

Regression and Classification in Matlab

EL2310 – SCIENTIFIC PROGRAMMING

Regression

- Often in statistics one has to find the relation between the dependent and the independent variables

$$Y \approx f(X, \beta)$$

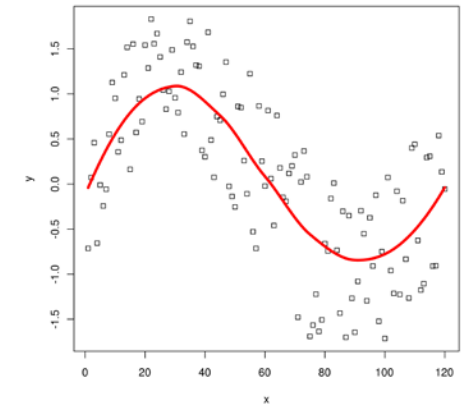
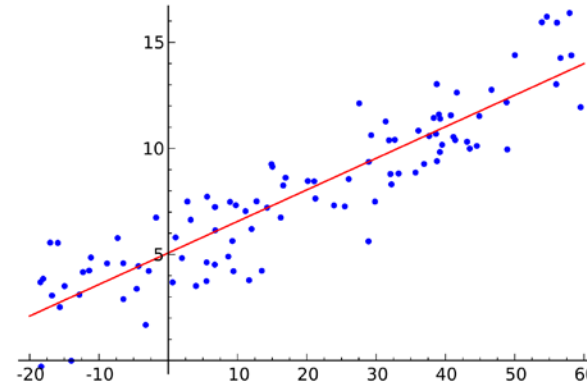
- Regression analysis helps in estimating the relationships among those variables (finding f).
- Widely used for prediction and estimation

Methods of Regression

- Linear

- Continuous outcome

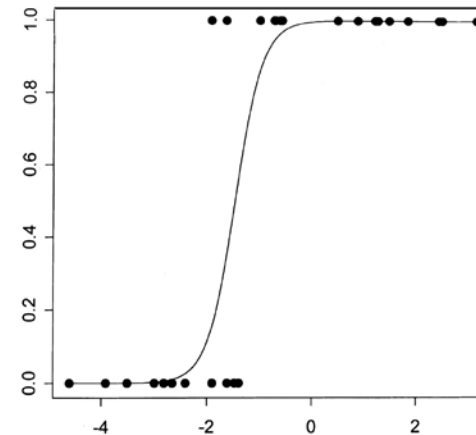
$$Y = \beta^T X$$



- Logistic

- Binary outcome

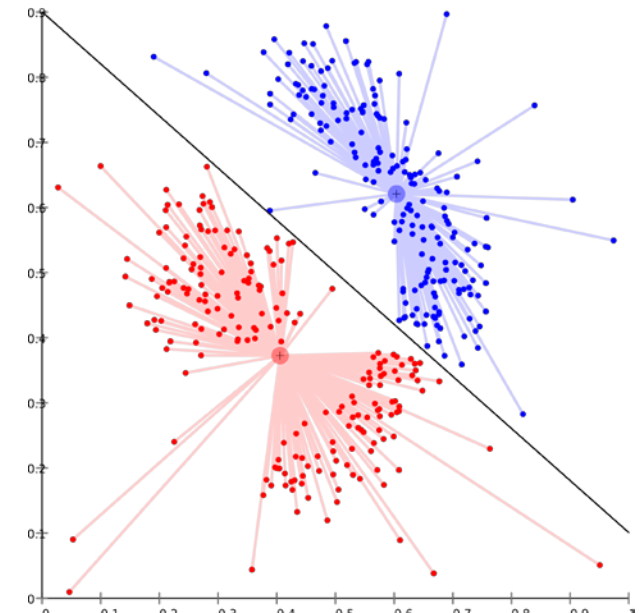
$$\sigma(t) = \frac{1}{1 + e^{-t}}, \text{ where } t = \beta_0 + \beta_1 x$$



- Neural Networks

Classification

- The problem of identifying to which, of set categories, each sample belongs.
- Classifiers are functions that map the data to a category.
- Training sets are used in order to build a classifiers for the given data.
 - The training sets can lead to supervised or unsupervised learning.



Supervised/Unsupervised Learning

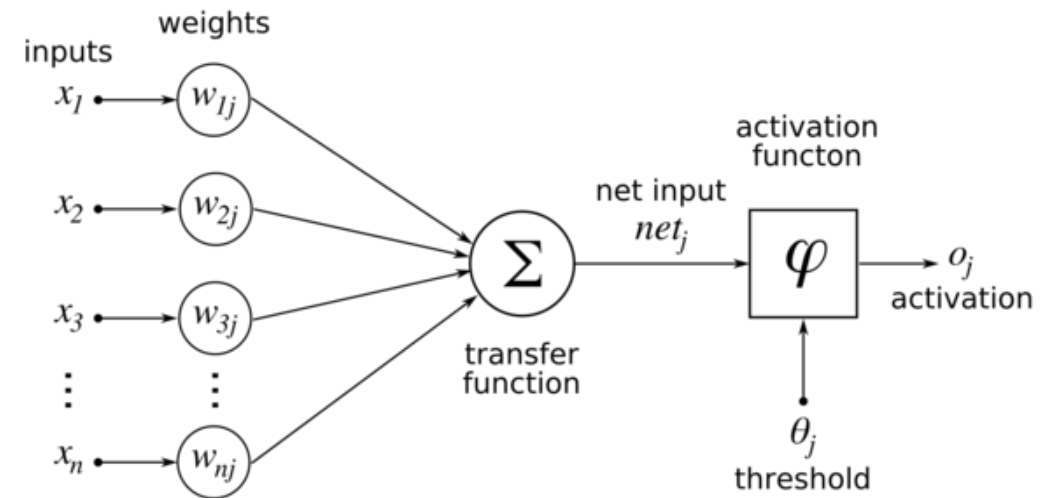
- The training sets used can either be labeled or unlabeled.
- Labeled training sets have the category of each data point pre-defined.
 - Algorithms using labeled training sets are said to use supervised learning.
 - Example: Neural Networks
- In unlabeled training sets, the category each point belongs to is unknown.
 - Results to unsupervised learning algorithms.
 - Algorithms try to categorize the data.
 - Examples: k-means Clustering, Expectation-Maximization (EM)

Popular Methods of Classification

- Clustering
- Neural Networks
- Decision Trees

Neural Networks

- Inspired by biological neural networks in animals.
- Increasingly popular method used for:
 - Robotics, control, image processing etc.
- Each node is a numeric weight that can be adjusted by training the network.
- Capable of approximating non-linear functions.



Neural Network Example

Wines – 0 is red, 1 is white

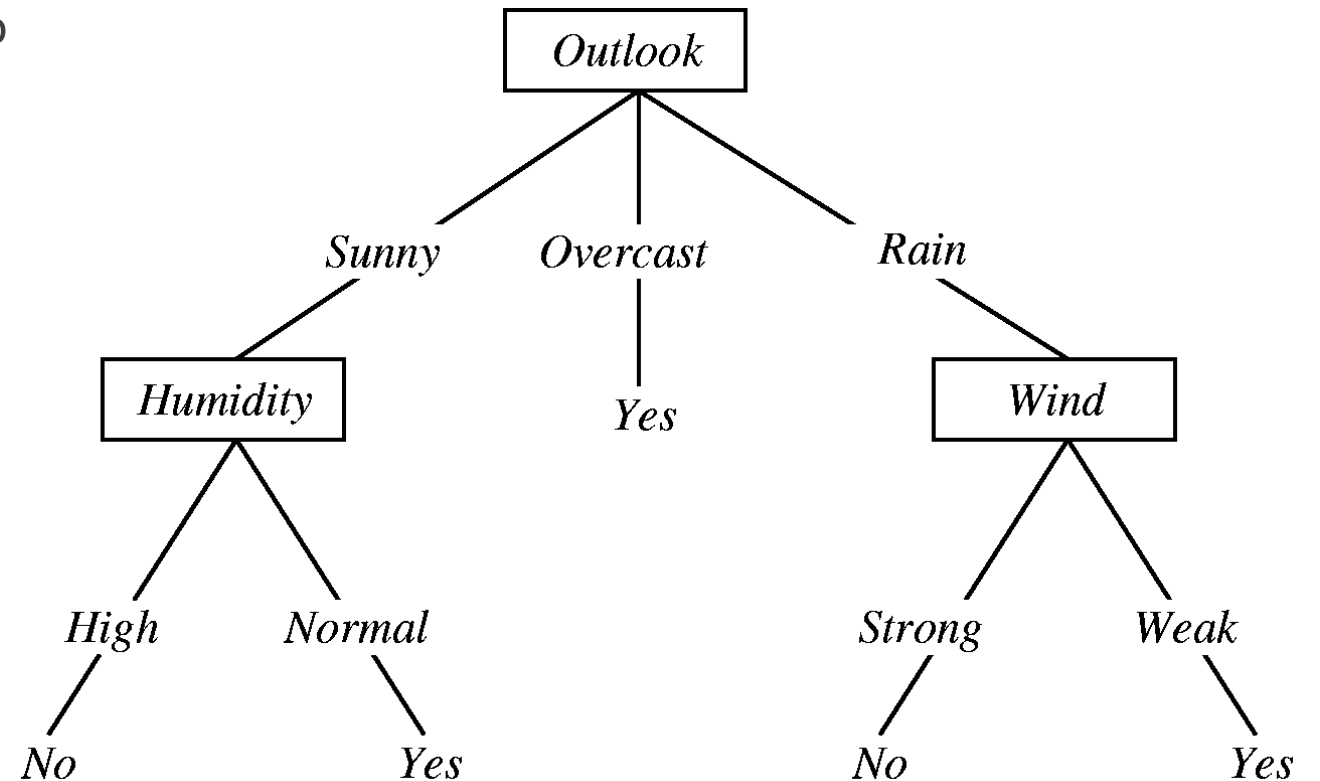
fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	wine type
7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6	0
6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6	1
8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6	0
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6	0
7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6	1
8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6	1
6.2	0.32	0.16	7	0.045	30	136	0.9949	3.18	0.47	9.6	6	1
7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6	1
6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6	0

12 variables (inputs), binary outcome

Source: UCI – Machine Learning Repository (<http://archive.ics.uci.edu/ml/datasets.html>)

Decision Trees

- Attribute definition spaces are divided so that an optimal decision can be made.
- Several algorithms available
 - ID3 (Iterative Dichotomiser 3)
 - C4.5
 - CART



Decision Tree Example

- 3 flower types
 - Setosa
 - Versicolor
 - Virginica
- 4 flower attributes
 - Sepal length (x_1)
 - Sepal width (x_2)
 - Petal length (x_3)
 - Petal width (x_4)
- 150 samples

