Data formats

In a supervised learning problem, there will always be a dataset, defined as a finite set of real vectors with **m** features each:

$$X = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n\} \; where \; \bar{x}_i \in \mathbb{R}^m$$

Considering that our approach is always probabilistic, we need to consider each **X** as drawn from a statistical multivariate distribution **D**. For our purposes, it's also useful to add a very important condition upon the whole dataset **X**: we expect all samples to be **independent andidentically distributed** (**i.i.d**). This means all variables belong to the same distribution **D,** and considering an arbitrary subset of **m** values, it happens that:

$$P(\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m) = \prod_{i=1}^{m} P(\bar{x}_i)$$

The corresponding output values can be both numerical-continuous or categorical. In the first case, the process is called **regression**, while in the second, it is called **classification**. Examples of numerical outputs are:

$$Y = \{y_1, y_2, \dots, y_n\} \; where \; y_n \in (0,1) \; or \; y_i \in \mathbb{R}^+$$

Categorical examples are:

$$y_i \in \{red, black, white, green\} \; or \; y_i \in \{0,1\}$$

We define generic **regressor**, a vector-valued function which associates an input value to a continuous output and generic **classifier**, a vector-values function whose predicted output is...

Previous Section (/book/big_data_and_business_intelligence/9781785889622/2)

Next Section (/book/big_data_and_business_intelligence/9781785889622/2/ch02lvl1sec1