Contents

1. Classification
    - Framework and notations

Contents

1. Classification
   - Framework and notations
   - Confusion matrix

Contents

1. Classification
   - Framework and notations
   - Confusion matrix
   - Logistic Regression motivation

Contents

1. Classification
   - Framework and notations
   - Confusion matrix
   - Logistic Regression motivation
   - Logistic Regression algorithm

Contents

1. Classification
    - Framework and notations
    - Confusion matrix
    - Logistic Regression motivation
    - Logistic Regression algorithm
    - Summary

In fact this is the same notation as in the linear models chapter, except that now Y_i is discrete and represent generally a class number $Y_i \in \{0, 1, ...K\}$
$Y_i = h(X^i) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + ... + \beta_i X_{i,p}$
Moreover the prediction is a probability to be in the class k. In a two-class problem, you can decide Y_i is true if proba.> 0.5

Array TBD , Vocabulary

- How to understand datas, correlation but not causation
- How to handle variety of datas (see 3V description later)
- see materials for supervised (*) and unsupervised (*) use-cases

(*) Definition to come later

- 3V definition :
    - Volumen, Velocity, Variety
    - + Veracity, ++
- Position of data mining vs ML vs Statistical Learning vs AI

Starting point :

- Outcome measurement Y (also called dependent variable, response, target) ;
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables). X is a matrix of dimension (N,p), where n is the number of measurements ;
- In the **regression problem**, Y is quantitative (e.g price, blood pressure) ;
- In the **classification problem**, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample) ;
- We have training data (x1,y1),..., (xN,yN). These are observations (examples, instances) of these measurements.