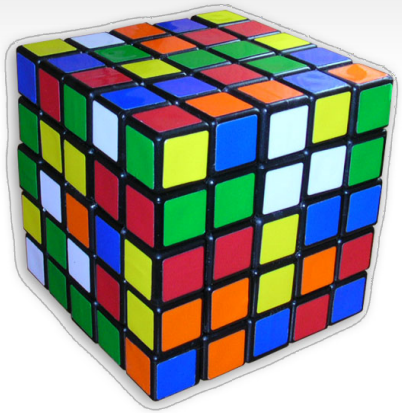
# Mineração de Dados Aplicada

## Dependence between Attributes



**Loïc Cerf**

March, 19th 2025

DCC – ICEx – UFMG

DCC
DEPARTAMENTO DE
CIÊNCIA DA COMPUTAÇÃO

UFMG

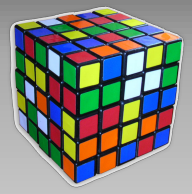
# Datasets

- Most datasets are (potentially infinite) *sets* of objects described with attributes;
- More structured datasets exist (one large graph, a video, etc.) and their components (vertices of a graph, scenes of a video, etc.) can be described with attributes derived from the structure;
- Data are usually incomplete, inconsistent, with some exceptions, uncertain/noisy or even plainly wrong;
- Understanding the application and the data generation/acquisition process is essential.



# Attributes

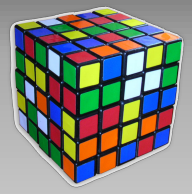
- Basic statistics (center, dispersion, etc.) and visualizations (histograms, box plots, etc.) help a lot in understanding the data; looking at specific objects too;
- The simplest attributes can be categorized into four types: nominal, ordinal, interval and ratio;
- The type tells what operations, statistics and data mining algorithms are applicable;
- More structured attributes exist (sequences, sounds, etc.) and simpler attributes (presences of sub-sequences, BPM, etc.) can be derived from the structures;
- Additional attributes, taken from other sources or derived from the existing data, may be more relevant.





# Outline

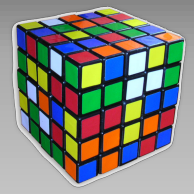
# Outline





# Stacked histogram

The *stacked histogram* allows to visualize a dependence between nominal attributes.



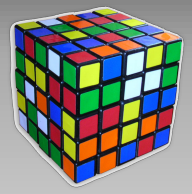
# Conditional box plot

The *conditional box plot* allows to visualize a dependence between a nominal attribute and an interval-scaled attribute.



# Scatter plot

A *scatter plot* allows to visualize a dependence between two (or three) interval-scaled attributes. In presence of many objects, a random sample can be used.





# Scatter plot

A *scatter plot* allows to visualize a dependence between two (or three) interval-scaled attributes. In presence of many objects, a random sample can be used.

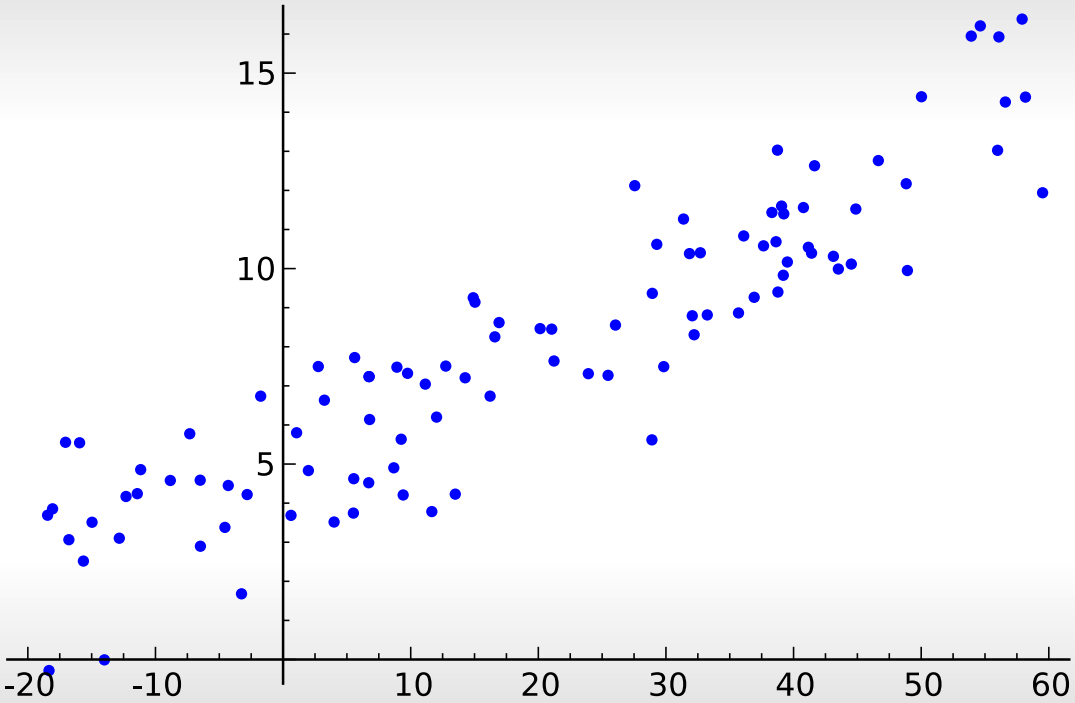
Simultaneously visualizing additional attributes of the objects:

Nominal shapes and/or colors with different hues;

Interval areas and/or colors on a brightness gradient.

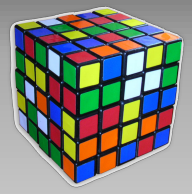


# Simple scatter plot





# Outline



# Regression

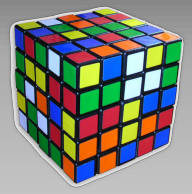
**Definition**

Learning from a set of objects how to predict the value of an attribute on an interval scale.

Input:

| | $a_1$ | $\ldots$ | $a_n$ | $y$ |
|---|---|---|---|---|
| $o_1$ | $d_{1,1}$ | $\ldots$ | $d_{1,n}$ | $y_1$ |
| $o_2$ | $d_{2,1}$ | $\ldots$ | $d_{2,n}$ | $y_2$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $o_m$ | $d_{m,1}$ | $\ldots$ | $d_{m,n}$ | $y_m$ |

A parametric function, the *regression model*, that takes the values of the *independent attributes* $a_1, \ldots, a_n$ at input and returns a value for the *dependent attribute* $y$.



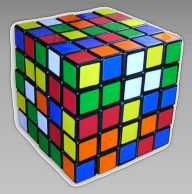
# Regression

**Definition**

Learning from a set of objects how to predict the value of an attribute on an interval scale.

Output:

Estimates of the parameters of the regression model. The regression model with its parameters substituted by their estimates is the *regression function*.



# Regression

**Definition**

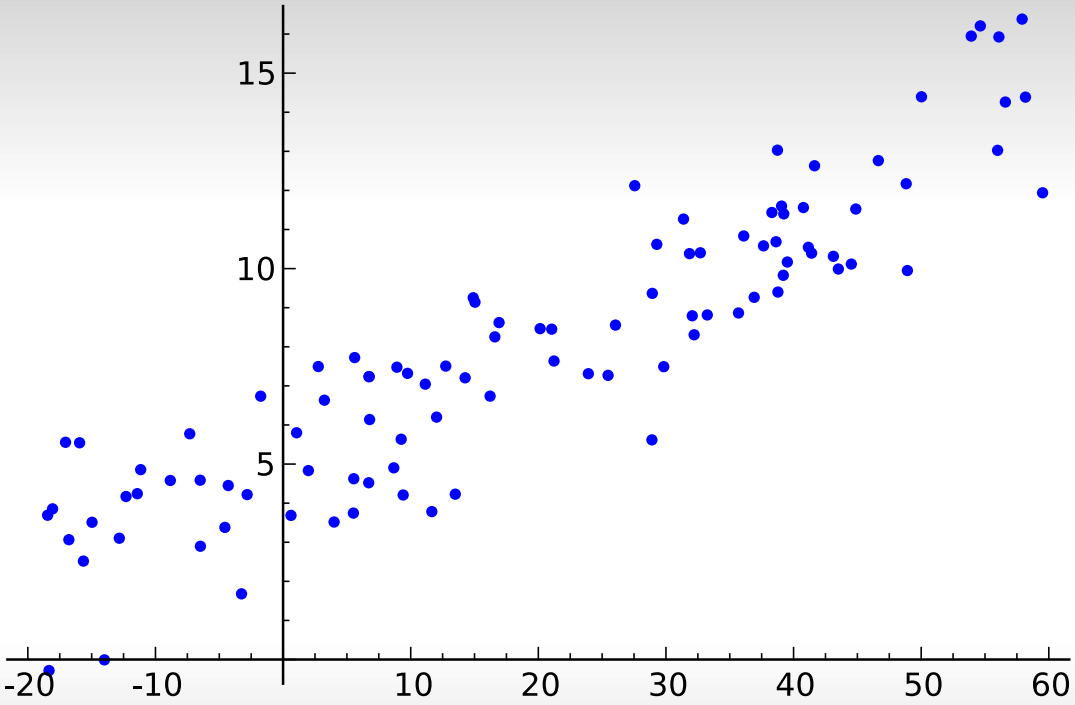
Learning from a set of objects how to predict the value of an attribute on an interval scale.

Output:

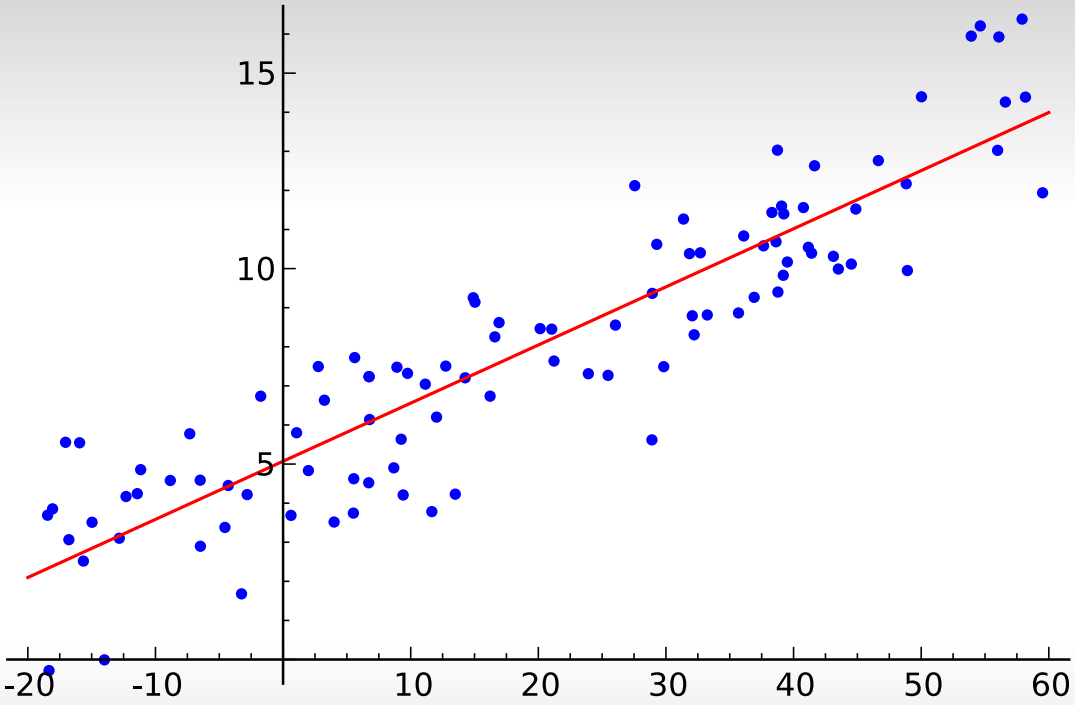
Estimates of the parameters of the regression model. The regression model with its parameters substituted by their estimates is the *regression function*.

If the model is *interpretable* and if no independent attribute is well predicted from the others, the regression function provides an understanding of the effect of every independent attribute on $y$.

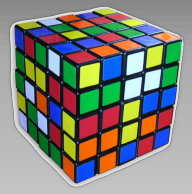
# Regression: graphical interpretation



# Regression: graphical interpretation




Simple linear regression ($y = c_0 + c_1 a + \epsilon$): $\begin{cases} \hat{c}_0 \approx 5 \\ \hat{c}_1 \approx 0.15 \end{cases}$ .



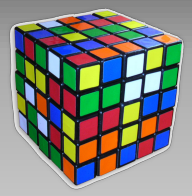


# Making predictions

Applying a regression function to an object, with the value of the dependent attribute missing, predicts it.

Input: a regression function of $a_1, \ldots, a_n$ and

| | $a_1$ | $\ldots$ | $a_n$ | $y$ |
|---|---|---|---|---|
| $o_{m+1}$ | $d_{m+1,1}$ | $\ldots$ | $d_{m+1,n}$ | |
| $o_{m+2}$ | $d_{m+2,1}$ | $\ldots$ | $d_{m+2,n}$ | |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | |
| $o_p$ | $d_{p,1}$ | $\ldots$ | $d_{p,n}$ | |





# Making predictions

Applying a regression function to an object, with the value of the dependent attribute missing, predicts it.

Output:

| | $a_1$ | $\dots$ | $a_n$ | $y$ |
|---|---|---|---|---|
| $o_{m+1}$ | $d_{m+1,1}$ | $\dots$ | $d_{m+1,n}$ | $y_{m+1}$ |
| $o_{m+2}$ | $d_{m+2,1}$ | $\dots$ | $d_{m+2,n}$ | $y_{m+2}$ |
| $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ | $\vdots$ |
| $o_p$ | $d_{p,1}$ | $\dots$ | $d_{p,n}$ | $y_p$ |



# Linear regression

**(Multi-attribute) linear regression model**

$$y = c_0 + c_1 a_1 + \cdots + c_n a_n + \epsilon$$

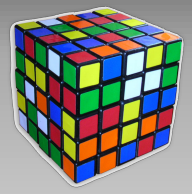


# Linear regression

**(Multi-attribute) linear regression model**

$y = c_0 + c_1 a_1 + \cdots + c_n a_n + \epsilon$

The (weighted) least-square method takes $O(mn^2)$ time to estimate the values of the parameters $c_0$, $c_1$, $\ldots$, $c_n$ minimizing $\left\|(\epsilon_1, \epsilon_2, \ldots, \epsilon_m)^T\right\|_2$ (respectively $\left\|\left(\frac{\epsilon_1}{\sigma_1}, \frac{\epsilon_2}{\sigma_2}, \ldots, \frac{\epsilon_m}{\sigma_m}\right)^T\right\|_2$).





# Linear regression

**(Multi-attribute) linear regression model**

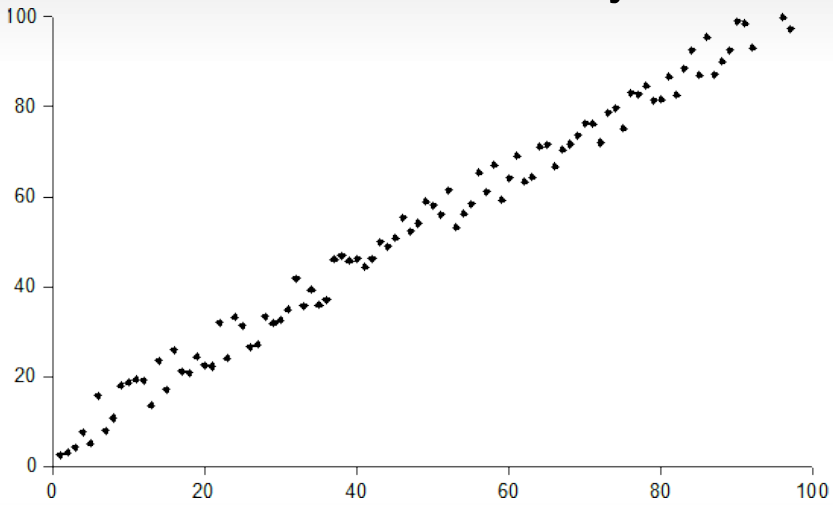
$y = c_0 + c_1 a_1 + \cdots + c_n a_n + \epsilon$

The (weighted) least-square method takes $O(mn^2)$ time to estimate the values of the parameters $c_0$, $c_1$, $\ldots$, $c_n$ minimizing $\left\|(\epsilon_1, \epsilon_2, \ldots, \epsilon_m)^T\right\|_2$ (respectively $\left\|\left(\frac{\epsilon_1}{\sigma_1}, \frac{\epsilon_2}{\sigma_2}, \ldots, \frac{\epsilon_m}{\sigma_m}\right)^T\right\|_2$).

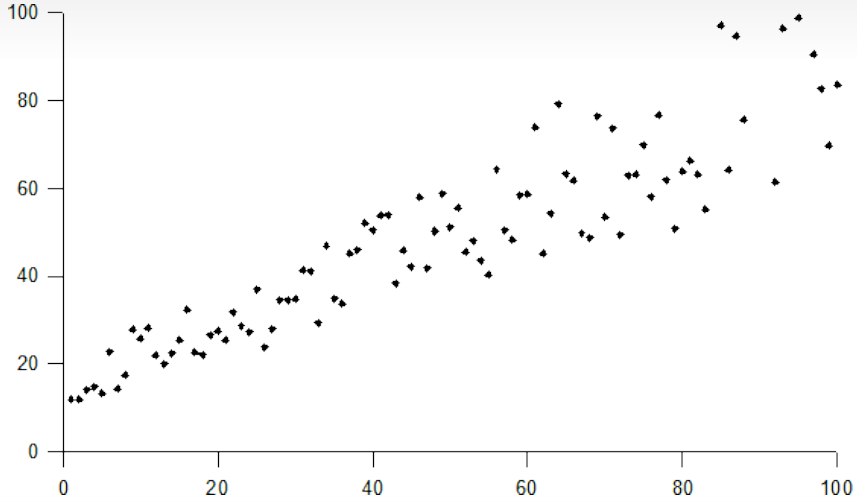
Deriving attributes allows to fit different models. E. g., a linear regression can estimate the parameters $c_0$, $c_1$ and $c_2$ of the model $y = c_0 + c_1 \log a_1 + c_2 \frac{a_1}{a_2} + \epsilon$ after the derivation of the attributes $\log a_1$ from $a_1$ and $\frac{a_1}{a_2}$ from $a_1$ and $a_2$.

# Homo and heteroscedasticity

## Homoscedasticity



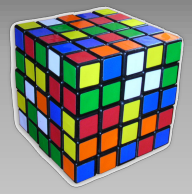

## Heteroscedasticity




Providing standard deviations of the errors leads to a better estimation of the parameters when data show heteroscedasticity.

# Other regression models

Other regression models do not assume that the dependent attribute *linearly* depends on the independent attributes. Some models (e. g., regression trees) are interpretable. Others (e. g., SVR) are not but may provide better predictions.



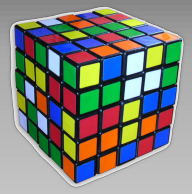


# Other regression models

Other regression models do not assume that the dependent attribute *linearly* depends on the independent attributes. Some models (e. g., regression trees) are interpretable. Others (e. g., SVR) are not but may provide better predictions.

When the dependent attribute is nominal, the regression is said *logistic*. It is also called *supervised classification*. We will study that topic at the end of the semester.

# Outline





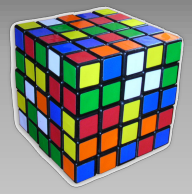
# Goodness of fit

If the chosen model is wrong or if there is no actual dependence between the independent attributes and the dependent attribute, the fit is bad, and so are predictions.

# Goodness of fit

If the chosen model is wrong or if there is no actual dependence between the independent attributes and the dependent attribute, the fit is bad, and so are predictions.

The *coefficient of determination*, in $[0; 1]$, quantifies how much the regression function fits the data. It is the proportion of the variance of the dependent attribute that the regression function explains. It tells nothing about the estimates of the parameters.





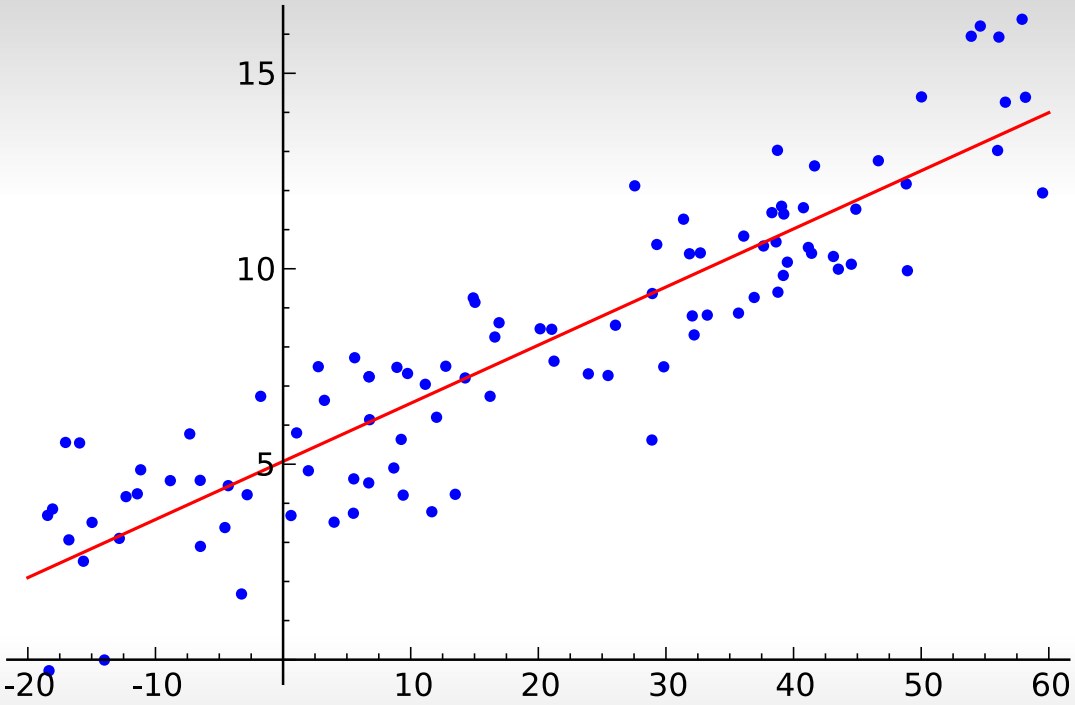
# Goodness of fit

If the chosen model is wrong or if there is no actual dependence between the independent attributes and the dependent attribute, the fit is bad, and so are predictions.

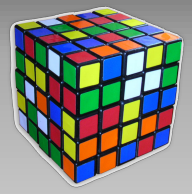
The *coefficient of determination*, in $[0; 1]$, quantifies how much the regression function fits the data. It is the proportion of the variance of the dependent attribute that the regression function explains. It tells nothing about the estimates of the parameters.

With one single independent attribute, the term *correlation* is preferred. Correlation measures are symmetric. They are in $[-1; 1]$, where the sign is that of the estimate of the parameter multiplying the independent attribute.

# Correlation: graphical interpretation




Pearson product-moment correlation coefficient: 0.91 (strong positive correlation).

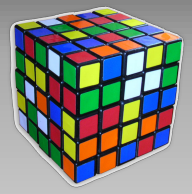




# Over-fitting and over-generalizing

The more parameters in the regression model, the better the fit. However, a model with too many parameters is bad: the regression function over-fits the data and its predictions are imprecise.

**Over-fitting** Fitting too much the data is learning the noise: the regression function poorly generalizes to new objects;





# Over-fitting and over-generalizing

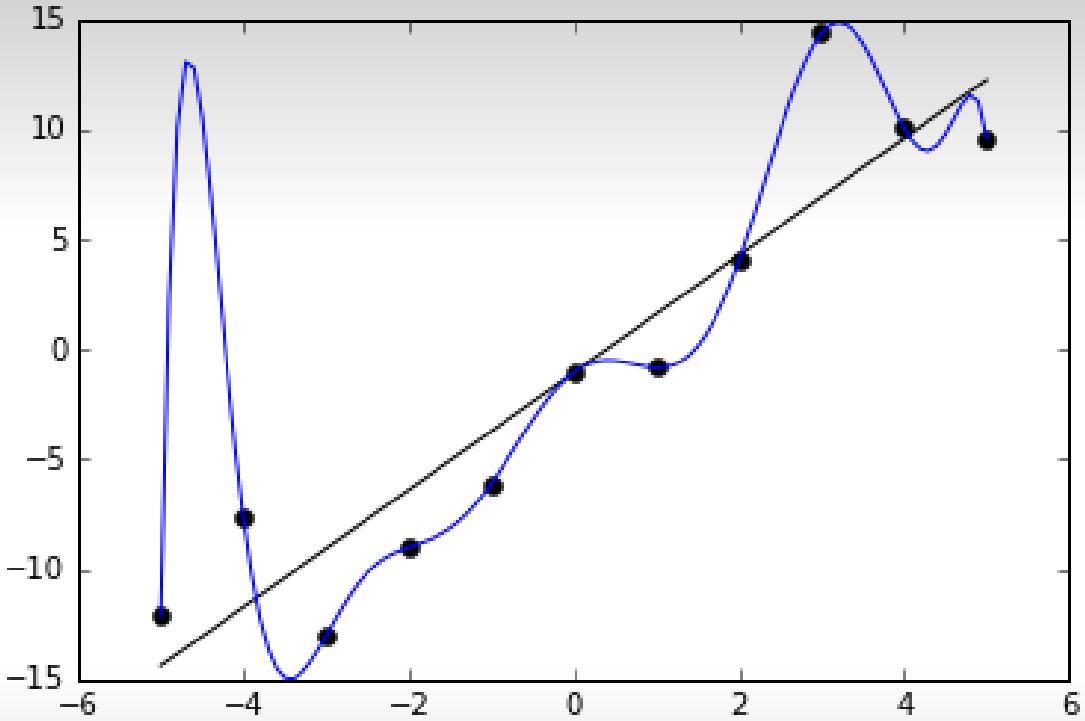
The more parameters in the regression model, the better the fit. However, a model with too many parameters is bad: the regression function over-fits the data and its predictions are imprecise.

**Over-fitting** Fitting too much the data is learning the noise: the regression function poorly generalizes to new objects;

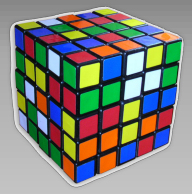
**Over-generalizing** Not fitting enough the data is missing significant details to explain the dependent variable.



# Over-fitting: graphical interpretation

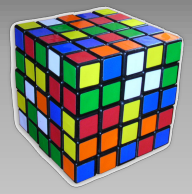



The blue regression function, $y = \sum_{j=0}^{10} \hat{c}_j a^j$, perfectly fits the data but over-fits them. It generalizes worse than the black function, $y = \hat{c}_0 + \hat{c}_1 a$.



# Learning the good amount

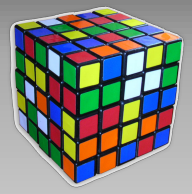
To compare regression models, measures such as the Bayesian and the Akaike information criteria quantify the trade-off between the goodness of fit and the number of parameters.



# Learning the good amount

To compare regression models, measures such as the Bayesian and the Akaike information criteria quantify the trade-off between the goodness of fit and the number of parameters.

*Stepwise regression* is greedily optimizing such a measure to select a model. It is criticized for leading to over-fitting by data dredging.

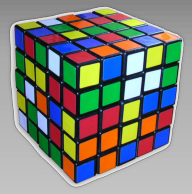


# Learning the good amount

To compare regression models, measures such as the Bayesian and the Akaike information criteria quantify the trade-off between the goodness of fit and the number of parameters.

*Stepwise regression* is greedily optimizing such a measure to select a model. It is criticized for leading to over-fitting by data dredging.

The fit can be computed on objects that are *not* used to estimate the parameters of the regression model. Different techniques will be detailed during the courses about supervised classification.





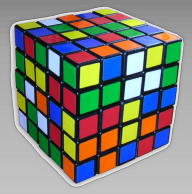
# Popular correlation measures

Correlation measures depend on the type of both attributes:

**nominal** the (adjusted) Rand index;

**ordinal** the *Kendall's tau rank correlation coefficient* and the *Spearman's rho rank correlation coefficient*;

**interval** the *Pearson product-moment correlation coefficient*, $\text{sgn}(\hat{c}_1)\sqrt{r^2}$, where $r^2$ is the coefficient of determination of the regression function $y = \hat{c}_0 + \hat{c}_1 a$.





# Popular correlation measures

Correlation measures depend on the type of both attributes:

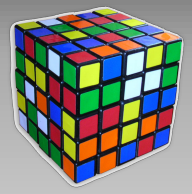
- **nominal** the (adjusted) Rand index;
- **ordinal** the *Kendall's tau rank correlation coefficient* and the *Spearman's rho rank correlation coefficient*;
- **interval** the *Pearson product-moment correlation coefficient*, $\mathrm{sgn}(\hat{c}_1)\sqrt{r^2}$, where $r^2$ is the coefficient of determination of the regression function $y = \hat{c}_0 + \hat{c}_1 a$.

Computing from $m$ objects the (adjusted) Rand index or Kendall's tau takes $O(m^2)$ time; $O(m)$ for Spearman's rho and Pearson's r.



# Pearson correlation coefficient

If the two attributes are not linearly dependent, the Pearson correlation coefficient does not apply. However, ignoring the error term, nonlinear models can be turned linear. E. g., $y = c_0 e^{c_1 a}$ becomes $\ln y = \ln c_0 + c_1 a$, a linear dependence between $a$ and $\ln y$.



# Pearson correlation coefficient

If the two attributes are not linearly dependent, the Pearson correlation coefficient does not apply. However, ignoring the error term, nonlinear models can be turned linear. E. g., $y = c_0 e^{c_1 a}$ becomes $\ln y = \ln c_0 + c_1 a$, a linear dependence between $a$ and $\ln y$.

Pearson's r is not robust. Kendall's tau and Spearman's rho are robust since they are based on ranks and not on actual values.

# License