Knowledge Discovery in Databases
TP2 : *Accuracy-precision tradeoff*
Numerical simulations with R

Julien Blanchard

April 2017

1 Simple regression problem

The dataset `1000.data` has 1000 rows described by two numerical variables `x` and `y`. The goal is to study models for explaining `y` from `x`.

1. Visualize the scatterplot (`x`, `y`).
2. Write a function `polynReg(degree, color, nbPoints=20)` which :
   - makes a training dataset by drawing `nbPoints` randomly from the whole dataset (use the `sample()` function),
   - computes the least squares polynomial for explaining `y` from `x` in the training dataset `lm(y ~ poly(x,degree))`
   - plots the polynomial with this `color` on the scatterplot (use `lines()` to draw lines linking the 1000 points).
3. Write a function `polynRepeat(degree, display=FALSE, nbPoints=20)` which calls `polynReg()` 1000 times and, if `display` is true, plots the ten first polynomials on the scatterplot
4. Visualize how the degree (model complexity) affects the accuracy and precision of the models.
5. Write a function `polynMeasures(degree, display=FALSE, nbPoints=20)` which calculates measures to assess precision and accuracy :
   - training accuracy with the MAE[1] on the training set,
   - generalization accuracy with the MAE on the whole population,
   - precision with the MAD[2] of the generated models.
6. Write a function `polynMeanModel(degree, nbPoints=20)` which calculates the average model A from the 1000 generated polynomials, plots A on the scatterplot and calculates the generalization accuracy of A on the whole population with the MAE.
7. Plot the four measures w.r.t. the degree.

[1] Mean Absolute Error $MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$

[2] Mean Absolute Deviation $MAD = \frac{1}{n}\sum_i |y_i - \bar{y}|$