**Supervised learning**

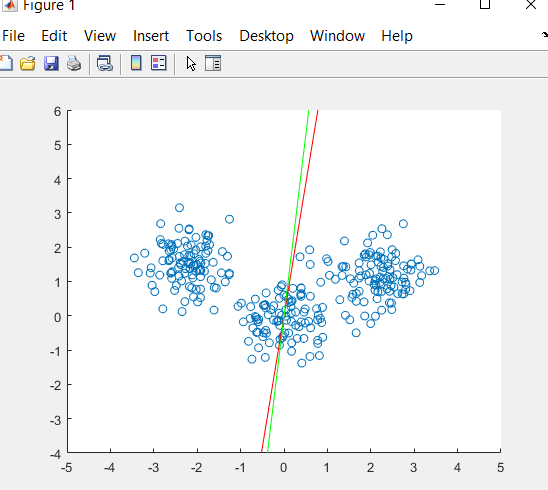
A training set is used in supervised learning to instruct models to produce the desired results. This training dataset has both the right inputs and outputs, enabling the model to develop over time. The loss function serves as a gauge for the algorithm's correctness, and iterations are made until the error is suitably reduced. When using data mining, supervised learning may be divided into two categories of issues: regression and classification.

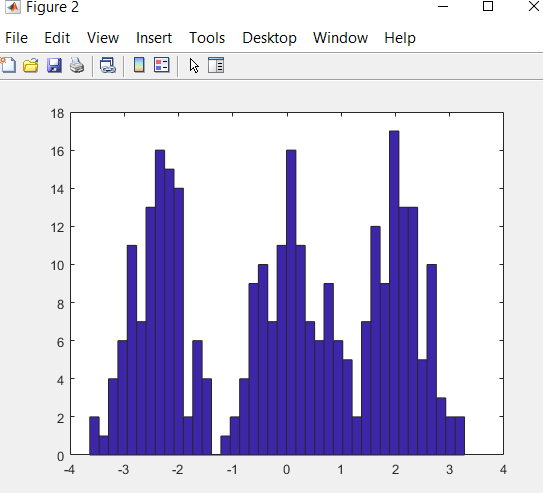
**Classification**

In order to correctly classify test data into different groups, classification employs an algorithm. It identifies certain entities in the dataset and makes an effort to determine how those things should be described or labelled. The following classification techniques are often used: decision trees, k-nearest neighbor, random forest, support vector machines (SVM), linear classifiers, and SVM.

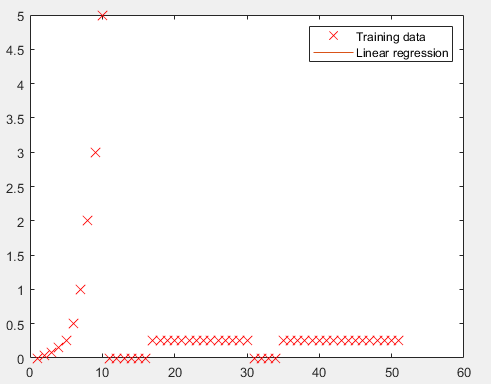
**Regression**

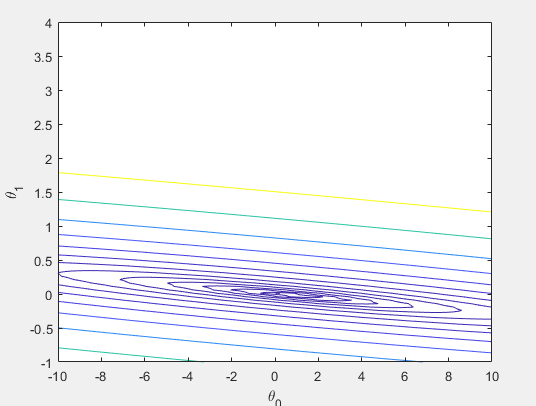
To comprehend the relationship between dependent and independent variables, regression is employed. It is frequently used to produce estimates, including those for a company's sales revenue. Popular regression techniques include linear regression, logistical regression, and polynomial regression.



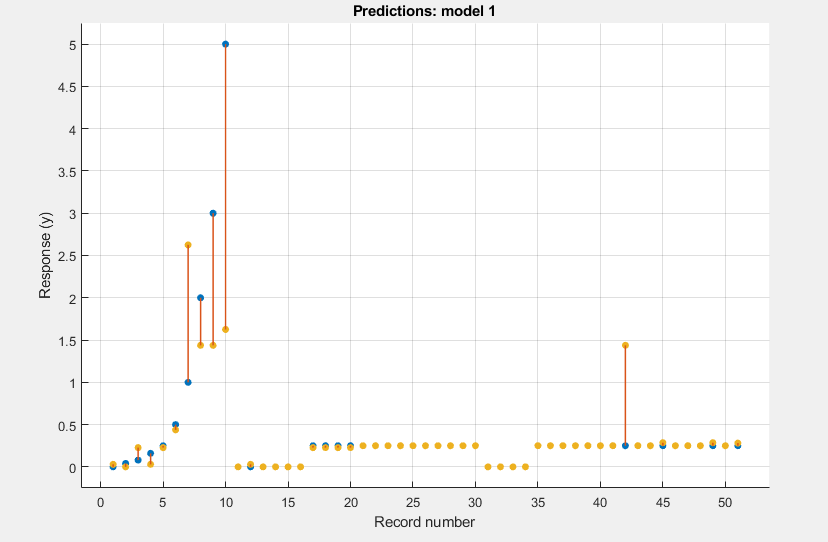


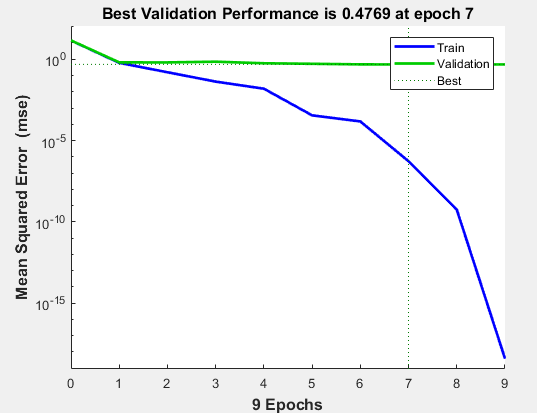
Neural networks handle training data by simulating the connection of the human brain using layers of nodes, which is mostly used for deep learning algorithms. Inputs, weights, a bias (or threshold), and an output make up each node. This "fires" or activates the node, sending data to the following layer in the network, if the output value exceeds a predetermined threshold. This mapping function is learned by neural networks using supervised learning, with gradient descent adjustments made in response to the loss function. We can be sure in the model's accuracy to get the right answer when the cost function is at or close to zero.





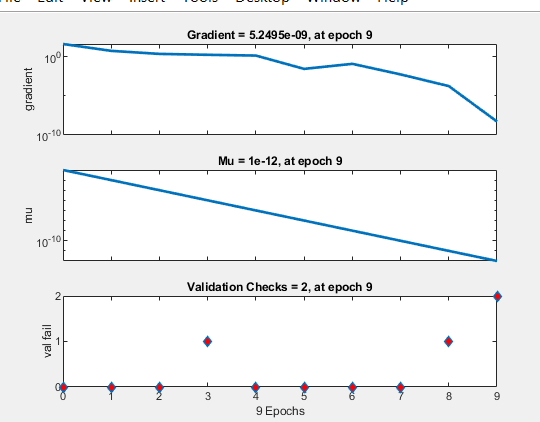
While logistical regression is used when the dependent variable is categorical, or has binary outputs, such as "true" and "false" or "yes" and "no," linear regression is used when the dependent variable is continuous. Despite the fact that both regression models aim to identify the connections between the data inputs, logistic regression is mostly employed to address binary classification issues, such as spam detection.

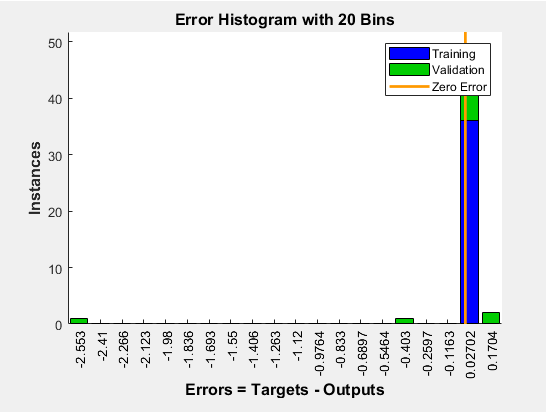




A lot has been written about PCA on the internet, and there are some excellent articles on it, but many of them spend too much time getting too technical when most of us just want to know how it works simply.

It is possible to divide principal component analysis into five phases. I'll walk you through each step while rationally explaining what PCA does and demystifying complicated mathematical ideas like standardization, covariance, eigenvectors, and eigenvalues without concentrating on how to calculate them.





Principal component analysis, or PCA, is a technique for reducing the number of dimensions in big data sets by condensing a large collection of variables into a smaller set that retains the majority of the large set's information.

Accuracy inevitably suffers as a data set's variables are reduced, but the answer to dimensionality reduction is to trade a little accuracy for simplicity. Since machine learning algorithms can analyses data much more quickly and easily with smaller data sets because there are less unnecessary factors to evaluate.

