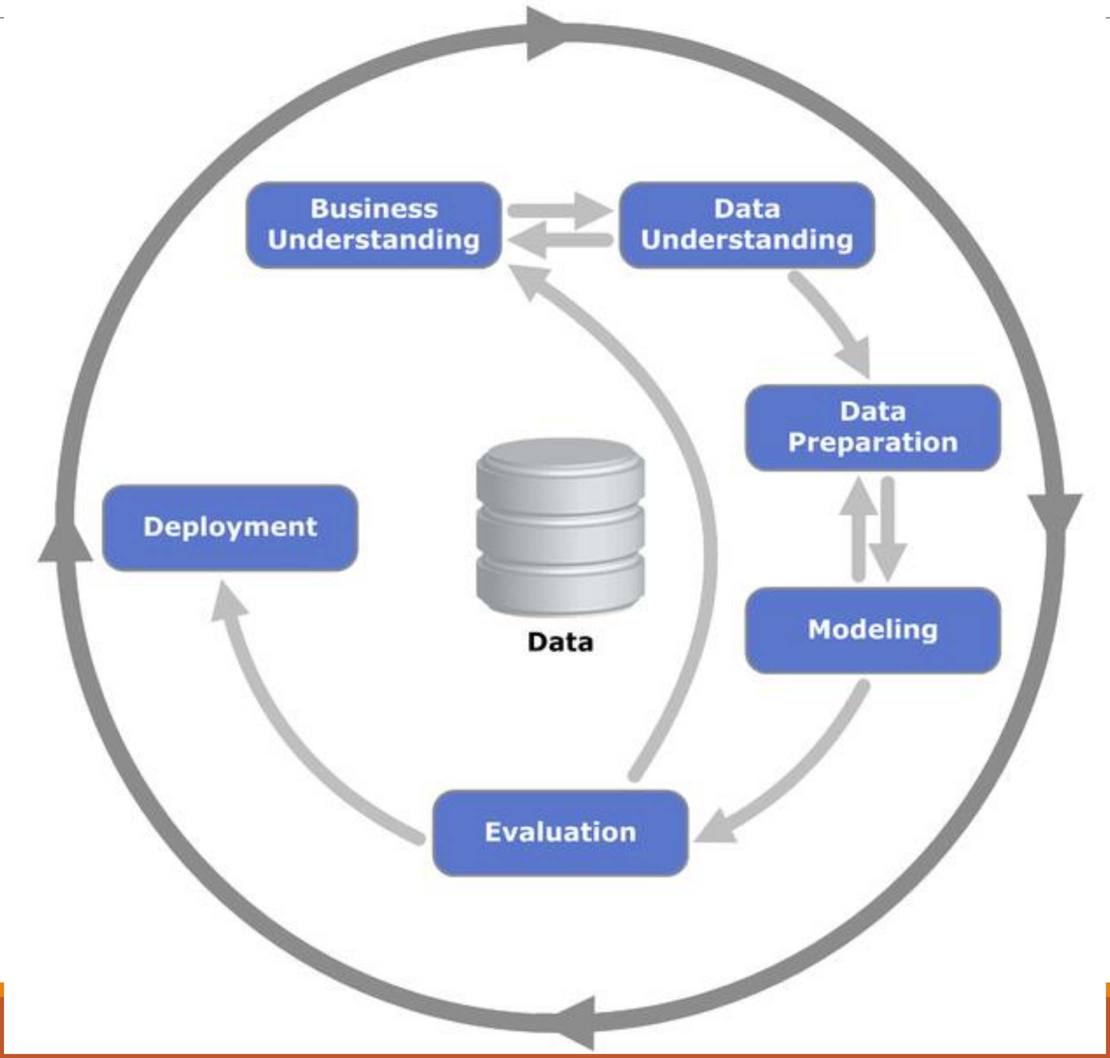
Model performance evaluation

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Supervised vs Unsupervised Data Mining

Supervised – there is a dependent variable – model is built to optimize prediction accuracy

Unsupervised – there is no dependent variable – looking for natural patterns in the data, e.g. K-means clustering

Model types

Prediction:

- Interval or ratio dependent variable

Classification

- Categorical (nominal or ordinal) dependent variable

Prediction models

Ratio dependent variables

- Customer acquisition cost
- Revenue per customer
- Response time
- ROI

Methods

- Regression
- Decision trees
- Neural networks

Classification models

Dependent variable type: ordinal (categories)

- Loan default (yes/no)
- Response to marketing campaign
- Hospital readmission

Methods

- Logit regression for binary categorical dependent variables
- Neural networks
- Bayes

Model building vs. scoring

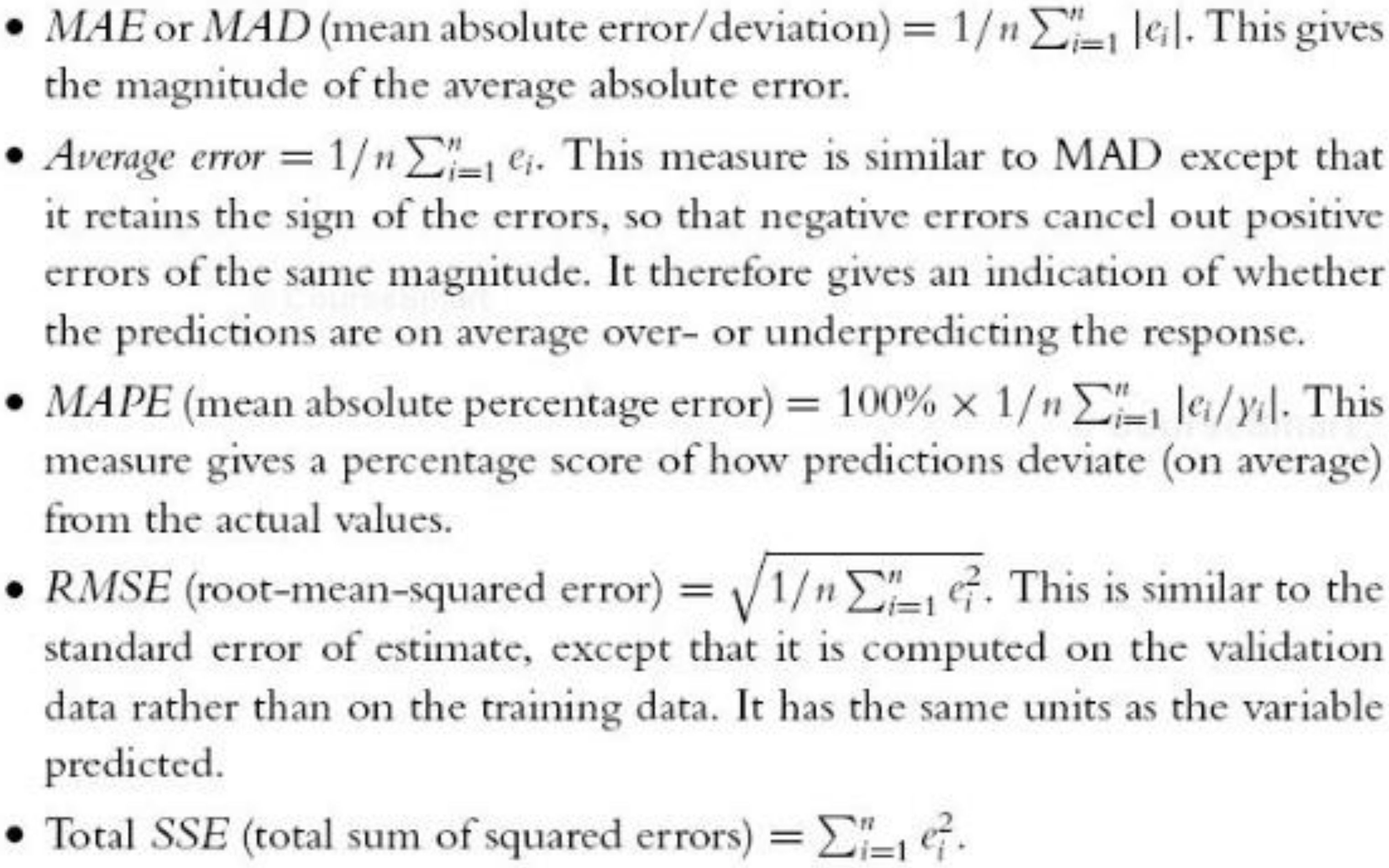
If data permits split into 3 sets

- Training
- Validation
- Test

N-fold validation is commonly used in practice

- Randomly sample training and validation subsets
- Train/validate
- Repeat N times – calculate average accuracy

Prediction model evaluation



- *MAE* or *MAD* (mean absolute error/deviation) $= 1/n \sum_{i=1}^{n} |e_i|$. This gives the magnitude of the average absolute error.
- *Average error* $= 1/n \sum_{i=1}^{n} e_i$. This measure is similar to MAD except that it retains the sign of the errors, so that negative errors cancel out positive errors of the same magnitude. It therefore gives an indication of whether the predictions are on average over- or underpredicting the response.
- *MAPE* (mean absolute percentage error) $= 100\% \times 1/n \sum_{i=1}^{n} |e_i/y_i|$. This measure gives a percentage score of how predictions deviate (on average) from the actual values.
- *RMSE* (root-mean-squared error) $= \sqrt{1/n \sum_{i=1}^{n} e_i^2}$. This is similar to the standard error of estimate, except that it is computed on the validation data rather than on the training data. It has the same units as the variable predicted.
- Total *SSE* (total sum of squared errors) $= \sum_{i=1}^{n} e_i^2$.

Classification model evaluation

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}} = \frac{\text{number of true positives}}{\text{number of positives}}$$

$$= \text{probability of a positive test, given that the patient is ill}$$

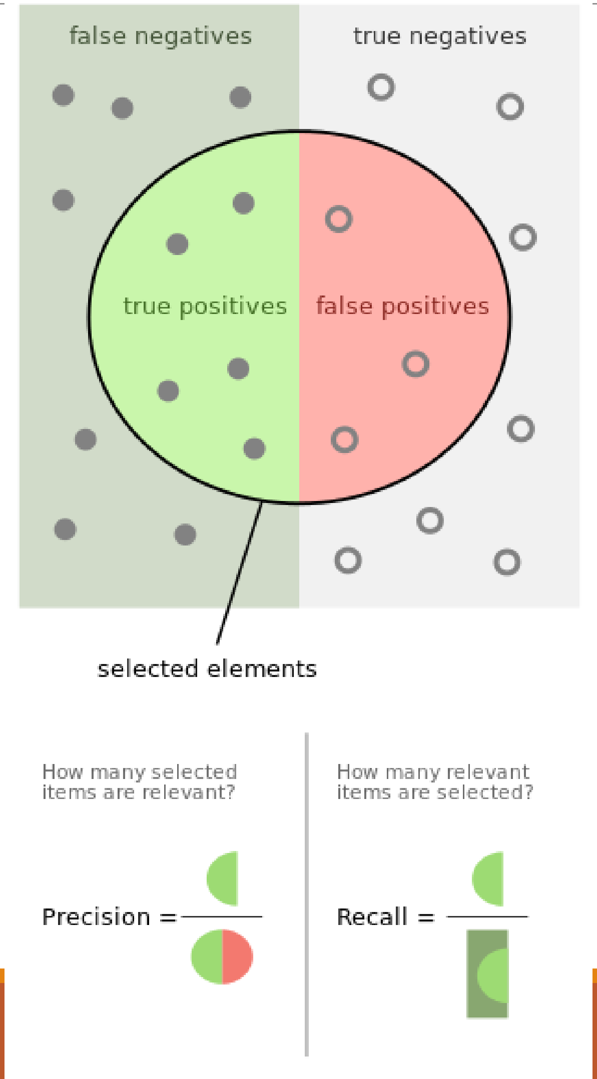
$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}} = \frac{\text{number of true negatives}}{\text{number of negatives}}$$

$$= \text{probability of a negative test given that the patient is well}$$

Precision and Recall

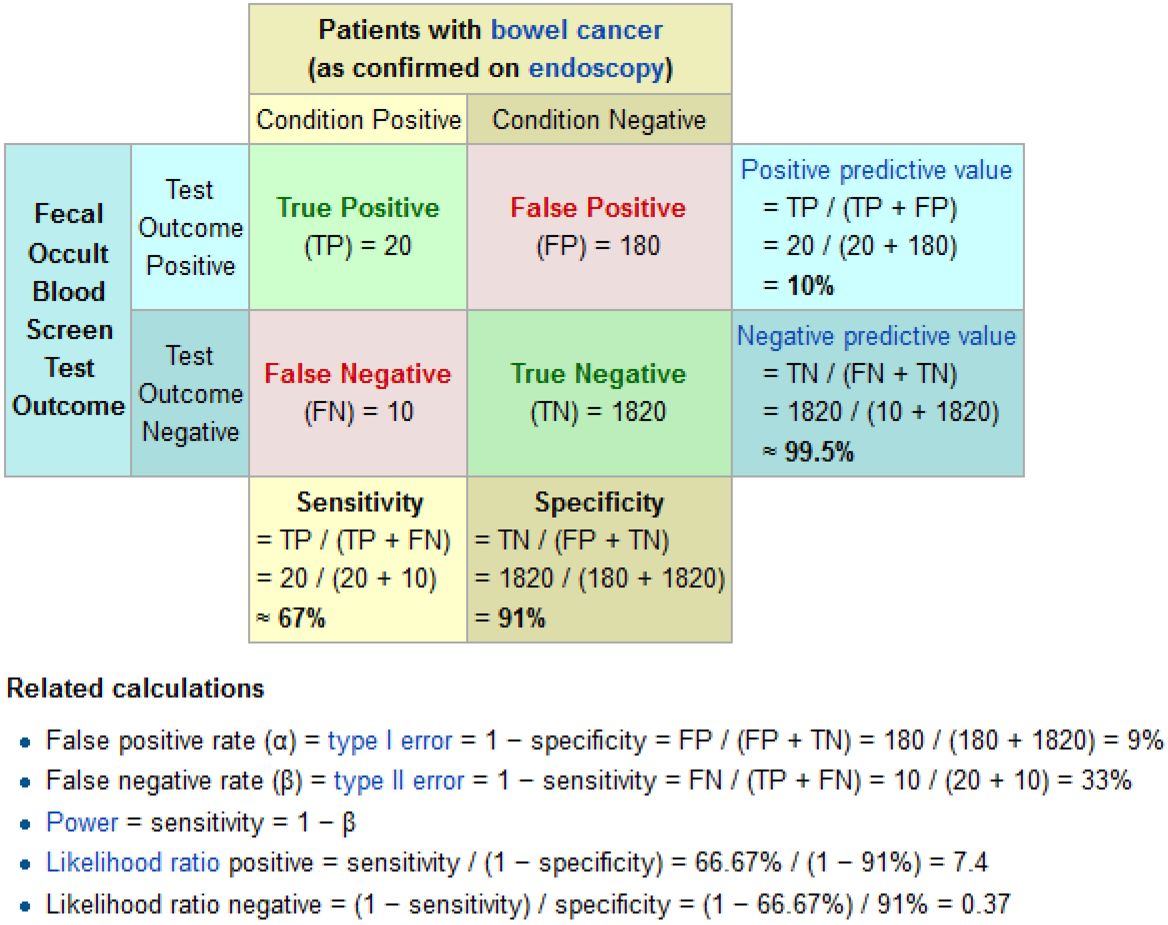
Precision = TP / (TP + FP)

Recall = TP / (TP + FN)



e

Sensitivity/Specificity Example



		Patients with bowel cancer **(as confirmed on endoscopy)**		
		Condition Positive	Condition Negative	
Fecal Occult Blood Screen Test Outcome	Test Outcome Positive	**True Positive** (TP) = 20	**False Positive** (FP) = 180	Positive predictive value = TP / (TP + FP) = 20 / (20 + 180) = **10%**
	Test Outcome Negative	**False Negative** (FN) = 10	**True Negative** (TN) = 1820	Negative predictive value = TN / (FN + TN) = 1820 / (10 + 1820) ≈ **99.5%**
		Sensitivity = TP / (TP + FN) = 20 / (20 + 10) ≈ **67%**	**Specificity** = TN / (FP + TN) = 1820 / (180 + 1820) = **91%**	

Related calculations

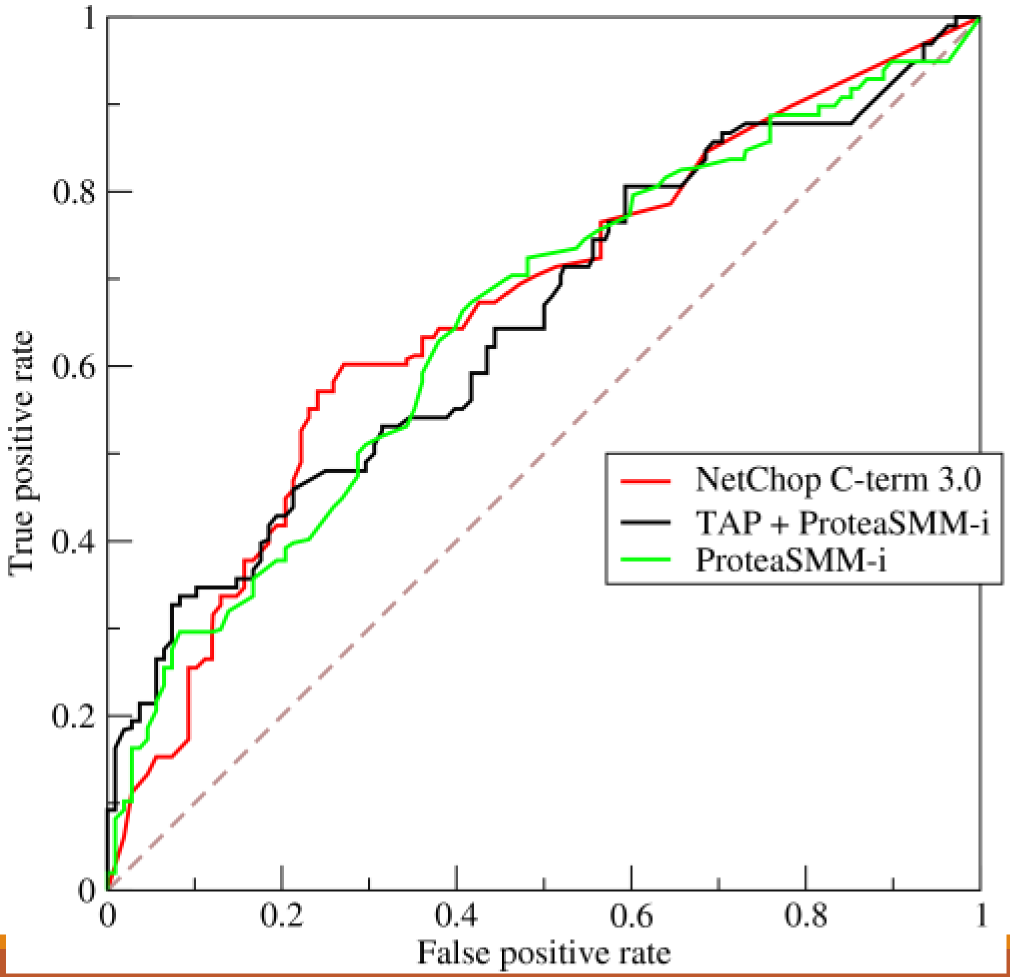
- False positive rate (α) = type I error = 1 − specificity = FP / (FP + TN) = 180 / (180 + 1820) = 9%
- False negative rate (β) = type II error = 1 − sensitivity = FN / (TP + FN) = 10 / (20 + 10) = 33%
- Power = sensitivity = 1 − β
- Likelihood ratio positive = sensitivity / (1 − specificity) = 66.67% / (1 − 91%) = 7.4
- Likelihood ratio negative = (1 − sensitivity) / specificity = (1 − 66.67%) / 91% = 0.37

Classification model performance - ROC graphs

Receiver Operating Characteristics

- Used in signal detection theory
- Tradeoffs in hits vs. false alarms
- Medical diagnosis
- Costs/tradeoffs in type-I, type-II errors

ROC Curve



Predict the probability of class for cases in the validation dataset

Order cases by decreasing probability

Traverse the data, calculate the TP and FP rates at each data point, plot the results

Comparing models – the graph closest to the top left corner is best

AUC – area under the curve

Other ways to evaluate classification model performance

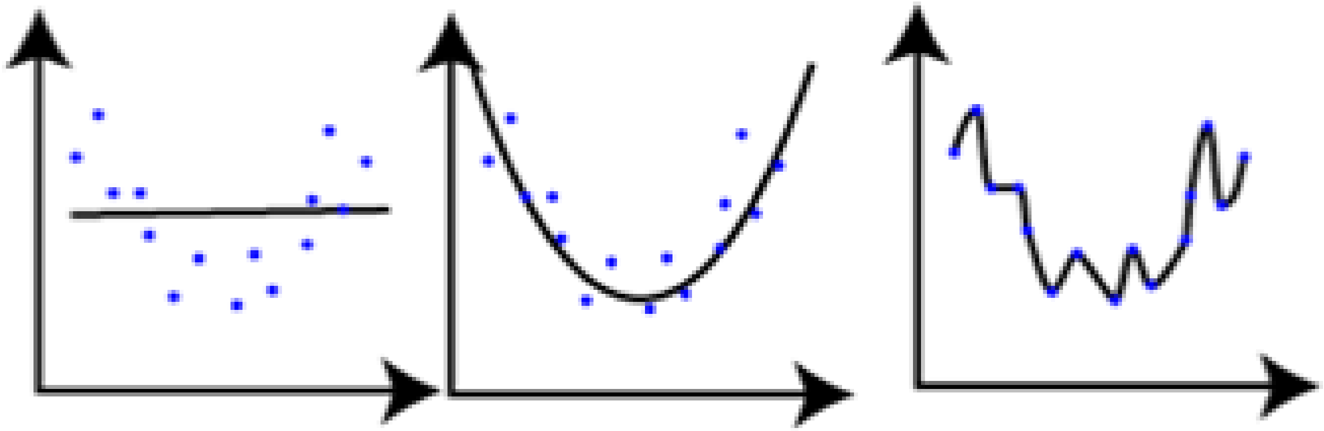
Profit function

- Mailer response ($29), non-response (-$0.50)

Lift – increase in the proportion of the target class

- 2x lift = double the % of respondents versus non-respondents compared to base rate (entire dataset)

The problem of over-fitting

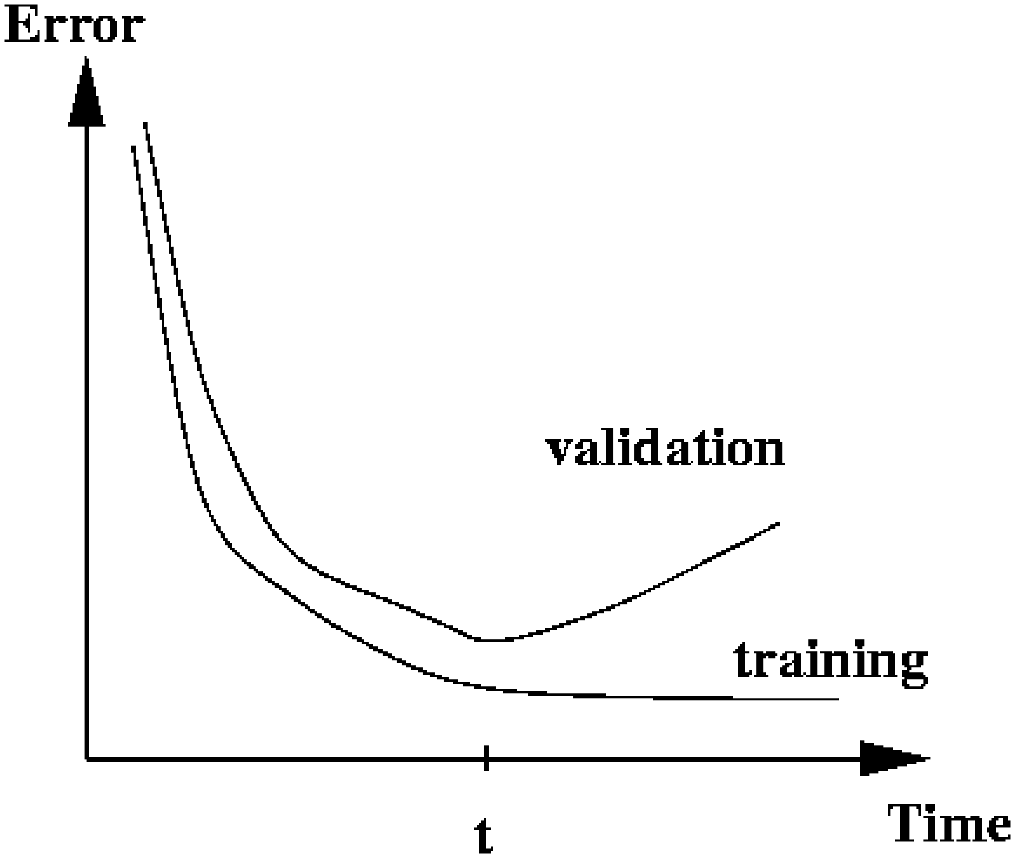


Occam’s razor

Principle of parsimony: among competing hypotheses the one with the fewest assumptions should be selected

Applied to data mining: among competing models with similar predictive power the one with the lowest number of explanatory factors is preferred

Validation data set helps to evaluate model performance


Error
validation
training
t
Time