Feature Selection

Sascha Niro, Prof. Dr. Stephan Trahasch

Offenburg University of Applied Sciences

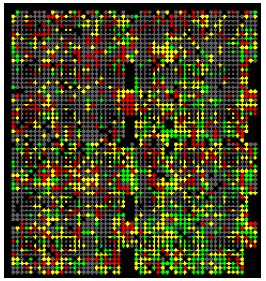
Motivation

Data sets can contain tens or hundreds of thousands of features:

- Text classification:

	I	love	dogs	hate	and	knitting	is	my	hobby	passion
Doc 1	1	1	1							
Doc 2	1		1	1	1	1				
Doc 3					1	1	1	2	1	1

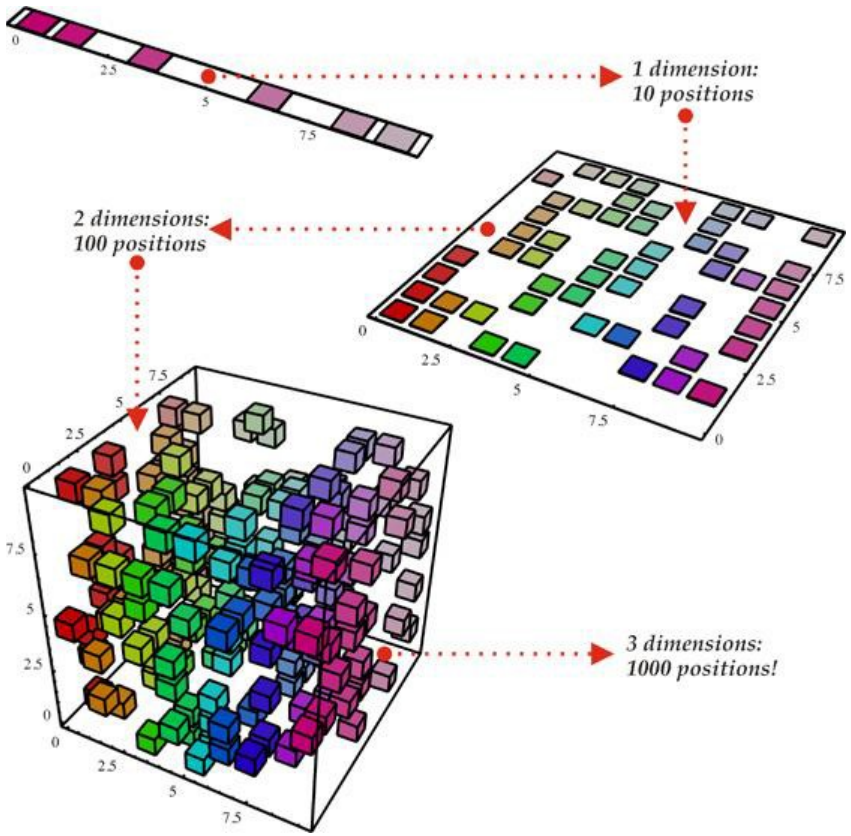
- Gene expression classification:
  m = 6000 … 60000,
  few number of examples (patients)



High dimensional data set

- $m \gg N$ (m: number of features, N: number of instances)

Curse of dimensionality - Richard E. Bellman



- Number of samples needed to describe a d-dimensional space grows exponentially with d
- d binary features, $O(2^d)$ combinations
- Distance functions become meaningless

$$\lim_{d\to\infty} E\left(\frac{\mathrm{dist}_{\max}(d) - \mathrm{dist}_{\min}(d)}{\mathrm{dist}_{\min}(d)}\right) \to 0.$$

- Rule of thumb: 5 training samples per dimension (feature)

Feature Selection goals

- Improve prediction performance (defy curse of dimensionality)
- Facilitate data visualization and data understanding
- Speed up model building process
- Improve understanding of the underlying process
- Reduce storage requirements

Feature Selection: Ranking/Filter Method

Algorithm:

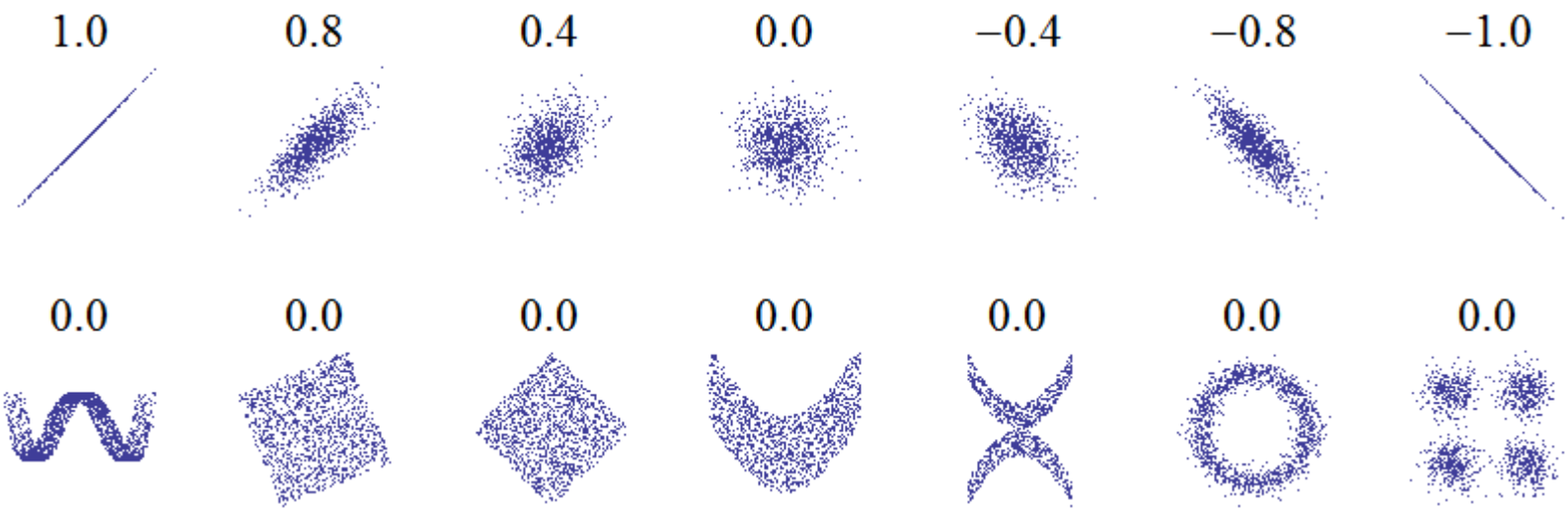
- Calculate the score of every feature x_i against the target variable y
- using a scoring criteria $S(i)$
- Sort the features by that score
- The rank is a used as a measure of variable importance

Rank	Feature	Score
1	B	10
2	A	5
3	C	4

Feature Selection: Ranking - Correlation

Criterion: Pearson correlation coefficient

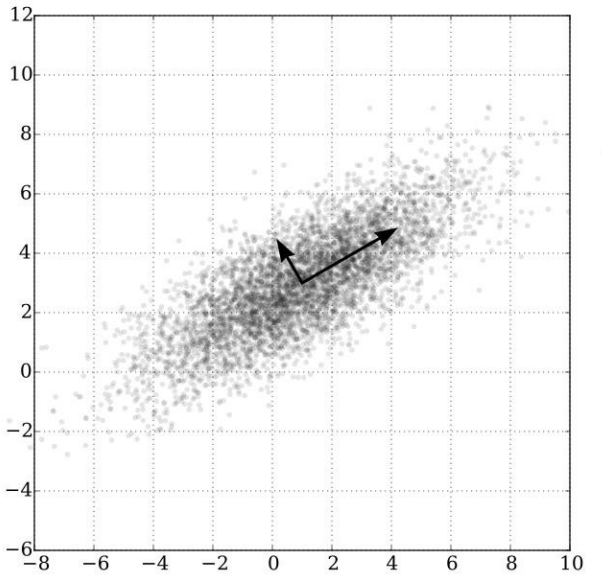
$$Cor(x_i, y) = \frac{cov(x_i, y)}{\sqrt{var(x_i)var(y)}} = \frac{\sum_{i=1}^{m}(x_i - x)(y_i - y)}{\sqrt{\sum_{i=1}^{m}(x_i - x)^2}\sqrt{\sum_{i=1}^{m}(y_i - y)^2}}$$



Principal Component Analysis

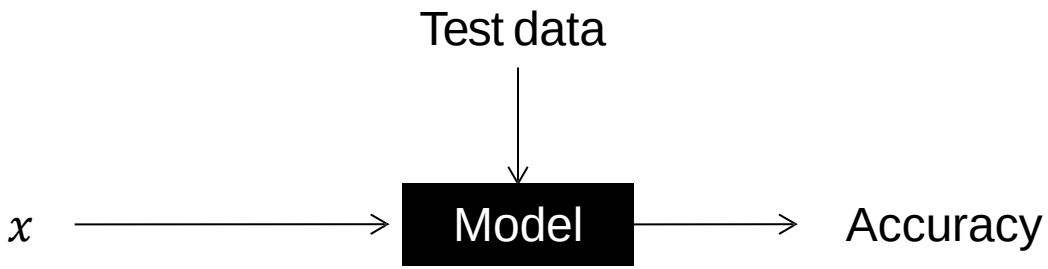
- Converts a set of possibly correlated variables into a set of values of uncorrelated variables called principal components

- Transforms data to a new coordinate system such that the first coordinate (first principal component) captures the greatest variance, the second coordinate the second greatest variance and so on
- Can be used to reduce the dimensionality of a dataset by keeping only the first principal components and discarding the rest
- Unsupervised technique



Ranking – Single Value Classifier

- Criteria: Performance of a classifier built with one variable
- E.g. the value of the variable itself (or its negative)
- Set threshold on the value of the variable
- Predictive power is measured in terms of error rate (or fpr /fnr)
- For regression the RMSE can be used



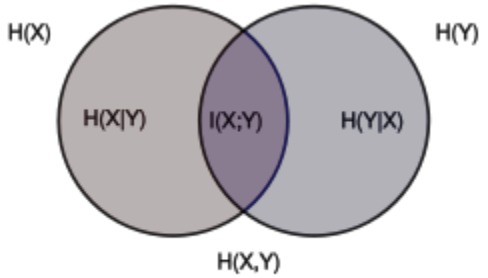
Ranking – Mutual Information

Amount of information obtained about one random variable, through the other random variable

$$I(x_i; y) = \int_{x_i} \int_{y} p(x_i, y) \cdot \log\left(\frac{p(x_i, y)}{p(x_i) \cdot p(y)}\right) \mathrm{d}x \, \mathrm{d}y$$

$$\sum_{x_i} \sum_{\mathrm{y}} P(X = x_i, Y = y) \cdot \log\left(\frac{P(\mathrm{X} = x_i, Y = y)}{P(\mathrm{X} = x_i) \cdot P(\mathrm{Y} = y)}\right)$$

Suitable for non continuos variables, e.g. categorical variables.



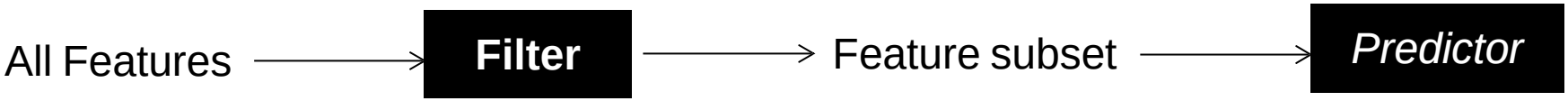
Ranking – Summary

Advantages

- Computationally efficient
- Scales linearly with the number of variables
- Gives an order of variable importance

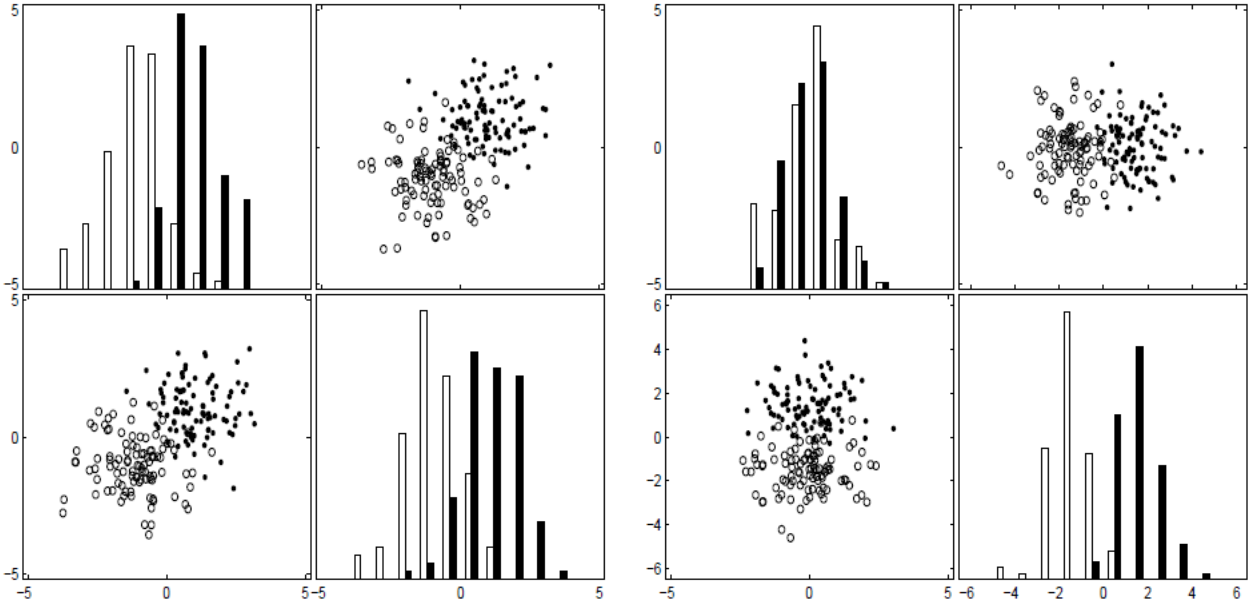
Disadvantages

- Leads to redundant feature sets
- Does not consider the
- relationships between variables



Can redundant variables help each other?

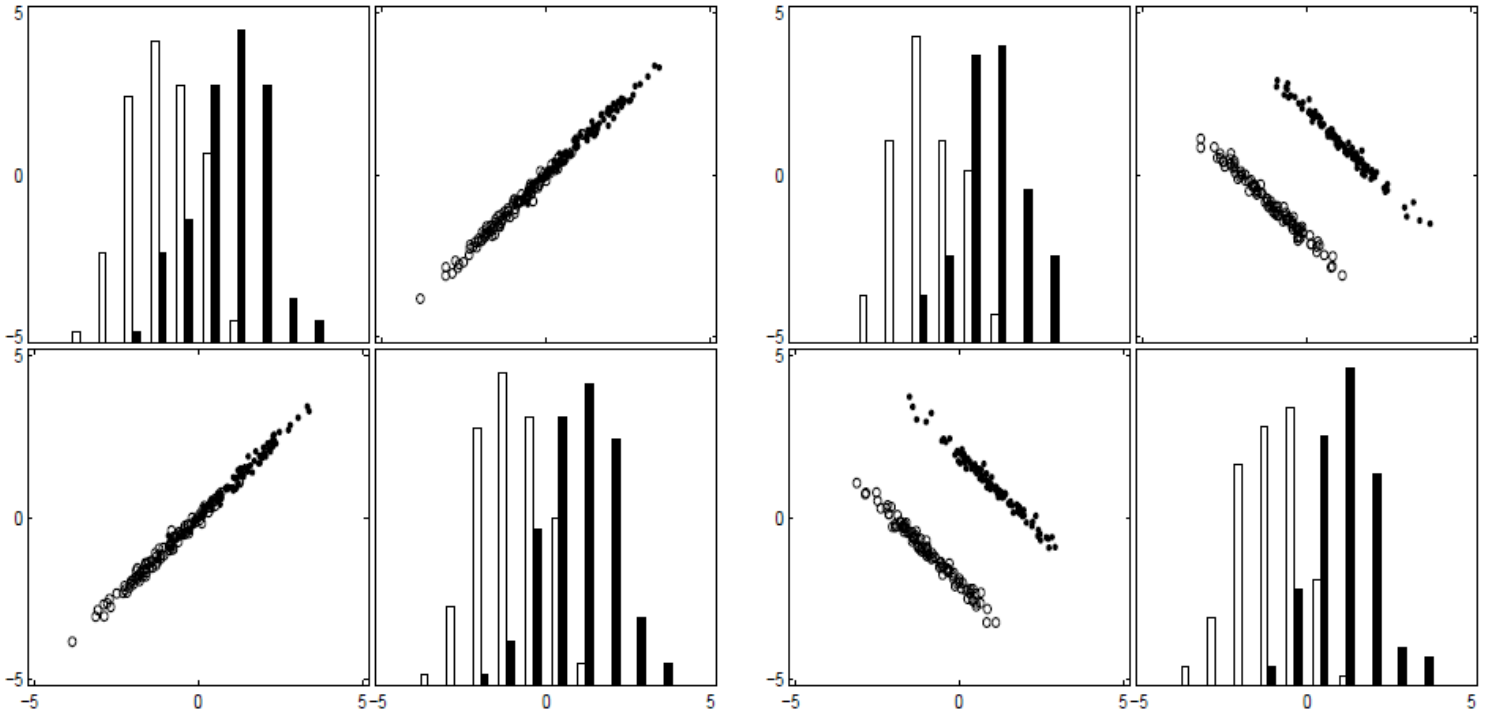
Two class classification problem with two are independently and identically distributed (i.i.d) variables



Noise reduction and consequently better class separation may be obtained by adding variables that are presumably redundant.

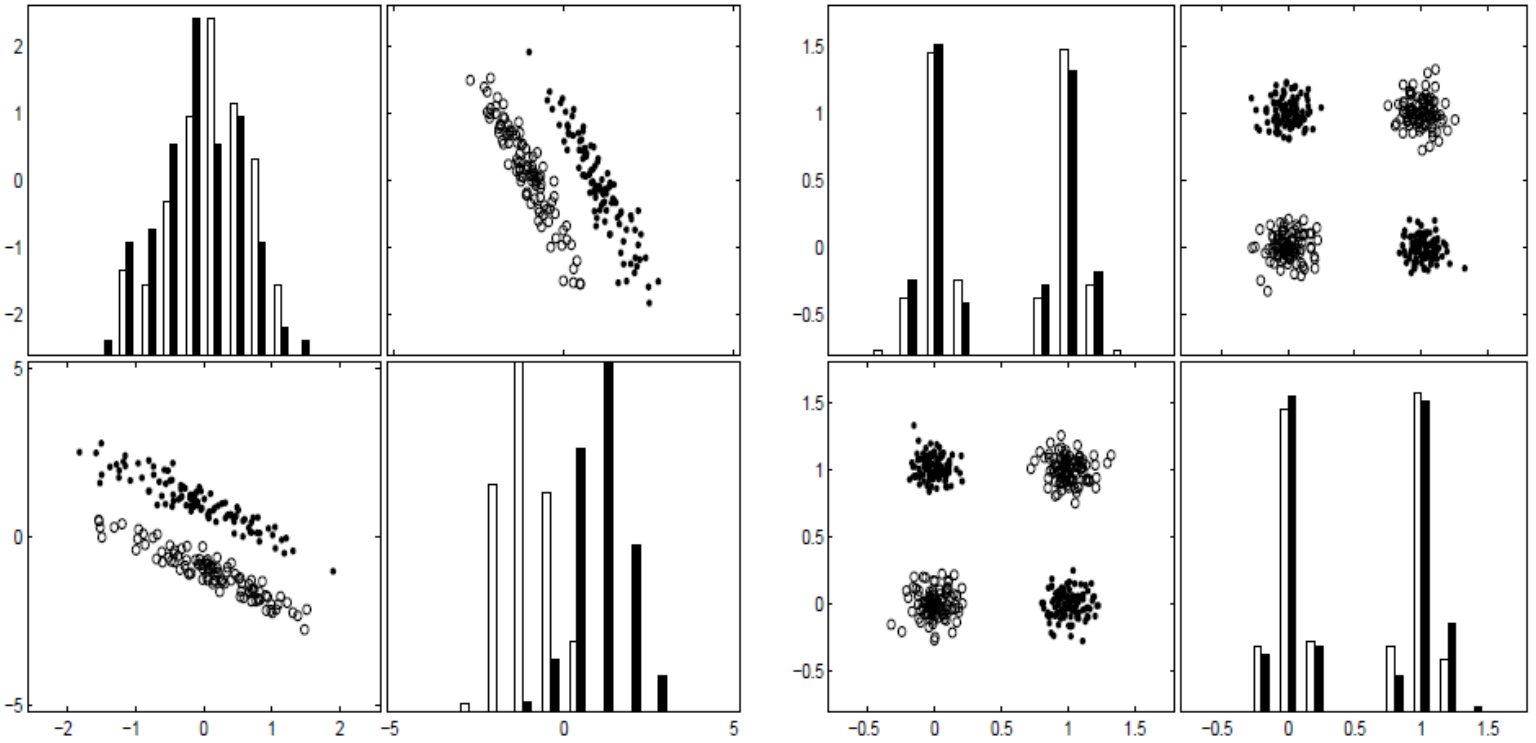
Variables that i.i.d. are not truly redundant.

How does correlation impact variable redundancy?



- Perfectly correlated variables are truly redundant (left)
- Very high variable correlation does not mean absence of variable complementarity (right)

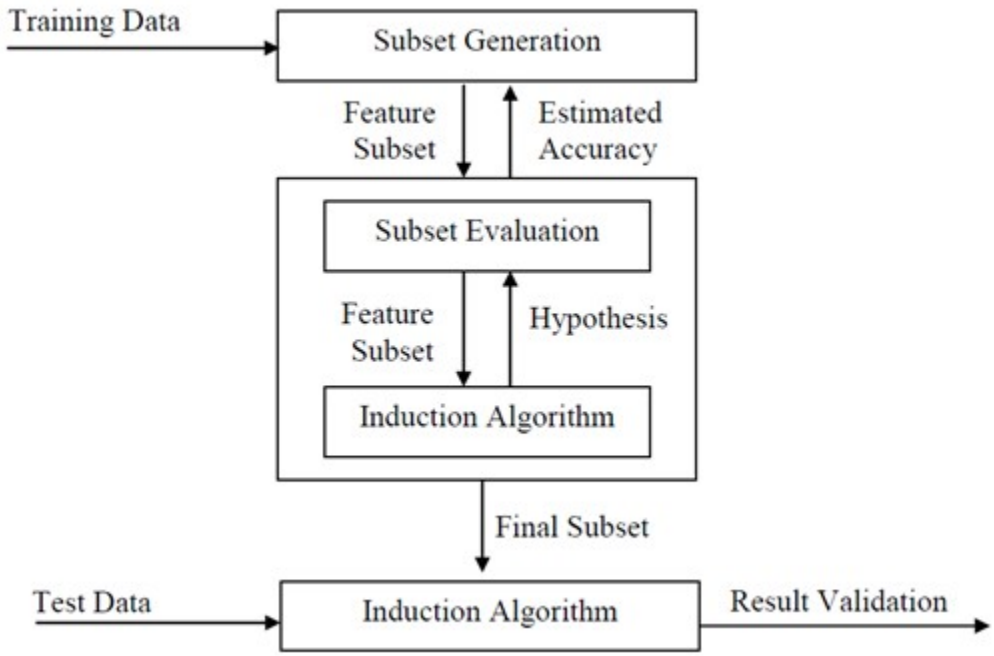
Can a variable that is useless by itself be useful with others?



- A variable that is completely useless by itself can provide a significant performance improvement when taken with others
- Two variables that are useless by themselves can be useful together

Wrapper Method

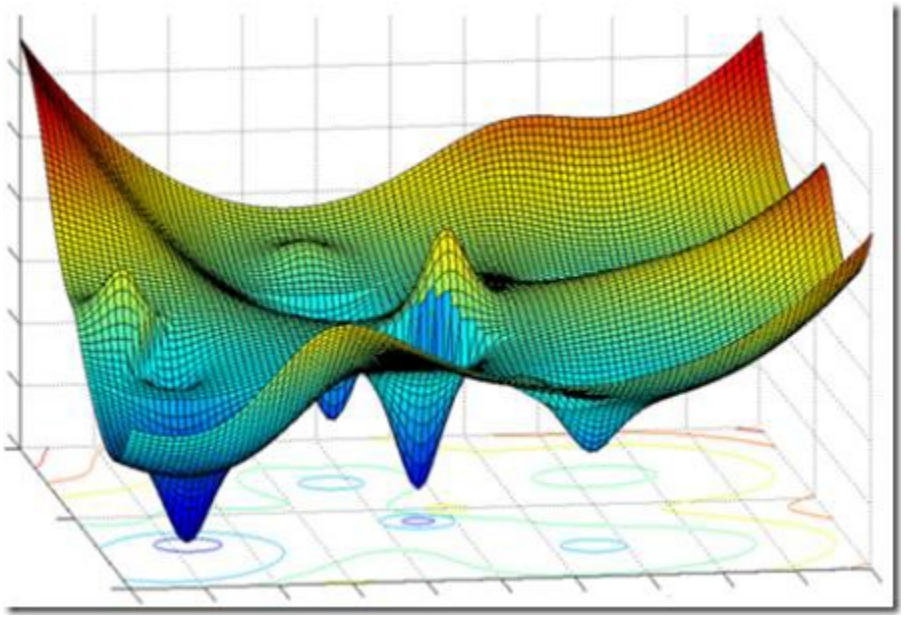
Use the prediction performance to assess the relative usefulness of subsets of variables while performing a search.



Wrapper Method

To be defined:

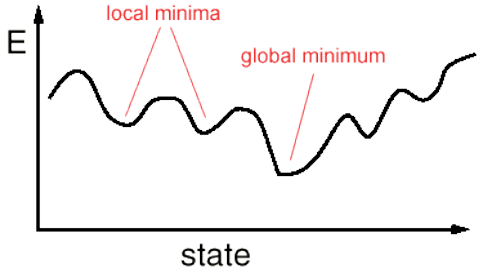
1. Search method
2. How to assess the prediction performance to guide the search and halt it
3. Which predictor to use

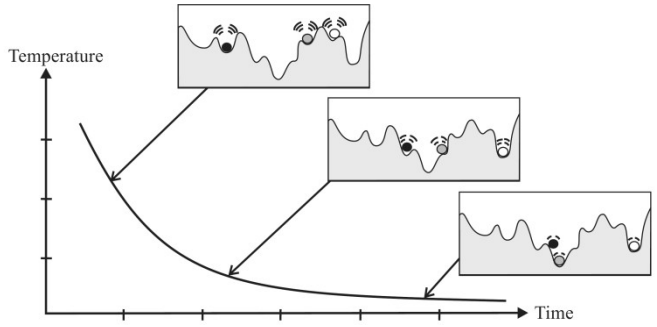


Wrapper Method – Search method

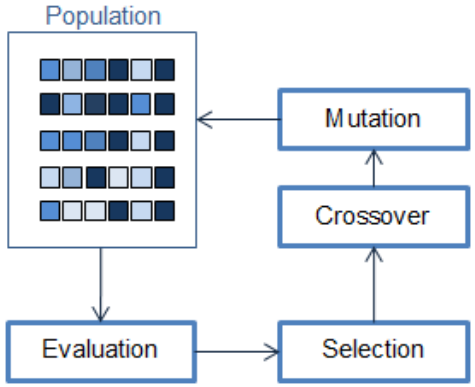
- Exhaustive Search (try all possible feature combinations)?

- Simulated Annealing





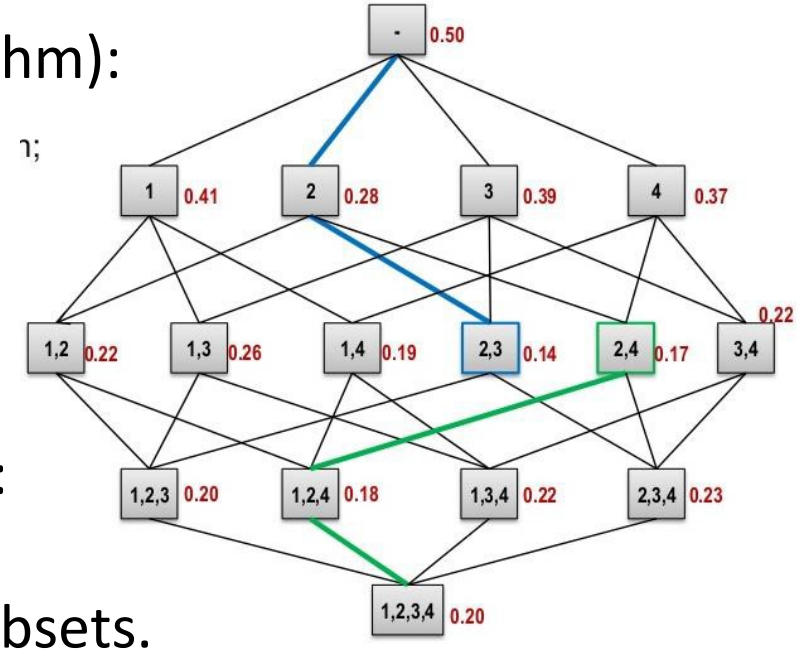
- Genetic algorithms



Wrapper Method – Search method

- Too intensive searching can lead to overfitting
- Simpler and more efficient strategies should be used such as greedy algorithms

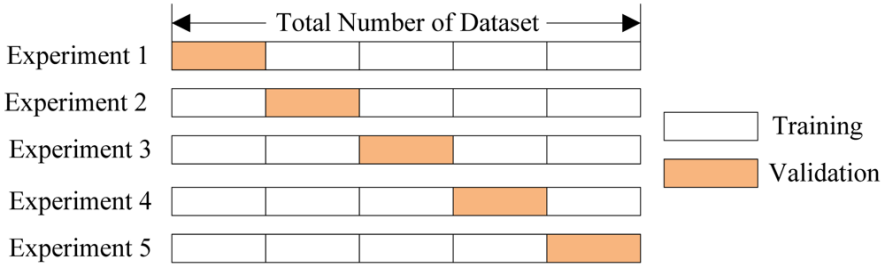
- Backward elemination (greedy algorithm):
  Starts with all features,
  least promising ones are
  progressively elimated.



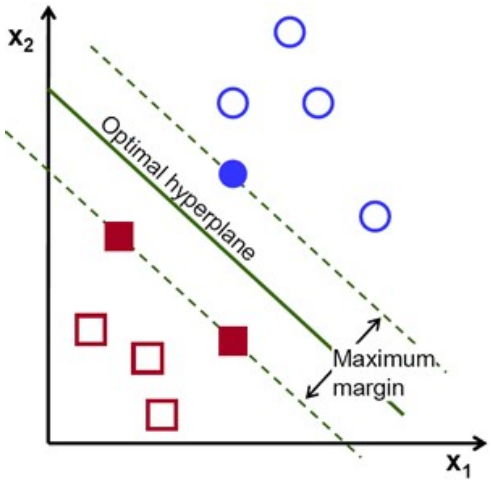
- Forward selection (greedy algorithm):
  Variables are progressively
  incorporated into larger and larger subsets.

Wrapper Method – Performance assessement, Predictor

- Performance Measure:
  - Accuracy on a validation set
  - Crossvalidation



- Predictors:
  - Decision tree
  - Naive Bayes
  - Least-square linear predictors
  - Support Vector Machines



SVM

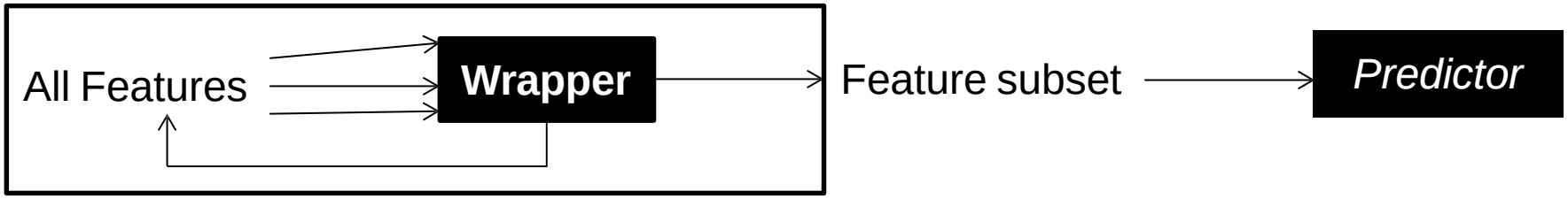
Wrapper Summary

Advantages

- Considers the relationships between variables
- Leads to good feature subsets

Disadvantages

- Computationally expensive
- Brute force method
- Danger of overfitting
- Variance of the feature subsets (Solution: Bootstrapping)



Embedded Method

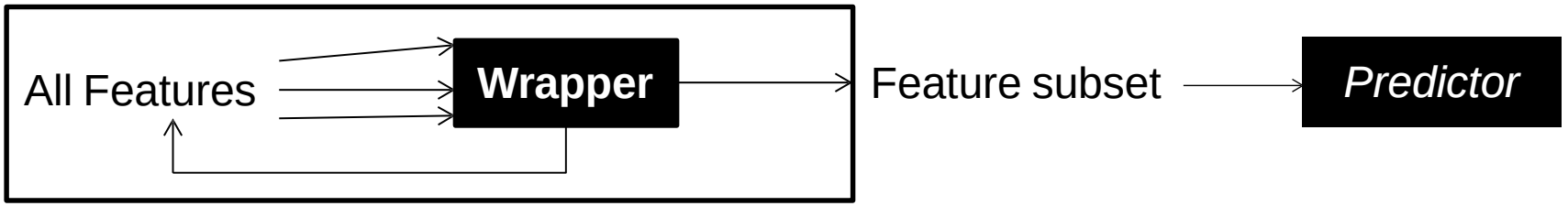
- Feature selection is part of the learning algorithm
- Examples:
  - Decision trees (CART)
  - Random Forest

- Universal and simple approach
- Makes better use of the available data
- More efficient than Wrapper methods

Feature Selection: Summary

- Filter/Ranking:
  Computationally efficient, leads to redundant feature subsets

- Wrapper:
  Computationally more demanding but takes relationship between features into consideration and leads to non redundant feature subsets



- Embedded:

Feature Selection: Lab Task

- Compare the three different feature selection methods using R on a data set of your choice from http://featureselection.asu.edu/datasets.php or the forecasting challenge data set

- Compare the performance of the full feature set vs. the three resulting feature subsets using a predictor of your choice

- Document the feature selection process in an R Markdown document

Literature

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), 1157-1182.