

# Classification

Mohammad Emtiyaz Khan  
EPFL

Oct 8, 2015



ÉCOLE POLYTECHNIQUE  
FÉDÉRALE DE LAUSANNE

©Mohammad Emtiyaz Khan 2015


# Classification

Similar to regression, [classification](#) relates input variables  $\mathbf{x}$  to the output variable  $y$ , but now  $y$  can take only discrete values, i.e.  $y$  is a categorical (or nominal) variable.

## Binary classification

When  $y$  can only take two discrete values, it is called [binary classification](#). We will denote these values as  $y \in \{\mathcal{C}_1, \mathcal{C}_2\}$ . These values are also called [class labels](#) or simply [classes](#). Other common notations are  $y \in \{-1, +1\}$  or  $y \in \{0, 1\}$ , although there may not necessarily be any ordering between the two classes.

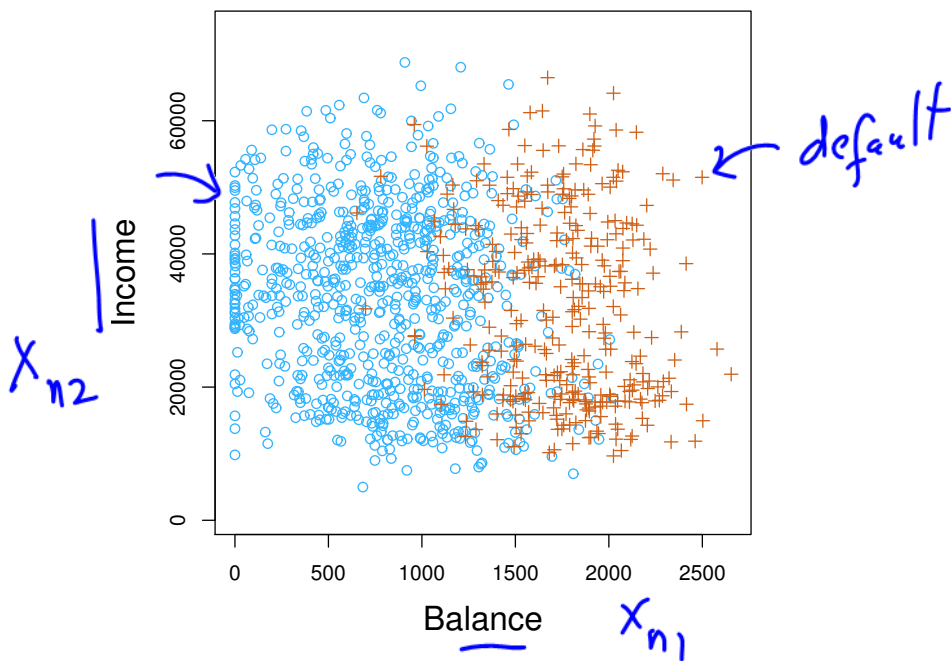
## Multi-class classification

In a [multi-class classification](#),  $y$  can take multiple discrete values i.e.  $y \in \{\mathcal{C}_0, \mathcal{C}_1, \dots, \mathcal{C}_{K-1}\}$  for a  $K$ -class problem. Again, there is no notion of ordering among these classes, but we may ignore this fact and may sometimes use the following notation for convenience:  $y \in \{0, 1, 2, \dots, K - 1\}$ . 

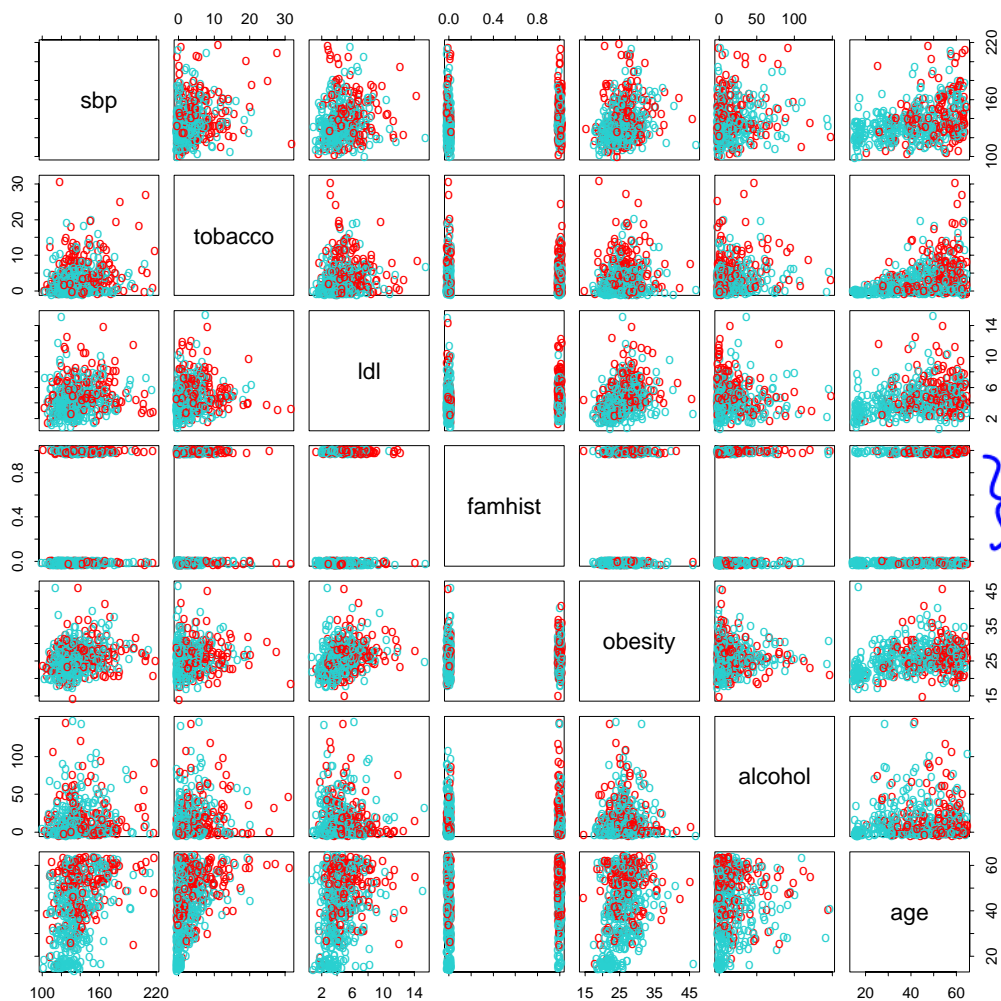
# Examples of classification problems

An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the users IP address, past transaction history, and so forth.

An example is shown below on the *default* dataset from JWHT. The annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.



A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of ~~three~~<sup>the two</sup> medical conditions. Which of the ~~three~~<sup>two</sup> conditions does the individual have?

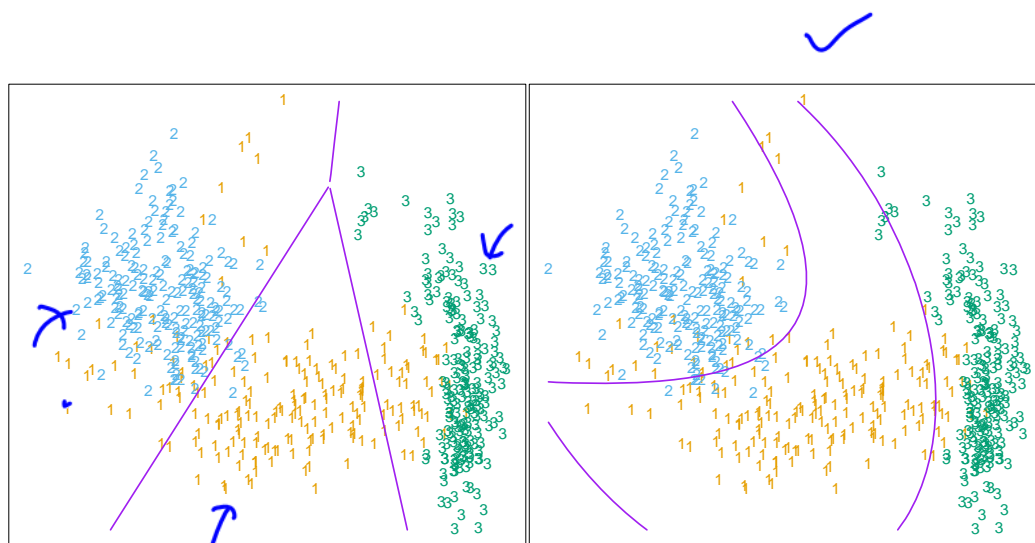


**FIGURE 4.12.** A scatterplot matrix of the South African heart disease data. Each plot shows a pair of risk factors, and the cases and controls are color coded (red is a case). The variable family history of heart disease (`famhist`) is binary (yes or no).

# Classifier

A **classifier** will divide the input space into a collection of regions belonging to each class. The boundaries of these regions are called **decision boundaries**. A classifier can be **linear or nonlinear**.

Elements of Statistical Learning (2nd Ed.) ©Hastie, Tibshirani & Friedman 2009 Chap 4



**FIGURE 4.1.** *The left plot shows some data from three classes, with linear decision boundaries found by linear discriminant analysis. The right plot shows quadratic decision boundaries. These were obtained by finding linear boundaries in the five-dimensional space  $X_1, X_2, X_1X_2, X_1^2, X_2^2$ . Linear inequalities in this space are quadratic inequalities in the original space.*