University of Dayton

Feature Selection

CPS 563 – Data Visualization

Dr. Tam Nguyen

tamnguyen@udayton.edu

What are Features and Classes?

Features

Petal length

Petal width

Sepal length

Sepal width

Class

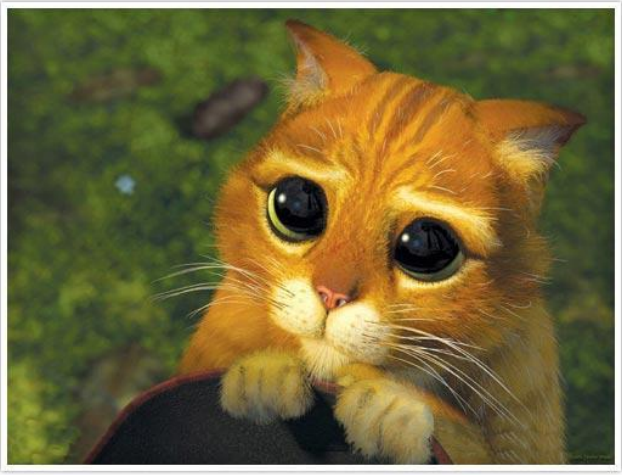
Iris setosa

Iris versicolor

Iris virginica

Another Example

Number of eyes
Number of ears
Number of teeth
Ear shape
Shape of the tail
Sound
Gender
Features



Class

DOG

CAT

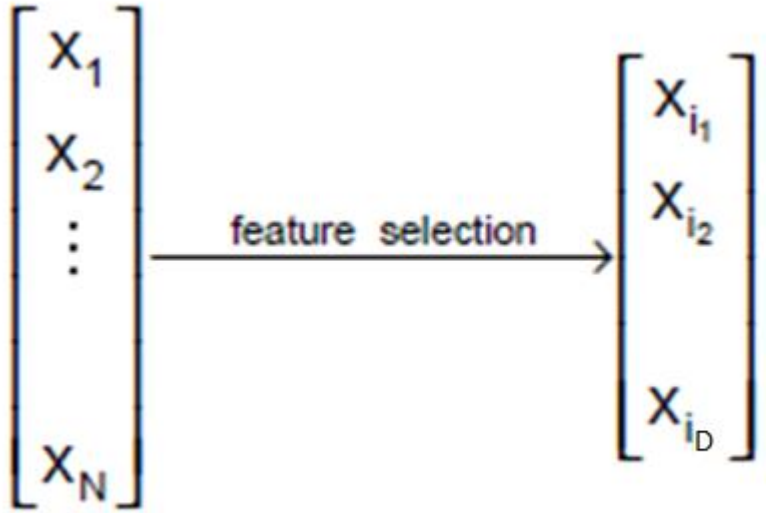
Feature Dimensionality Reduction

- Most machine learning and data mining techniques may not be effective for high-dimensional data
  - **Curse of Dimensionality**
  - Query accuracy and efficiency degrade rapidly as the dimension increases.

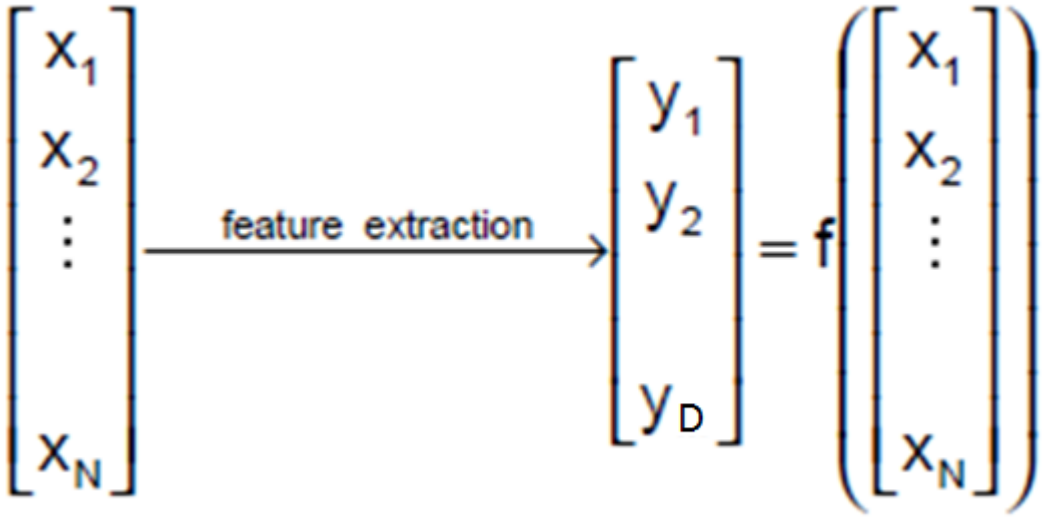
- The **intrinsic** feature dimension may be small.
  - For example, the number of genes responsible for a certain type of disease may be small.

Feature Dimensionality Reduction

- Given a set of **n** features, the goal of **feature selection** is to select a subset of **d** features (**d** < **n**) in order to minimize the classification error.



- Fundamentally different from dimensionality reduction (e.g., PCA or LDA) based on feature combinations (i.e., **feature extraction**).



Dimensionality Reduction and Data Visualization

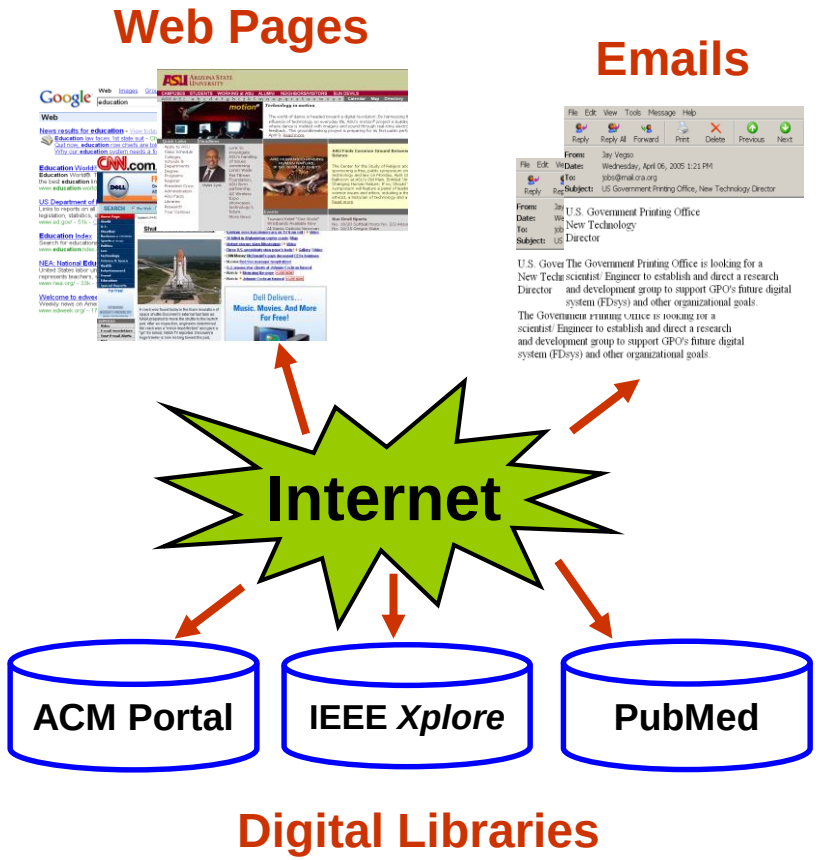
- Data visualization is very important for human to understand the structural relation among variables in a system.
- Visualization: projection of high-dimensional data onto 2D or 3D.

Application of Dimensionality Reduction

Paper "Visualizing High Dimensional and Big Data"

- **Data Visualization**
- Document Classification
- Microarray data analysis
- Protein classification
- Text mining
- Image retrieval
- Face recognition
- Handwritten digit recognition

Document Classification



Terms:Features

	T_1	T_2	.....	T_N	C
D_1	12	0	.....	6	Sports
D_2	3	10	.....	28	Travel
⋮	⋮			⋮	⋮
D_M	0	11	.....	16	Jobs

Documents

- **Task:** To classify unlabeled documents into categories
- **Challenge:** thousands of terms
- **Solution:** to apply dimensionality reduction

The need of feature selection
An illustrative example: online shopping prediction

Features (predictive variables, attributes)

Class

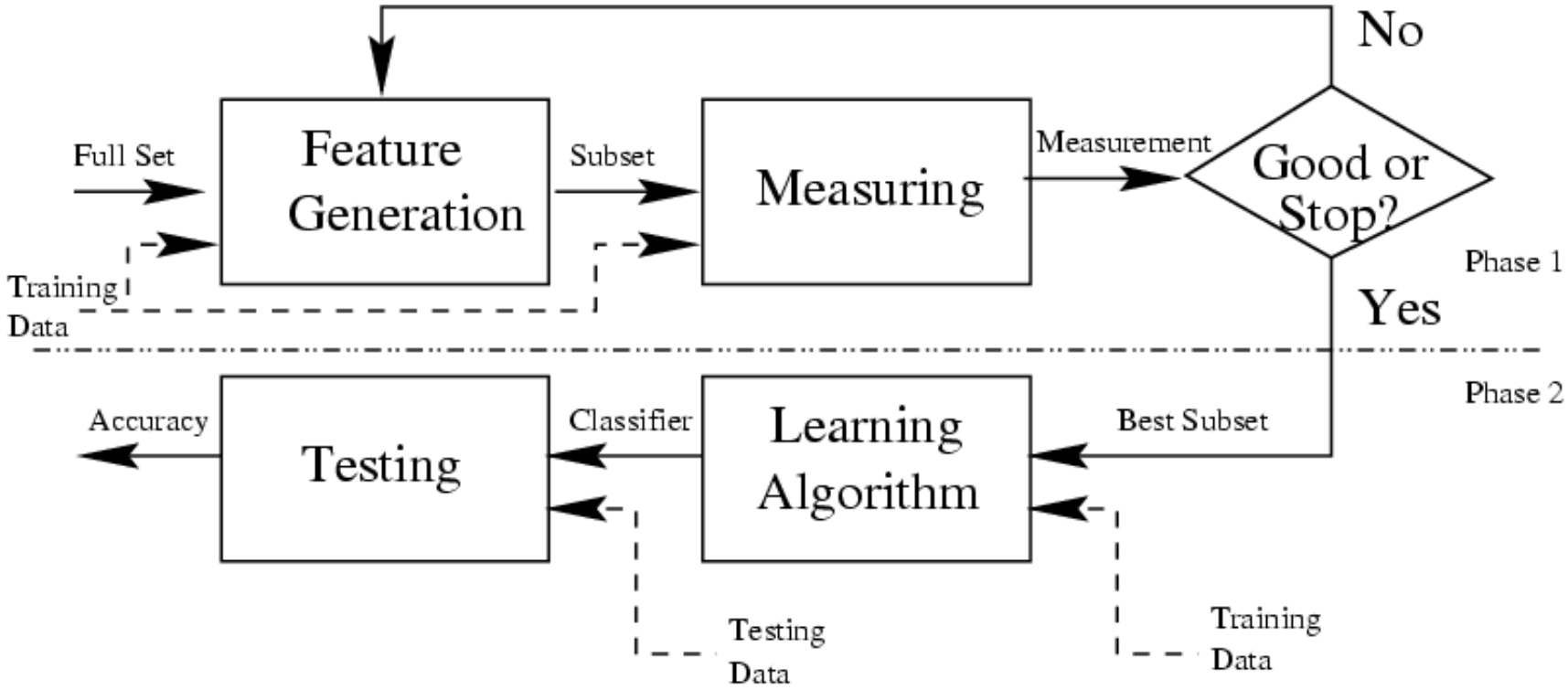
Customer	Page 1	Page 2	Page 3	….	Page 10,000	Buy a Book
1	1	3	1	….	1	Yes
2	2	1	0	….	2	Yes
3	2	0	0	….	0	No
…	…	…	…	…	…	…

- Difficult to understand
- Maybe only a small number of pages are needed, e.g. pages related to books and placing orders

Feature Selection Algorithms

- **Filter algorithm**
  - Separating feature selection from classifier learning
  - Relying on general characteristics of data, i.e., *information measurement*
  - No bias toward any learning algorithm, fast
- **Wrapper algorithm**
  - Relying on a predetermined classification algorithm
  - Using predictive accuracy as goodness measure
  - High accuracy, computationally expensive
- Wrapper methods are usually slower than filter methods but offer better performance.

Filter Algorithm



Feature Selection Methods

- Univariate Filter Methods
  - Consider one feature's contribution to the class at a time, e.g., entropy.
  - Advantages
    - Computationally efficient and parallelable
  - Disadvantages
    - May select low quality feature subsets

Feature Selection Methods

- Multivariate Filter methods
  - Consider the contribution of a set of features to the class variable.
  - Advantages:
    - Computationally efficient
    - Select higher-quality feature subsets than univariate filters
  - Disadvantages:
    - Not optimized for a given classifier

Prune of input variables

Features with the same value for all samples (variance=0) were eliminated.

Back seat	Compact Handlebar	Number of wheels
1	0	2
1	0	2
0	1	2
0	1	2

Prune of input variables

The variance (σ^2) is a measure of how far each value in the data set is from the mean.

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

Example: 3, 4, 4, 5, 6, 8

N	ΣX	ΣX^2	μ	μ^2	σ^2
6	30	166	5	25	2.67

Measures of Information

- Shannon proposed variant (Shannon's Entropy)

$$H = \sum_i p_i \cdot \log \frac{1}{p_i} = -\sum_i p_i \cdot \log p_i$$

  - weighs the information based on the probability that an outcome will occur
  - second term shows the amount of information an event provides is inversely proportional to its probability of occurring

Interpretations of Entropy

- The amount of information an event provides
  - An infrequently occurring event provides more information than a frequently occurring event
- The uncertainty in the outcome of an event
  - Systems with one very common event have less entropy than systems with many equally probable events

Example Data Set

	Hair	Height	Weight	Lotion	Result
i_1	1	2	1	0	1
i_2	1	3	2	1	0
i_3	2	1	2	1	0
i_4	1	1	2	0	1
i_5	3	2	3	0	1
i_6	2	3	3	0	0
i_7	2	2	3	0	0
i_8	1	1	1	1	0

Sunburn data

	Result (Sunburn)	
	No	Yes
P(Result)	5/8	3/8
P(Hair=1|Result)	2/5	2/3
P(Hair=2|Result)	3/5	0
P(Hair=3|Result)	0	1/3
P(Height=1|Result)	2/5	1/3
P(Height=2|Result)	1/5	2/3
P(Height=3|Result)	2/5	0
P(Weight=1|Result)	1/5	1/3
P(Weight=2|Result)	2/5	1/3
P(Weight=3|Result)	2/5	1/3
P(Lotion=0|Result)	2/5	3/3
P(Lotion=1|Result)	3/5	0

Priors and class conditional probabilities

Feature Ranking

- Weighting and ranking individual features
- Selecting top-ranked ones for feature selection
- Advantages
  - Efficient: $O(n)$ in terms of dimensionality n
  - Easy to implement
- Disadvantages
  - Hard to determine the threshold
  - Unable to consider correlation between features

Joint Entropy for Feature Selection

- Using joint entropy for feature selection:
  - Again define joint entropy to be:

$$H(A, B) = -\sum_{i,j} p(i, j) \cdot \log[\, p(i, j)]$$

  - Select sets of features that have **maximum joint entropy** since these will be the least aligned
  - These features will provide the most additional information

Search Strategies

- Assuming n features, an exhaustive search would require:
  - Examining all $\binom{n}{d}$ possible subsets of size d.
  - Selecting the subset that performs the best according to the criterion function.
- The number of subsets grows combinatorially, making exhaustive search impractical.
- In practice, heuristics are used to speed-up search but they **cannot** guarantee optimality.

Example Data Set

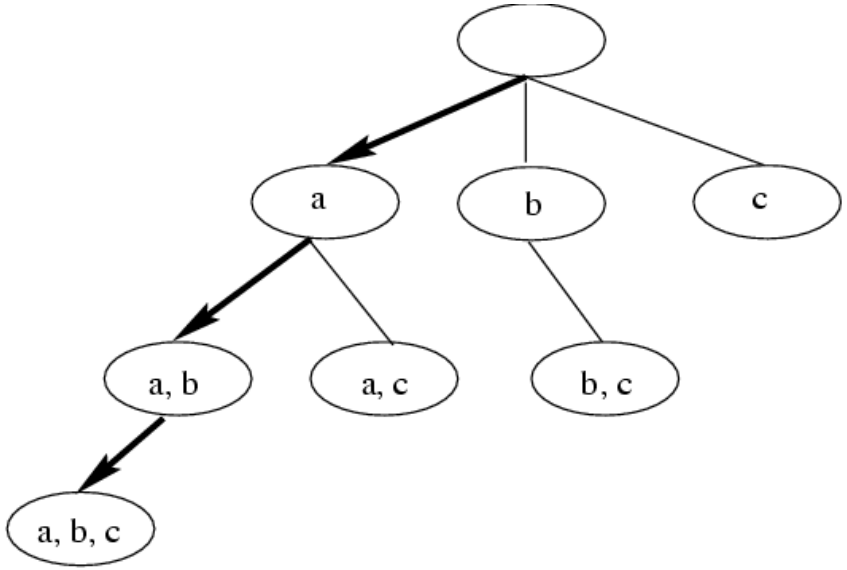
	Hair	Height	Weight	Lotion	Result
i_1	1	2	1	0	1
i_2	1	3	2	1	0
i_3	2	1	2	1	0
i_4	1	1	2	0	1
i_5	3	2	3	0	1
i_6	2	3	3	0	0
i_7	2	2	3	0	0
i_8	1	1	1	1	0

Sunburn data

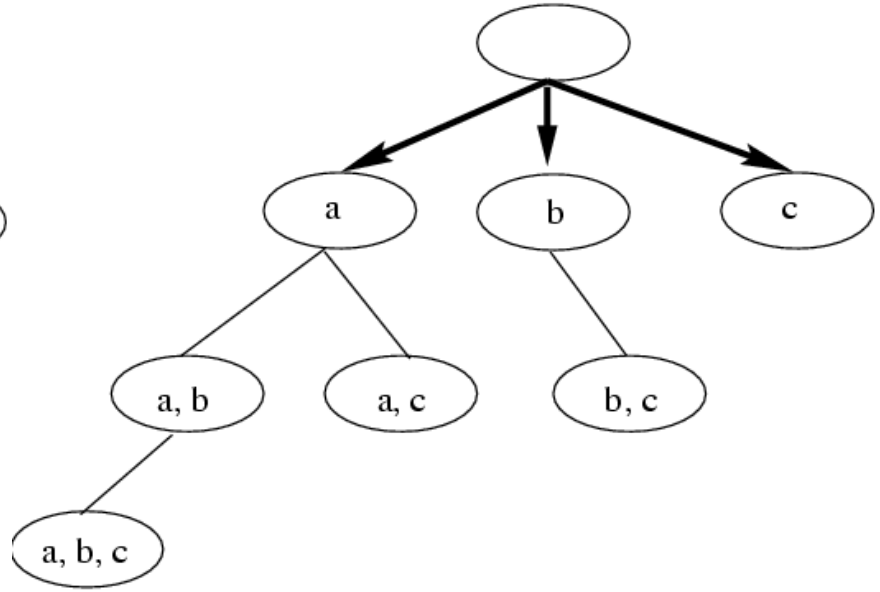
	Result (Sunburn)	
	No	Yes
P(Result)	5/8	3/8
P(Hair=1|Result)	2/5	2/3
P(Hair=2|Result)	3/5	0
P(Hair=3|Result)	0	1/3
P(Height=1|Result)	2/5	1/3
P(Height=2|Result)	1/5	2/3
P(Height=3|Result)	2/5	0
P(Weight=1|Result)	1/5	1/3
P(Weight=2|Result)	2/5	1/3
P(Weight=3|Result)	2/5	1/3
P(Lotion=0|Result)	2/5	3/3
P(Lotion=1|Result)	3/5	0

Priors and class conditional probabilities

Illustrations of Search Strategies



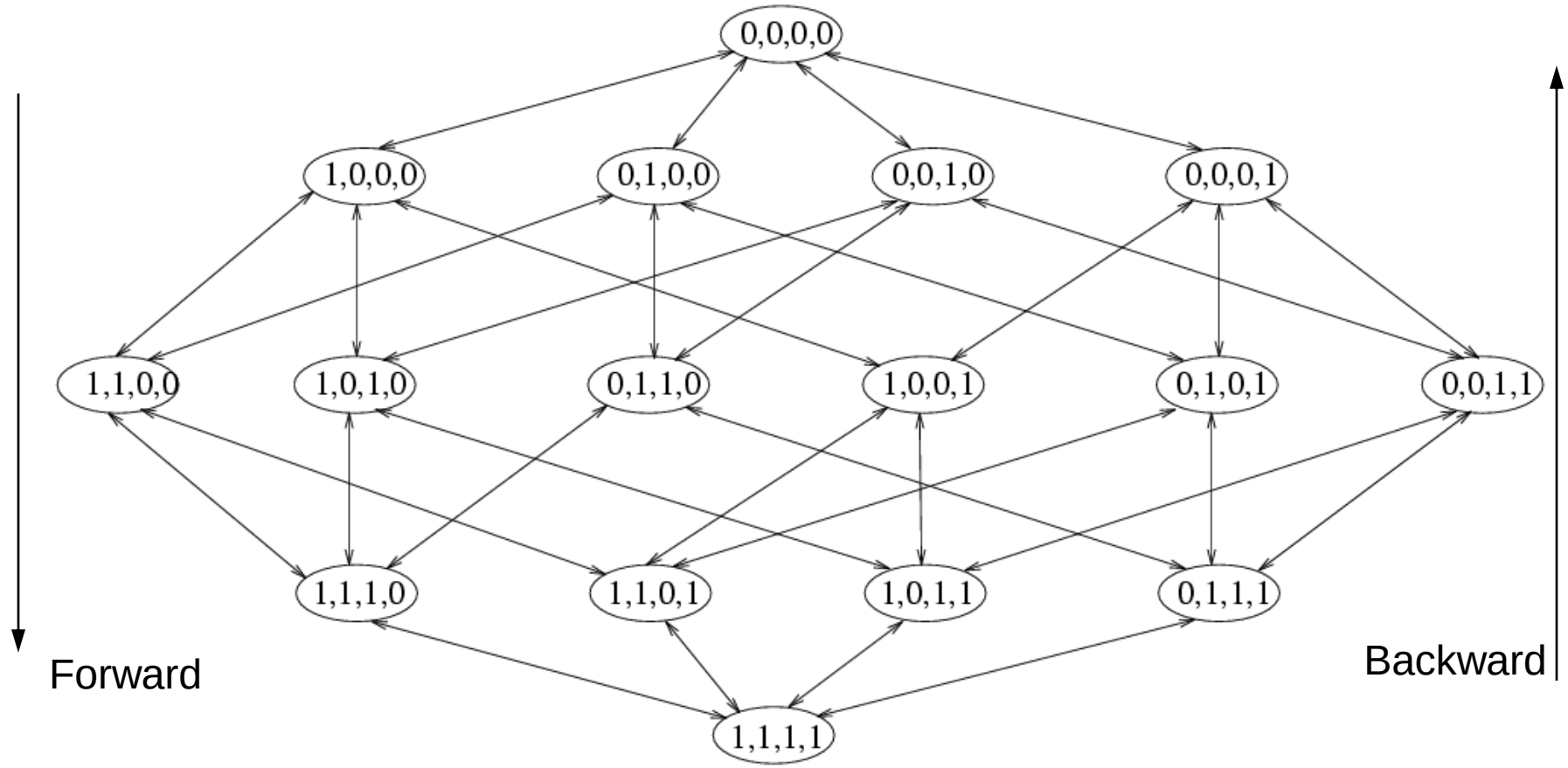
Depth-first search



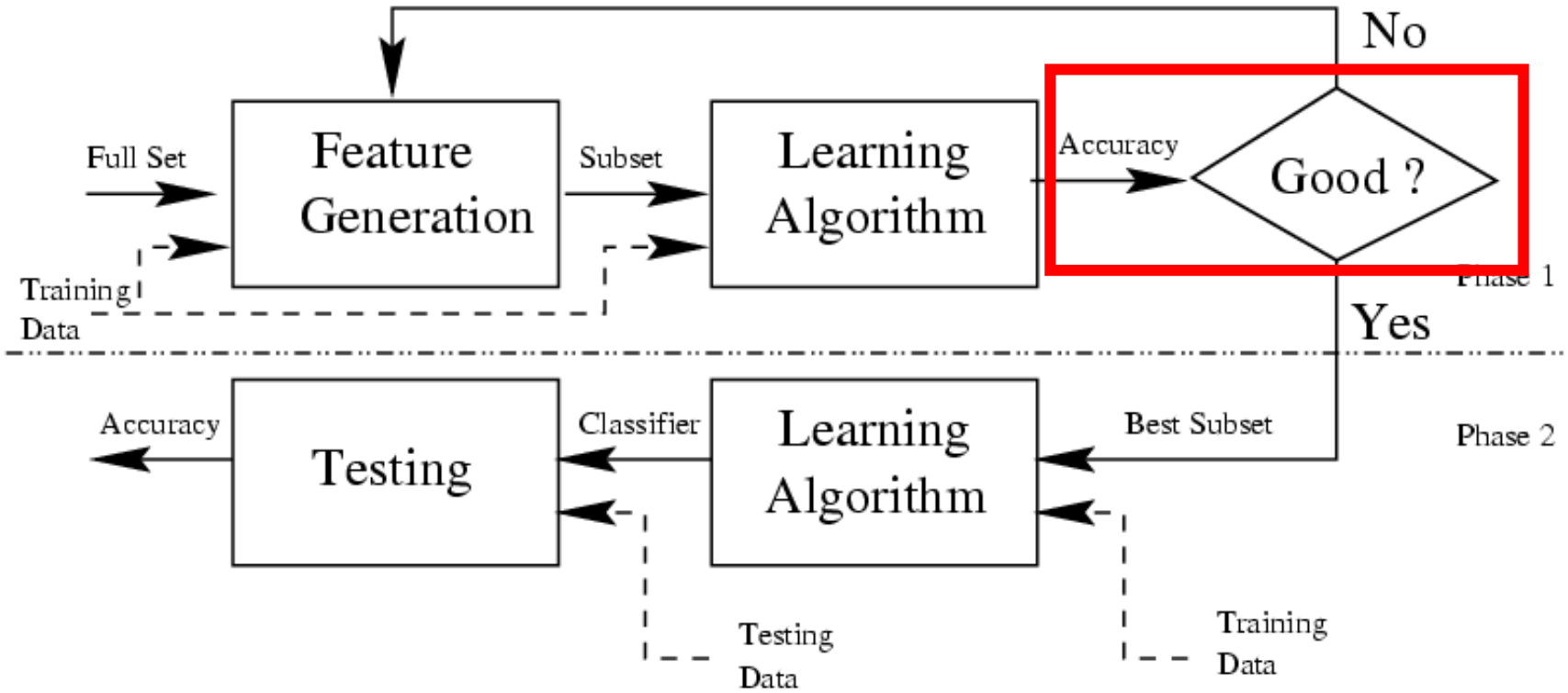
Breadth-first search

A Subset Search Problem

- An example of search space (*Kohavi & John 1997*)



Wrapper Algorithm

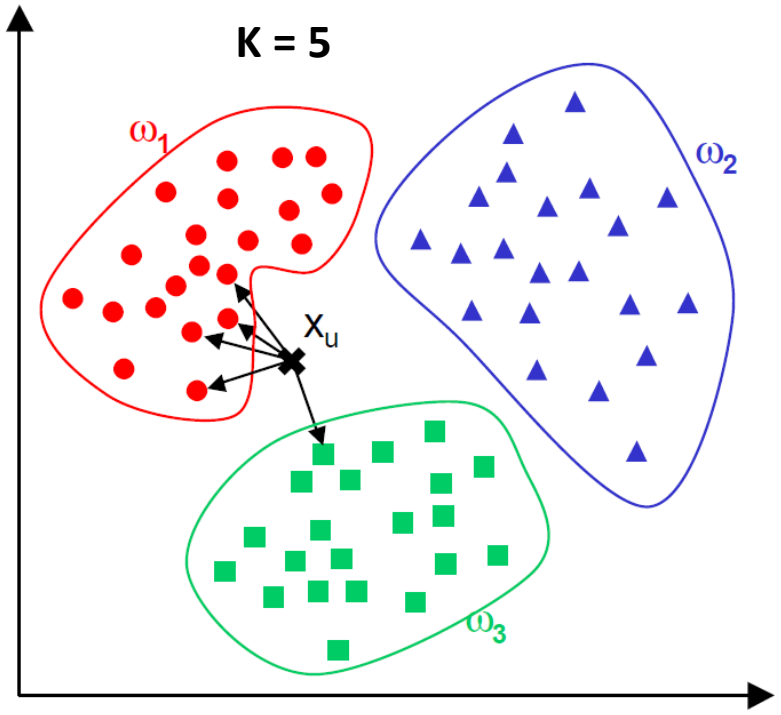


Wrapper Algorithm

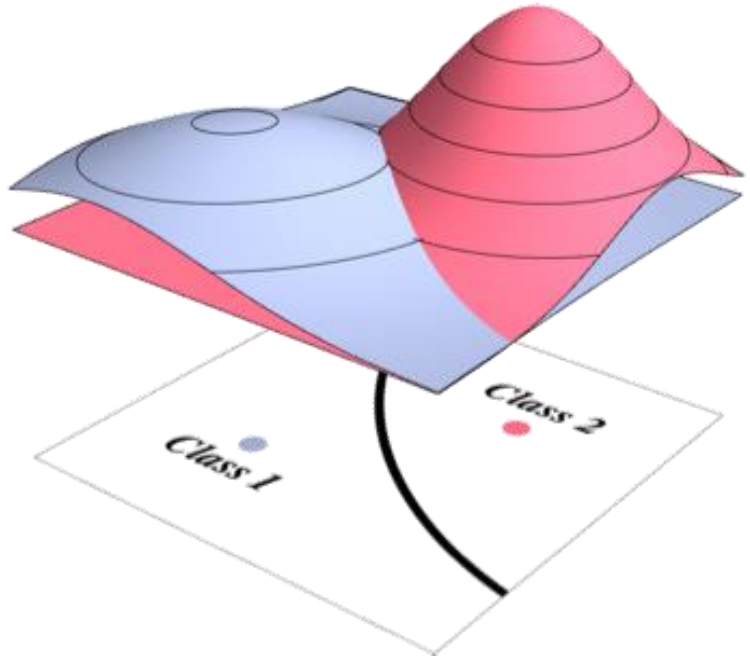
- Select a feature subset by building classifiers e.g.
  - Bayesian classifier
  - K Nearest Neighbors
  - Neural Network
  - SVM
- Advantages:
  - Select high-quality feature subsets for a particular classifier
- Disadvantages:
  - Classifiers are relatively computationally expensive.

Classifiers

- K nearest neighbors, Bayesian classifier



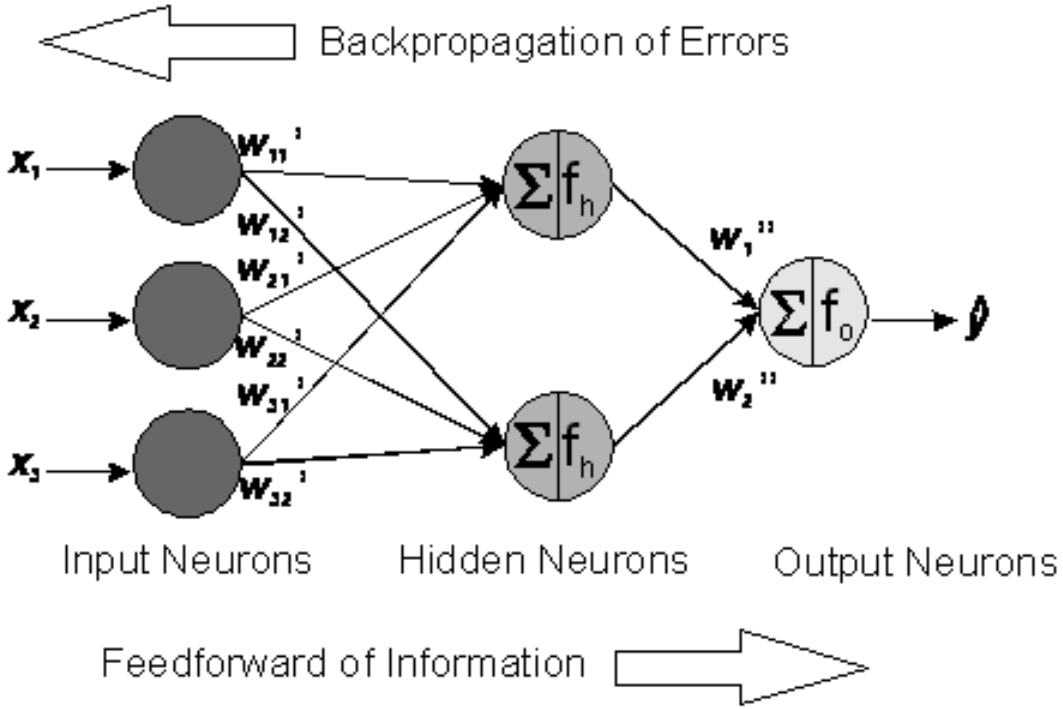
K nearest neighbors



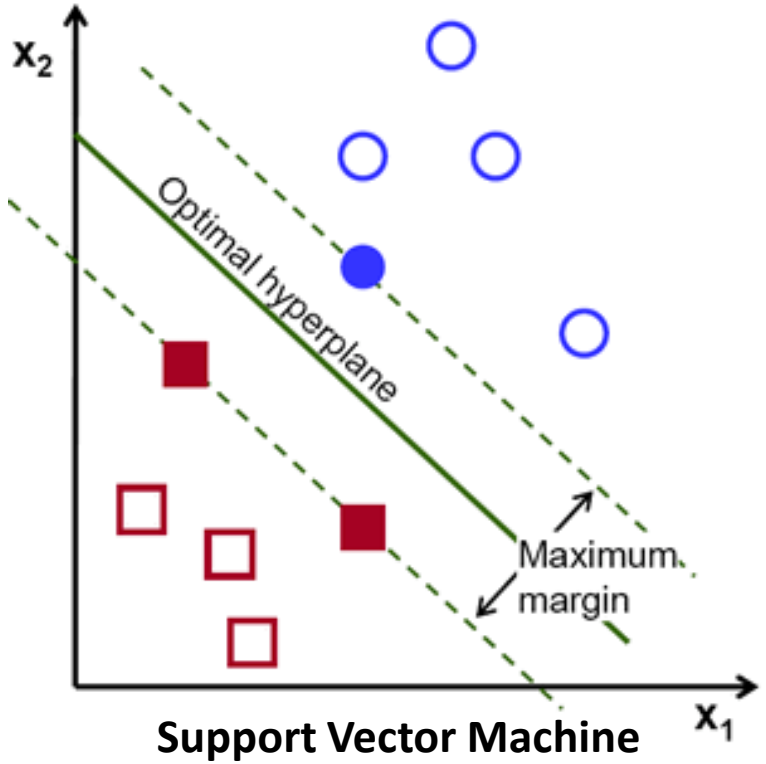
Bayesian classifier

Classifiers

- Neural Network, Support Vector Machine



Neural Network



Support Vector Machine

Feature Selection - Wrappers

- Optimizes for a specific learning algorithm
- The feature subset selection algorithm is a "wrapper" around the learning algorithm
  1. Pick a feature subset and pass it in to learning algorithm
  2. Create training/test set based on the feature subset
  3. Train the learning algorithm with the training set
  4. Find accuracy (objective) with validation set
  5. Repeat for all feature subsets and pick the feature subset which led to the highest predictive accuracy (or other objective)
- This approach is simple.

An Example for Optimal Subset

F_1	F_2	F_3	F_4	F_5	C
0	0	1	0	1	0
0	1	0	0	1	1
1	0	1	0	1	1
1	1	0	0	1	1
0	0	1	1	0	0
0	1	0	1	0	1
1	0	1	1	0	1
1	1	0	1	0	1

- Data set (whole set)
  - Five Boolean features
  - $F_3 = \neg F_2$, $F_5 = \neg F_4$
  - $C = F_1 \vee F_2$
  - Optimal subset:

    $\{F_1, F_2\}$ or $\{F_1, F_3\}$
- Combinatorial nature of searching for an optimal subset

CLASS LABEL C IS NOT CONSIDERED AS FEATURE!

Hybrid Algorithms

- Combine the best properties of filters and wrappers.
- Usual approach:
  - First, a filter method is used in order to reduce the feature space dimension space, possibly obtaining several candidate subsets.
  - Then, a wrapper is employed to find the best candidate subset.
- Highly used in recent years
  - E.g. fuzzy random forest feature selection, hybrid genetic algorithms.

Summary

- Dimensionality reduction
  - Feature Selection
  - Feature Extraction
- Feature Selection
  - Filter method
  - Wrapper method
  - Hybrid method