Predictive Analytics

Outline

12h Lectures (4*3h)

16h Practical Lectures (4*4h)

- Regression
- Linear Discriminant Analysis
- Classification
  - K-means (Data Mining Lectures)
  - Hierarchical Clustering (Data Mining Lectures)
- SVM
- Decision Trees
- Baysian Networks
- Artificial Neural Networks
  - Multi Layer Perceptron
  - ...

Regression

see Correlations in Data Mining Lecture

Introduction

Calculate a linear regression between two random variables - calculate the equation of the line that represents the relationship between these variables.

Estimation

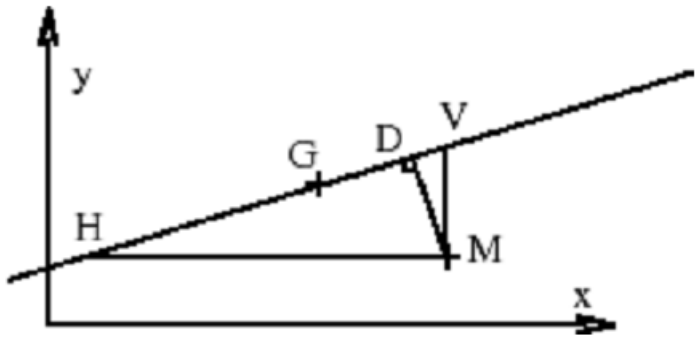
Consider two random variables X and Y strongly correlated. We want to obtain the "best" linear equation $Y = aX + b$ from a cloud of n points, X and Y are the coordinates of these points.

line passes through the centroid $G = \begin{pmatrix} m_x \\ m_y \end{pmatrix}$. So $b = m_y - am_x$.

We also want each item to be as close as possible to the right/line (optimization criterion).

Estimation

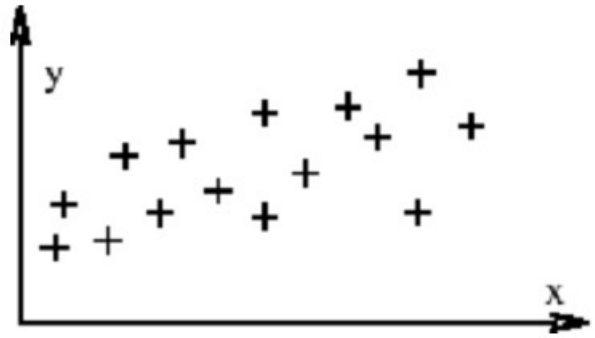
Optimizations criteria :

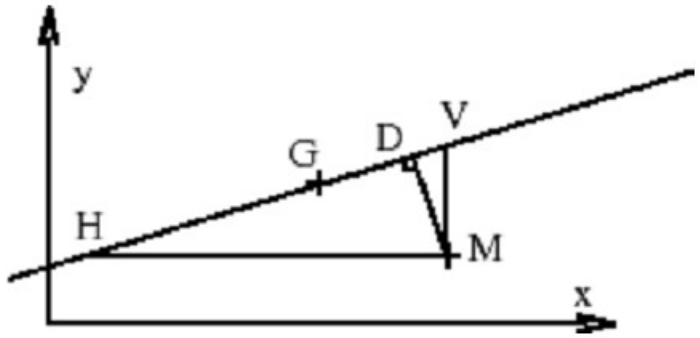


1 : Minimize $\sum MD^2$ (the most intuitive) : distance MD has no physical reality if x and y are not expressed in the same unit. Its optimization is not then nothing severe.

Estimation

Optimizations criteria :



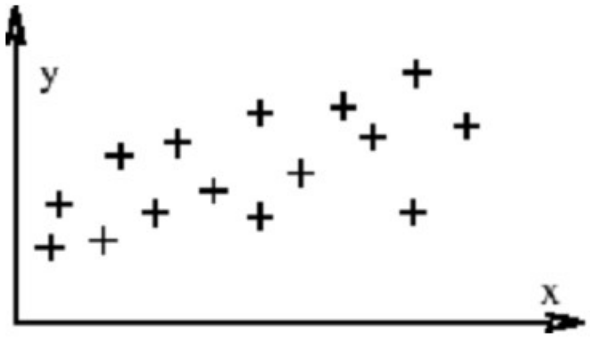


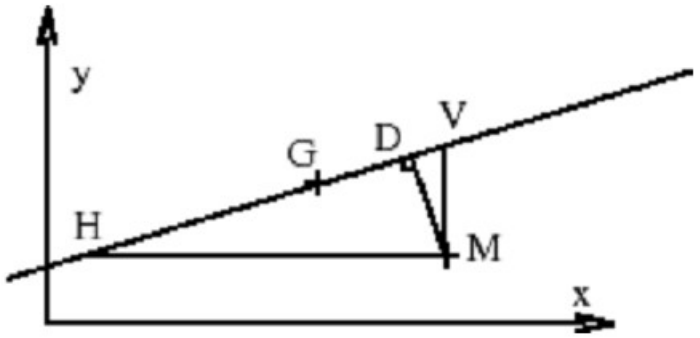
2 : Minimize $\sum MV^2$ (high variance of Y when X fixed, the most used) :

$$a_1 = \frac{Cov(X, Y)}{s_x^2}$$

Estimation

Optimizations criteria :



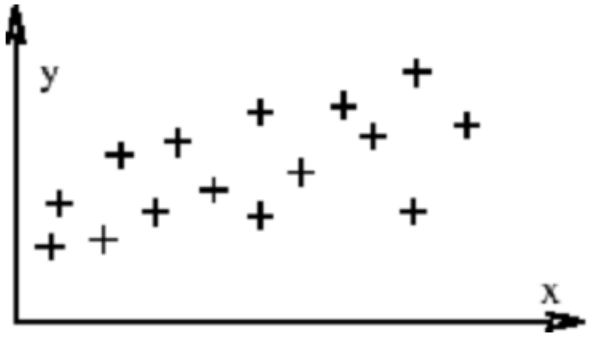


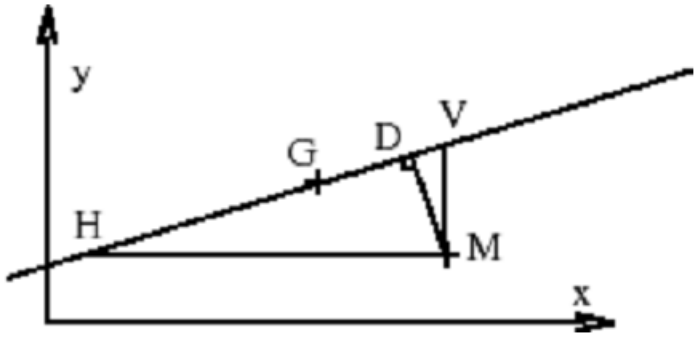
3 : Minimize $\sum MH^2$ (high variance of X when Y fixed) :

$$a_2 = \frac{s_y^2}{Cov(X,Y)}$$

Estimation

Optimizations criteria :





4 : Minimize $\sum MH \times MV$ (high variance of X and Y when the other is fixed) :

$$a_3 = signe(Cov(X, Y)) \times \frac{s_y}{s_x}$$

Interpretation

- If $r^2 = 1$,three criteria are equivalent since by definition neither X nor Y vary greatly when the other is fixed.
- If $r^2 < 1$, must use the knowledge which was on the data to determine which variable is most responsible for the dispersion of the points around the line. In case of doubt is used a^3.
- In any case, if $r^2 << 1$ three criteria differ but the correlation is low and the regression is uninformative. Instead if r^2 is close to 1 the correlation is good, the regression is informative and three criteria are equivalent.

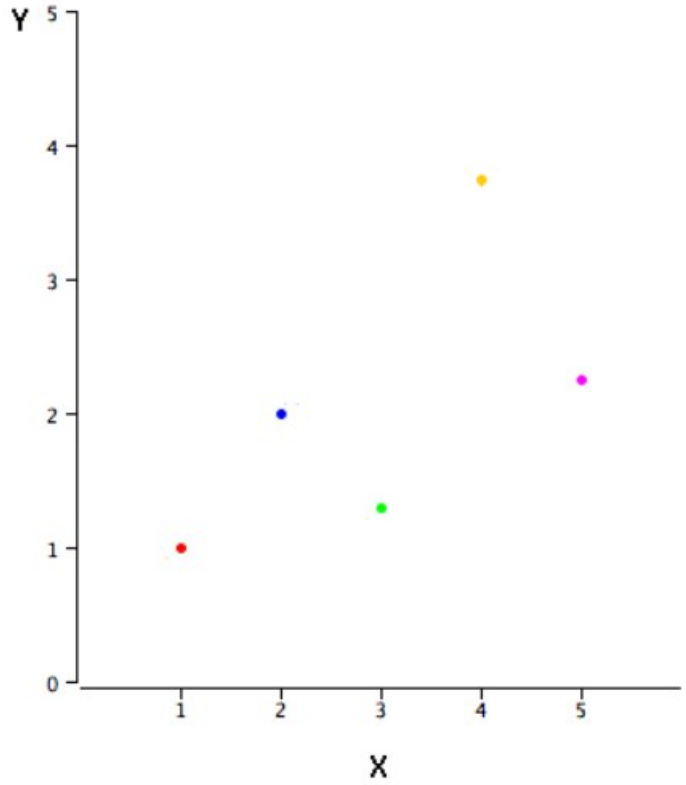
Linear Regression

- In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the criterion variable and is referred to as Y. The variable we are basing our predictions on is called the predictor variable and is referred to as X. When there is only one predictor variable, the prediction method is called simple regression.

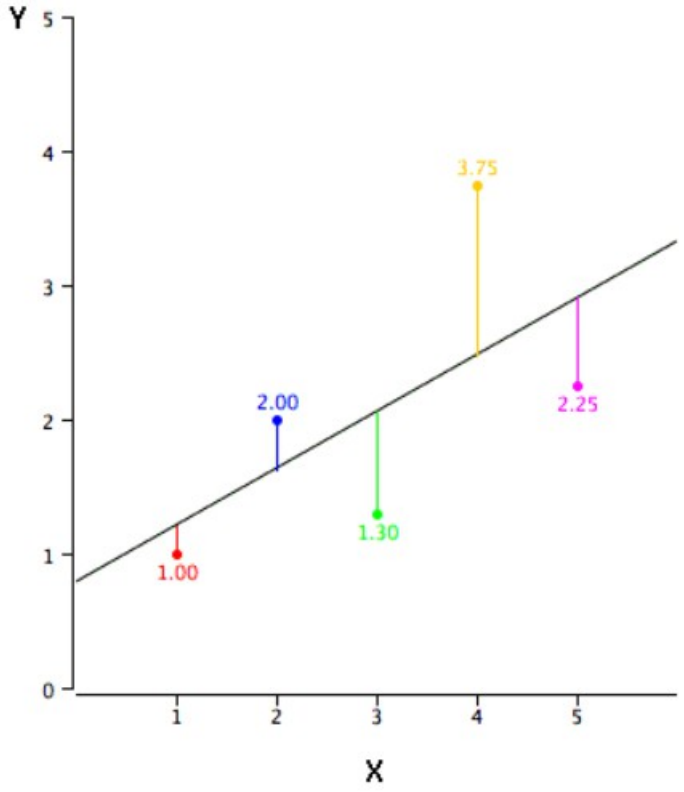
- You can see that there is a positive relationship between X and Y. If you were going to predict Y from X, the higher the value of X, the higher your prediction of Y.

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

Exemple: visualisation of the data



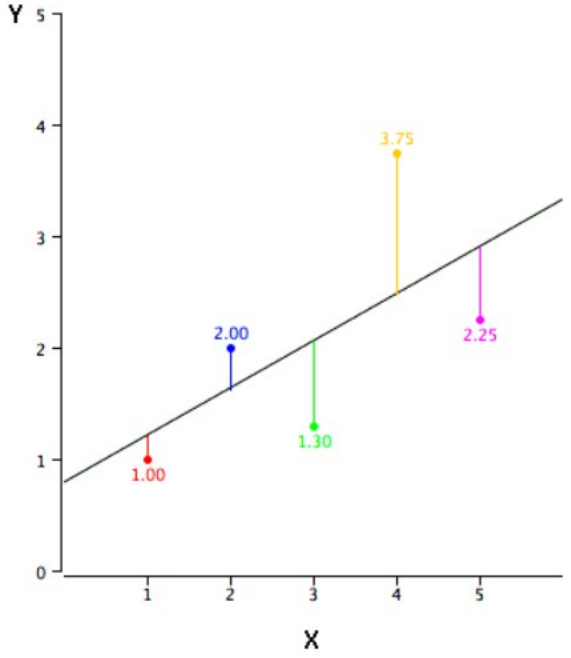
- Linear regression consists of finding the best-fitting straight line through the points.
- The best-fitting line is called a regression line.
- The black diagonal line is the regression line and consists of the predicted score on Y for each possible value of X.
- The vertical lines from the points to the regression line represent the errors of prediction.
- The red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.



The error of prediction for a point is the value of the point minus the predicted value (the value on the line).

Here we shows the predicted values (Y') and the errors of prediction (Y-Y').

For example, the first point has a Y of 1.00 and a predicted Y (called Y') of 1.21. Therefore, its error of prediction is -0.21.



X	Y	Y'	Y-Y'	$(Y-Y')^2$
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436

- The best-fitting line is the line that minimizes the sum of the squared errors of prediction.

The formula for a regression line is

$$Y' = bX + A$$

where Y' is the predicted score, b is the slope of the line, and A is the Y intercept. The equation for the line in Figure 2 is

$Y' = 0.425X + 0.785$

- For X = 1,

$Y' = (0.425)(1) + 0.785 = 1.21.$

- For X = 2,

$Y' = (0.425)(2) + 0.785 = 1.64.$

Computing the Regression Line

The slope (b) can be calculated as follows:

$b = r * sY/sX$

and the intercept (A) can be calculated as

$A = MY - bMX.$

For these data,

$b = (0.627)(1.072)/1.581 = 0.425$

$A = 2.06 - (0.425)(3) = 0.785$

M_X	M_Y	s_X	s_Y	r
3	2.06	1.581	1.072	0.627

X	Y	Y'	Y-Y'	$(Y-Y')^2$
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436