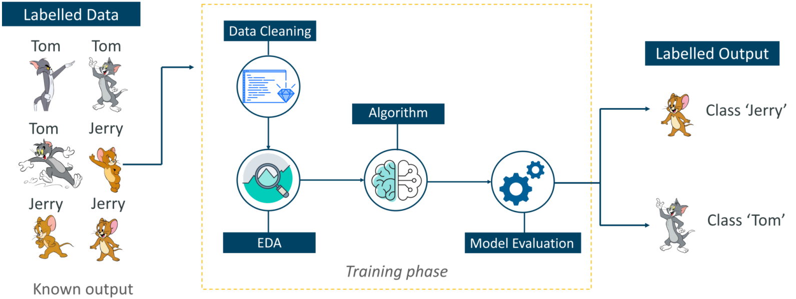
Week 18 -Supervised Learning

Supervised Learning



Supervised Learning

- Dataset contains problem instances coupled with the correct solutions. Each case of the training data consists of an input object (typically a vector of attribute-value pairs) and a desired output value (the supervisory signal or the teaching signal).
- Models use what they have learned to predict the correct solutions for new problems not contained in the dataset.

Examples of Problems suitable for Supervised Learning

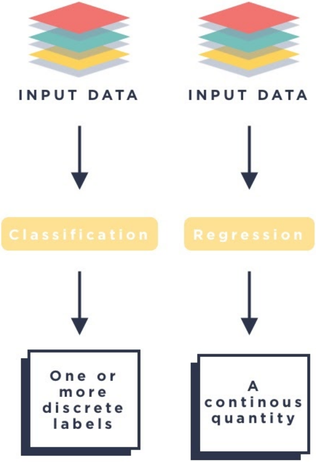
- Medical diagnosing: each patient is a problem instance described by a set of attributes (e.g. age, gender, height, weight, known past diseases, genetic risks, values from blood tests, symptoms, ...) and the solution corresponding to this problem instance is a binary target attribute whether or not the patient has diabetes.

- Credid Risk Assignment: Every credit card customer is a problem instance with attributes such as age, gender, average account balance, area of residence, occupation, purchase history,... and the solution corresponding to this problem instance is a target attribute of "low risk," "medium risk" or "high risk" (based on e.g. repayment history)

Examples of Problems suitable for Supervised Learning

- News Recommendation: every registered user has a particular reading history, formed by the topics of the news he has read, where he has clicked, and how much time was spent reading each article (or watching each video). An article is represented as a problem instance (attributes: topic of article, keywords of article, opinion targets in the article, sentiments about opinion targets, subscription fee of registered user, average subscription duration of readers of the article, average time readers of the article spent on website, average time readers of the article spent on each section of website, keywords of other articles read, ...) and the solution corresponding to this problem instance is either "interesting" or "not interesting" for a particular registered user (depending on the time spent, if any, reading the article).

Supervised Learning Algorithms



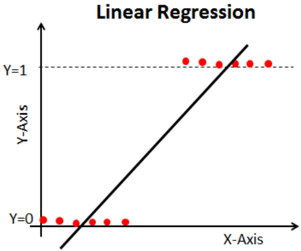
Classification vs Regression

- Are there only a few (or 2) options the value that my outcome can be? If yes, you are in a classification situation. If your outcome is continuous, or could be any value in a range, you are in a situation that requires regression.
- In classification: distinguish binary problems (is this a cat? yes / no) or multi-class problems (what type of animal is this?).

Supervised Learning: Classification

- Logistic Regression
- Decision Trees
- Support Vector Machines
- K-Nearest Neighbors

Logistic Regression



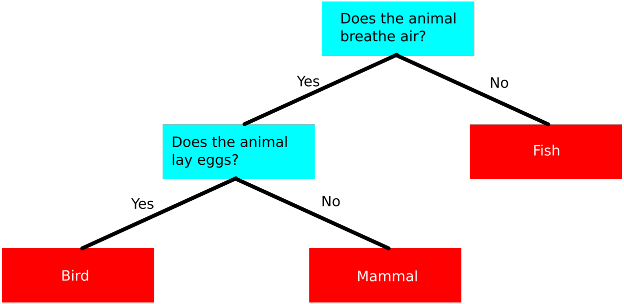
Logistic Regression line represents the likelihood that a particular outcome will occur based on the features provided. The likelihood is then rounded to either 0 or 1 (each of the two classes) based on a cutoff (usually 0.5).

Logistic Regression: Assumptions

- Large outliers are rare
- No two variables in the features that are highly correlated to each other (any variable can, of course, be highly correlated to the outcome).

Depending on the results of our tests during the data exploration phase of our project, we might not be able to apply a logistic regression, or have to transform our data before doing so.

Decision Trees



Type of algorithm that break down the training data into smaller and smaller subsets by asking more specific questions until the data can be classified. New data is classified sequentially, at each node (question) of the decision tree by the attributes that it has for that feature.

Decision Trees

- Decision trees are a useful method for classification because they can easily model non-linear relationships, and are not easily influenced by extreme data points (outliers).

- However, they can also learn the data too well, by asking too many or too specific questions (overfitting!)

Advantages of Decision Trees

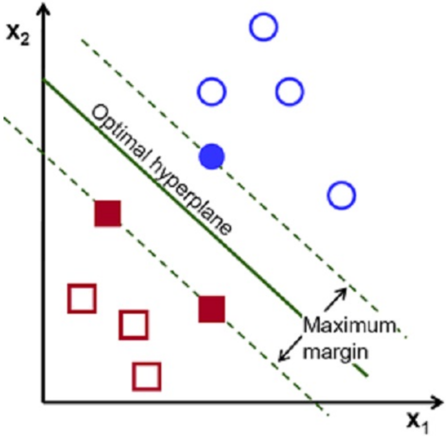
- Decision tree algorithms are good at modelling complex (nonlinear) relationships without needing to specify the format of those relationships beforehand.
- Examining the structure of the decision tree can also provide us with interpretable results to understand how each feature determines the outcome.

Problems of Decision Trees: Overfitting

- The more nodes that are created, the more 'deeply' the tree follows the patterns in the training data.

- If it goes too deep, it will pick up more noise, meaning that the algorithm could overfit to the training data, and will not be able to generalize well when it is applied to new data.

Support Vector Machines

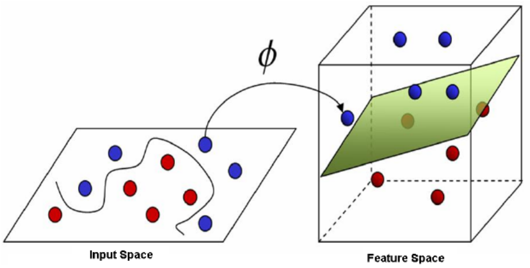


Support Vector Machines draw the line between the two classes on our graph that maximizes the distance between the data and the line.

Support Vector Machines

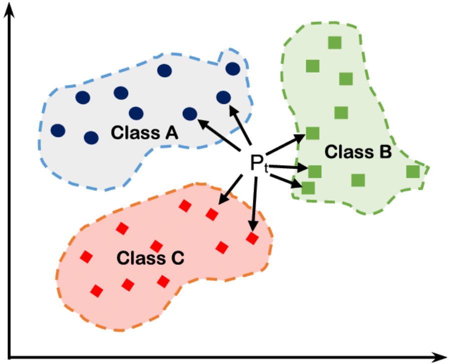
- SVMs use distance measurements to find the optimal classification.
- Support Vector Machines draw the line between the two classes on our graph that maximizes the distance between the data and the line. (in contrast to 0.5 cutoff for logistic regression)
- This ensures that when the algorithm is asked to classify new data, even if this new data lies a little bit closer to the wrong class than the training data from its class, it is likely to be classified into the correct class.

Support Vector Machines



It is possible for SVM to model non-linear boundaries both in two or higher dimensions,

K Nearest Neighbors (KNN)



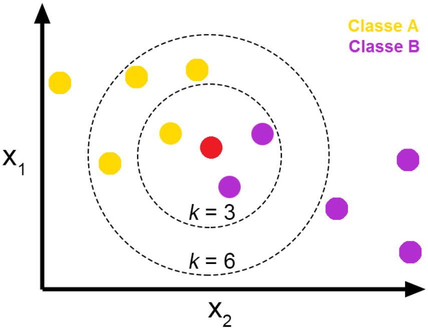
This algorithm that uses the nearest points to determine the output class of new data points..

K Nearest Neighbors (KNN)

Example: Voting Data

- We have data from a survey of voters, including demographics, opinions on the issues, and voting preference.
- The new, unseen data (or test data) for which we would like to predict the outcome is compared to the training data.
- The label for the new datapoint is assigned based on the 'vote' of the labels of a predetermined number (K) of data points that are closest to the test data point.
- For instance, for K=5, the label of the test data point will be determined by the majority vote of the classes of the points closest to it.

K Nearest Neighbors (KNN)



The number we pick for K plays an important role in determining the class of the point in some cases.

K Nearest Neighbors

Advantages

- KNN benefits from the fact that it uses all the data available to assign the new datapoint its outcome value.

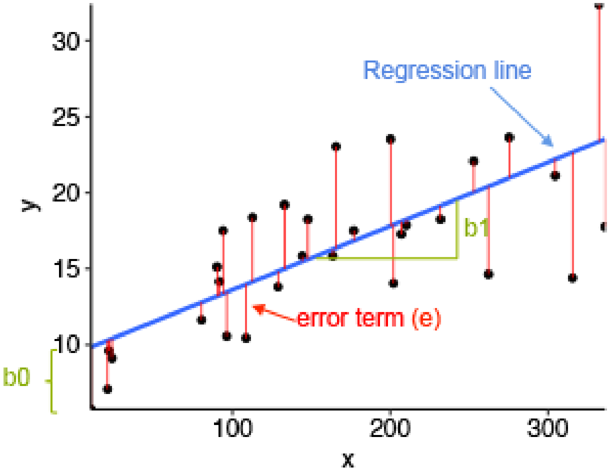
Disdvantages

- Depending on the size of the dataset, KNN can be a memory-intensive approach and can result in underperformance if the data contains many features.

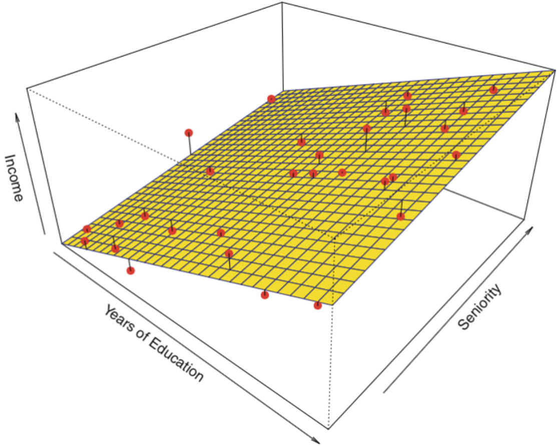
Supervised Learning: Regression

- Linear Regression
- Decision Trees
- K-Nearest Neighbors
- Time Series

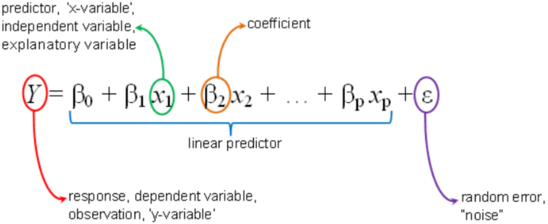
Linear Regression



Linear Regression



Linear Regression



predictor, 'x-variable', independent variable, explanatory variable

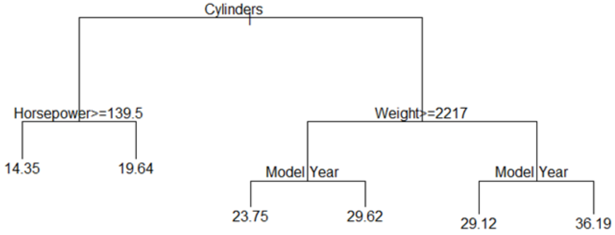
coefficient

$$Y = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p}_{\text{linear predictor}} + \varepsilon$$

response, dependent variable, observation, 'y-variable'

random error, "noise"

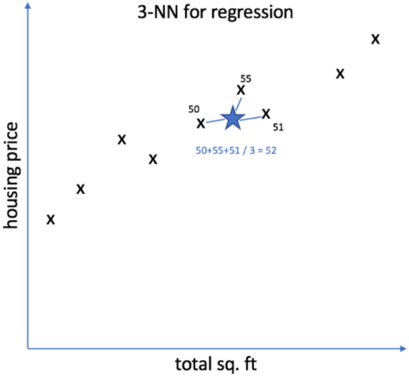
Decision Trees



Decision tree uses the training data to create nodes (splits) based on those features that have the most determining power on the outcome.

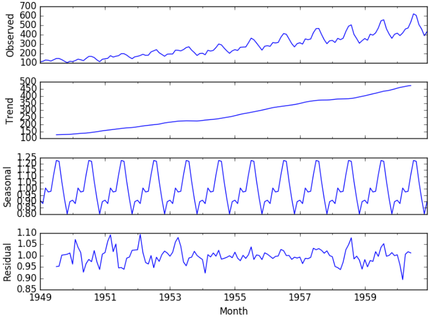
Each tree returns a value or an average value for the trees that are in the group at the final leaf of our tree.

K Nearest Neighbors (KNN)



The outcome for the new or test datapoint is determined by calculating the mean of the outcome variable for those 'k' number of datapoints closest to the datapoint in question.

Time Series



A time-series regression problem typically takes the form of predicting a value in a future period. Usually, only the past values of the outcome, not related features, are used to predict the future outcome.