General Mining Issues

a.j.m.m. (ton) weijters

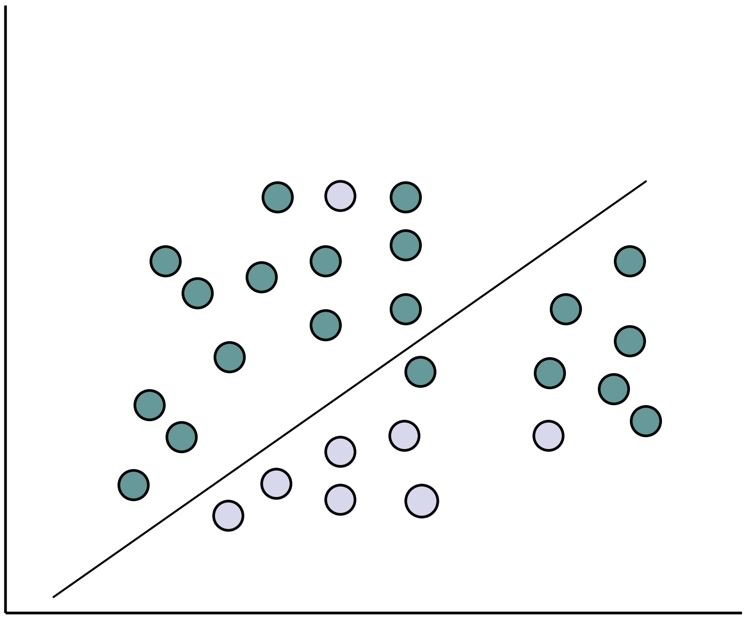
Overfitting
Noise and Overfitting
Quality of mined models

(some figures are based on the ML-introduction of Gregory Piatetsky-Shapiro)

Overfitting

- Good performance on learning material, weak performance on new material.
- Linear regression VS. Artificial Neural Network.
- Decision tree with many leafs VS. decision tree with few leafs.

Classification: Linear Regression

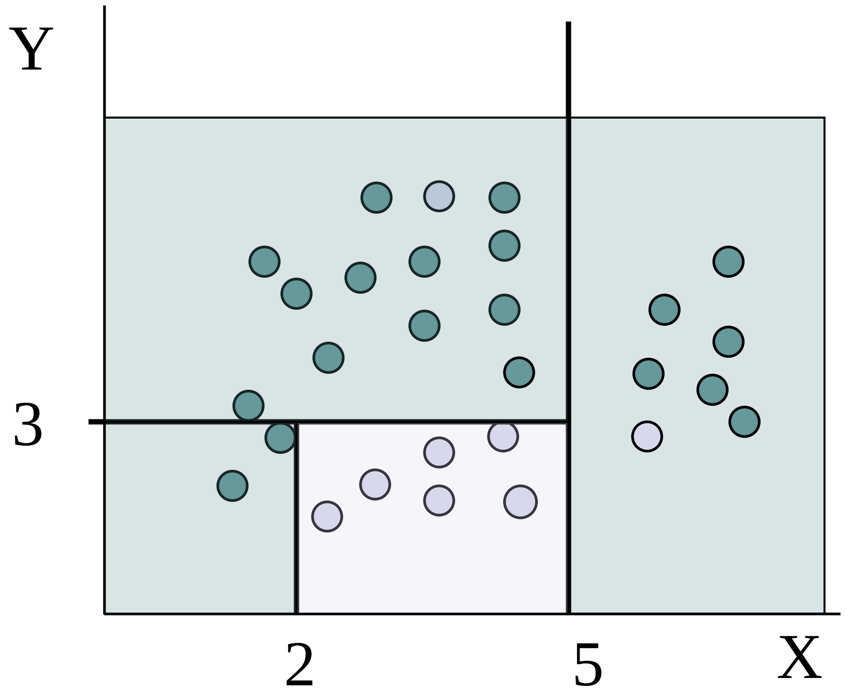


- Linear Regression

  $w_0 + w_1 x + w_2 y >= 0$

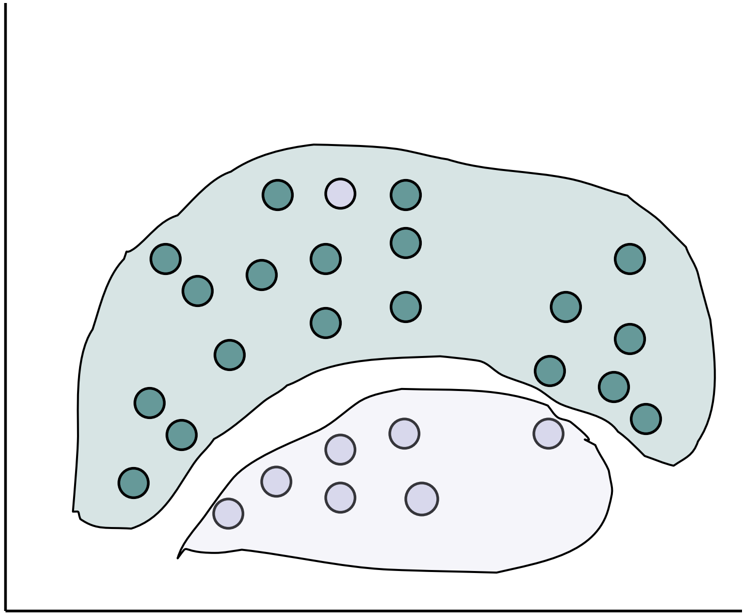
- Regression computes w_i from data to minimize squared error to 'fit' the data
- Not flexible enough

Classification: Decision Trees



if X > 5 then blue
else if Y > 3 then blue
else if X > 2 then green
else blue

Classification: Neural Nets



- Can select more complex regions
- Can be more accurate
- Also can overfit the data – find patterns in random noise

Overfitting and Noise

- Specially the combination of noise (errors) in the learning material and a mined model that attempts to fit all learning material can result in weak models (strong over fitting).

Reliability of a classification rule

- Based on many observations (covering)
- The classification of all the covered cases is correct
- 220/222 rule versus 2/2 rule
- Example of a simple quality measure for classification rules: OK/N+1
  220/222+1 = 0.9865 VS 2/2+1=0.666

Performance of a mined model (always on test material)

- Classification tasks
  - Classification error
  - Classification matrix
  - Weighted classification error
- Estimation tasks
  - MSE

$$\sum_{i=1}^{n}(target_i - result_i)^2$$

- Process Mining ...

K-fold-CV (cross validation) I

- Within the ML community there is a relative simple experimental framework called k-fold cross validation. Starting with a ML-technique and a data set the framework is used
- to build, for instance, an optimal classification model (i.e. with the optimal parameter settings),
- to report about the performance of the ML-technique on this data set,
- to estimate the performance of the definitive learned model, and
- to compare the performance of the ML-technique with other learning techniques.

K-fold-CV (cross validation) II

- In the first step a series of experiments is performed to determine an optimal parameter setting for the current learning problem.
- The available data is divided into k subsets of roughly equal size.
- The ML-algorithm is trained k times. In training n, subset n is used as test material, the rest of the material is used as learning material.
- The performance of the ML-algorithm with a specific parameter setting is the average classification error over the k test sets.
- Based on the best average performance in Step 1, the optimal parameter setting is selected.

K-fold-CV (cross validation) III

- The goal of a second series of experiments is to estimate the expected classification performance of the ML-technique. The available data is again divided in k subsets and again the ML-algorithm is trained k times (in combination with the parameter setting as selected in Step 1).
- The average classification performance on the k test sets is used to estimate the expected classification performance of the ML-technique on the current data set and the T-test is used to calculate a confidence interval.
- If useful, a definitive model is build. All the available material is used in combination with the parameter setting as selected in Step 1. The performance results of Step 2 are used to predict the performance of the definitive model.