Data Mining

Laboratory

Ewa Figielska

WARSAW SCHOOL
OF COMPUTER SCIENCE

Supervised and unsupervised methods

- Data mining methods may be categorized as either **supervised** or **unsupervised**.

- In **unsupervised methods**, no target variable is identified as such. The unsupervised method searches for patterns and structure among all the variables. The most common unsupervised data mining method is clustering.

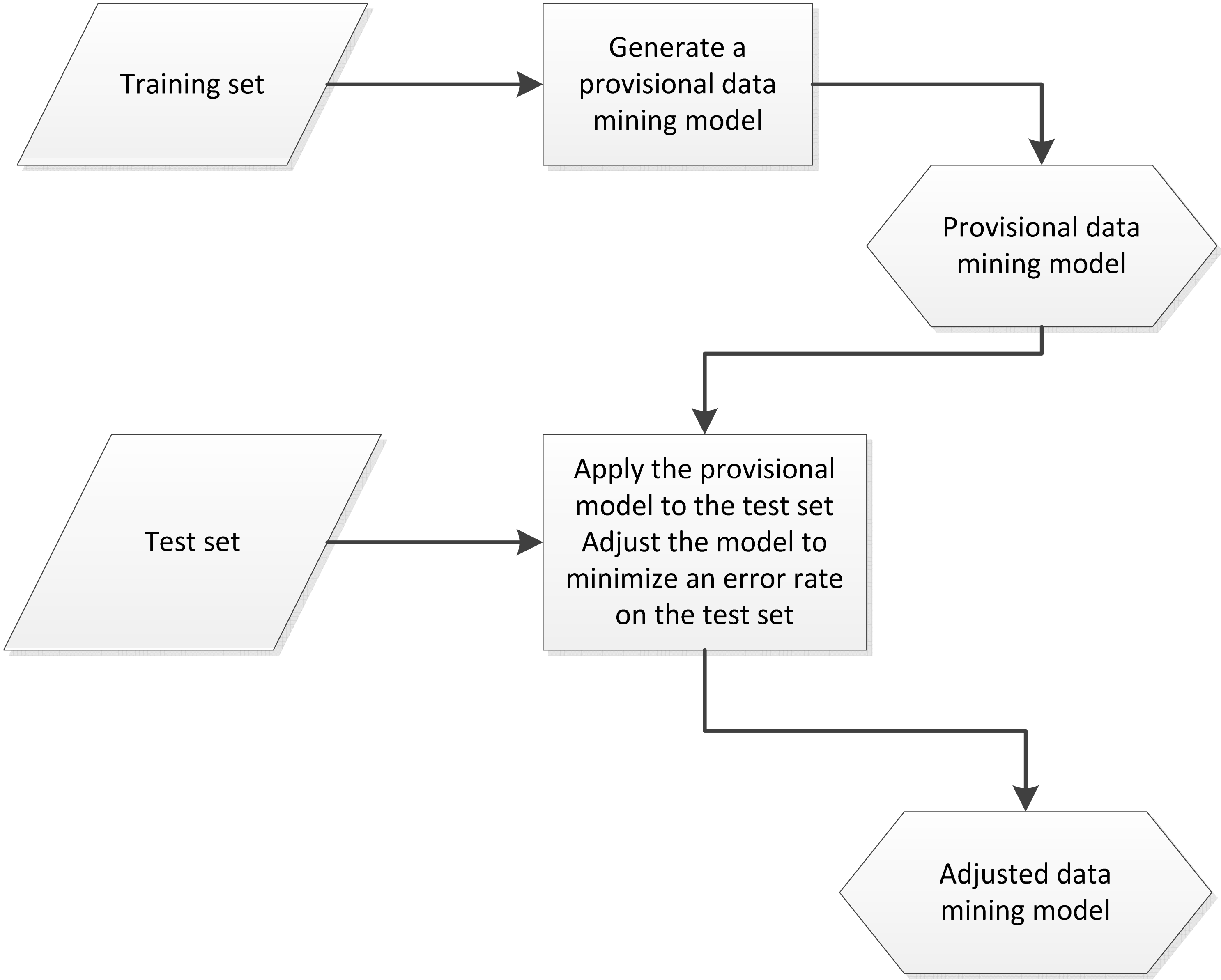
- Most data mining methods are **supervised methods:**
  - there is a particular prespecified **target variable**,
  - the algorithm is given many examples where the value of the target variable is provided, so that the algorithm may learn which values of the target variable are associated with which values of the **input (predictor) variables**. The classification and regression methods belong to supervised methods.

- For example: consider the following excerpt from a data set:

Subject	Age	Gender	Occupation	Income Bracket
001	47	F	Software engineer	High
002	28	M	Marketing consultant	Middle
003	35	M	Unemployed	Low
…				

  - **Target variable:** Income Bracket; **Input variables:** Age, Gender, Occupation
  - The researcher would like to be able to classify the income bracket of a person **not** currently in the database, based on the other characteristics associated with that person, such as age, gender, and occupation (the example of a classification task).

Methodology for supervised modeling



Training set

Generate a provisional data mining model

Provisional data mining model

Test set

Apply the provisional model to the test set Adjust the model to minimize an error rate on the test set

Adjusted data mining model

Cross-validation

- Cross-validation is a technique for insuring that the results uncovered in an analisys are *generalizable* to an *independent, unseen data set.*
- The most common methods:
  - **Twofold cross-validation**
  - **k-fold cross-validation**

- **Twofold cross-validation**:
  - The data are partitioned, using a random assignment, into **training data set** and **test data set**;
    - Training data **include** target variable (the records in the training set need to be preclassified);
    - Test data **do not include** target variable ;
    - Neither training nor test data sets include the “new” or future data that the modelers are interested in classifying.
  - A **provisional data model** is built using the training data set.
  - The provisional data mining model is examined on the test set of data.
    - The efficacy of the classification is evaluated by comparing the model results with the true values of the target variable (which are known for the test data).
    - The provisional data model is adjusted to minimize the error rate on the test set.
  - The performance of the model for the future, unseen data is estimated using various evaluation techniques.

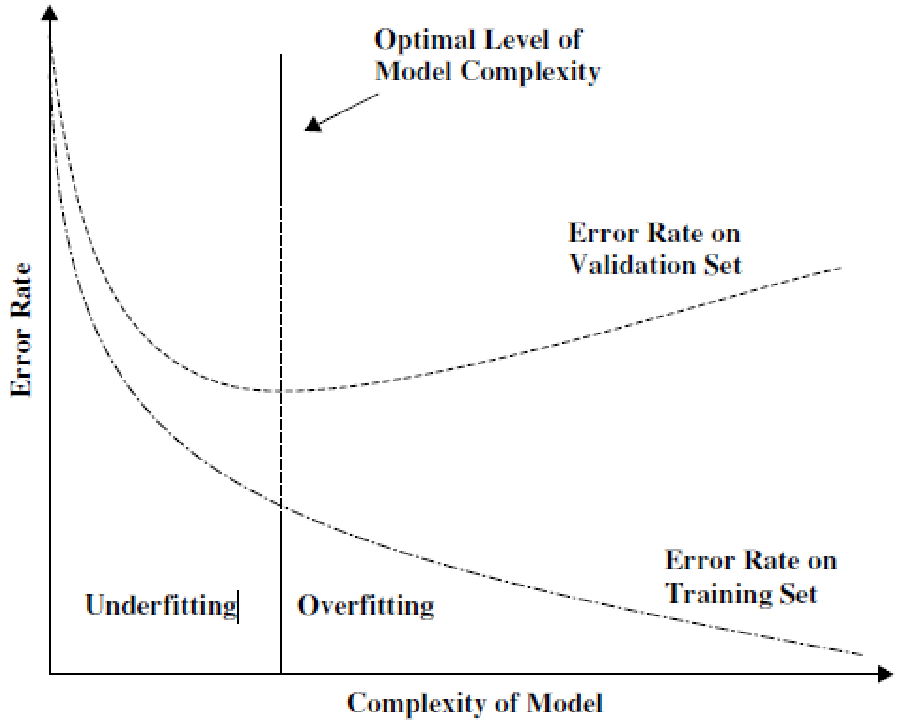
Cross-validation

- **k-fold cross-validation**
  - The original data are partitioned into *k* independent and similar subsets.
  - The model is then built using the data from *k-1* subsets and tested using the *k*th subset.
  - This is done iteratively until we have *k* different models.
  - The results from the *k* models are then combined using averaging or voting.
  - A popular choice for *k* is 10.

- Due to k-fold cross-validation, each record appears once in the test set.

Overfitting

- Often the provisional model is **overfitting** on the training set.



Overfitting (source: Larose “Discovering knowledge in data” (2015))

- At the beginning, the error rates on both the training set and the test set fall, as the provisional model grows in complexity
- As the model complexity still increases, the error rate on the training set continues to fall but the test set error rate begins to increase (because the provisional model has memorized the training set rather than leaving room for generalizing to unseen data).
- The point where the minimal error rate on the test set is encountered is the optimal level of model complexity.
- Complexity greater than this is considered to be overfitting; complexity less than this is considered to be underfitting.