is products that are often purchased together, such as the famous beer and diaper story. An analysis of

behavior of grocery shoppers discovered that men who buy diapers often also buy beer. Un-supervised learning algorithms include: K-means, Latent Dirichlet allocation (LDA), Gaussian Mixture Model (GMM), Alternating least squares (ALS), FP-growth.

Bibliography

[1] Nvidia