Feature Selection

CE-725: Statistical Pattern Recognition

Sharif University of Technology

Soleymani

Fall 2016

Outline

- Dimensionality reduction
- Filter univariate methods
- Multi-variate filter & wrapper methods
- Evaluation criteria
- Search strategies

Avoiding overfitting

- Structural risk minimization
- Regularization
- Cross-validation
  - Model-selection
- Feature selection

Dimensionality reduction: Feature selection vs. feature extraction

- Feature **selection**
  - Select a subset of a given feature set
- Feature **extraction** (e.g., PCA, LDA)
  - A linear or non-linear transform on the original feature space

$$\begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \rightarrow \begin{bmatrix} x_{i_1} \\ \vdots \\ x_{i_{d'}} \end{bmatrix}$$

Feature Selection ($d' < d$)

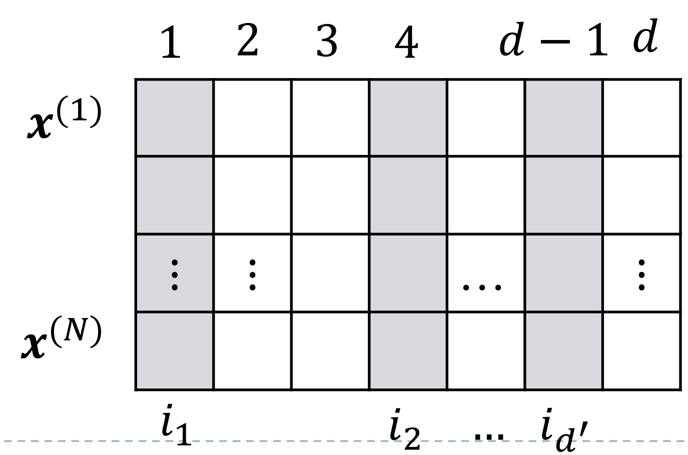
$$\begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \rightarrow \begin{bmatrix} y_1 \\ \vdots \\ y_{d'} \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}\right)$$

Feature Extraction

Feature selection

- Data may contain many irrelevant and redundant variables and often comparably few training examples

- Consider supervised learning problems where the number of features d is very large (perhaps $d \gg n$)
  - E.g., datasets with tens or hundreds of thousands of features and (much) smaller number of data samples (text or document processing, gene expression array analysis)



Why feature selection?

- FS is a way to find **more accurate, faster, and easier to understand** classifiers.
  - Performance: enhancing generalization ability
    - alleviating the effect of the curse of dimensionality
    - the higher the ratio of the no. of training patterns N to the number of free classifier parameters, the better the generalization of the learned classifier
  - Efficiency: speeding up the learning process
  - Interpretability: resulting a model that is easier to understand

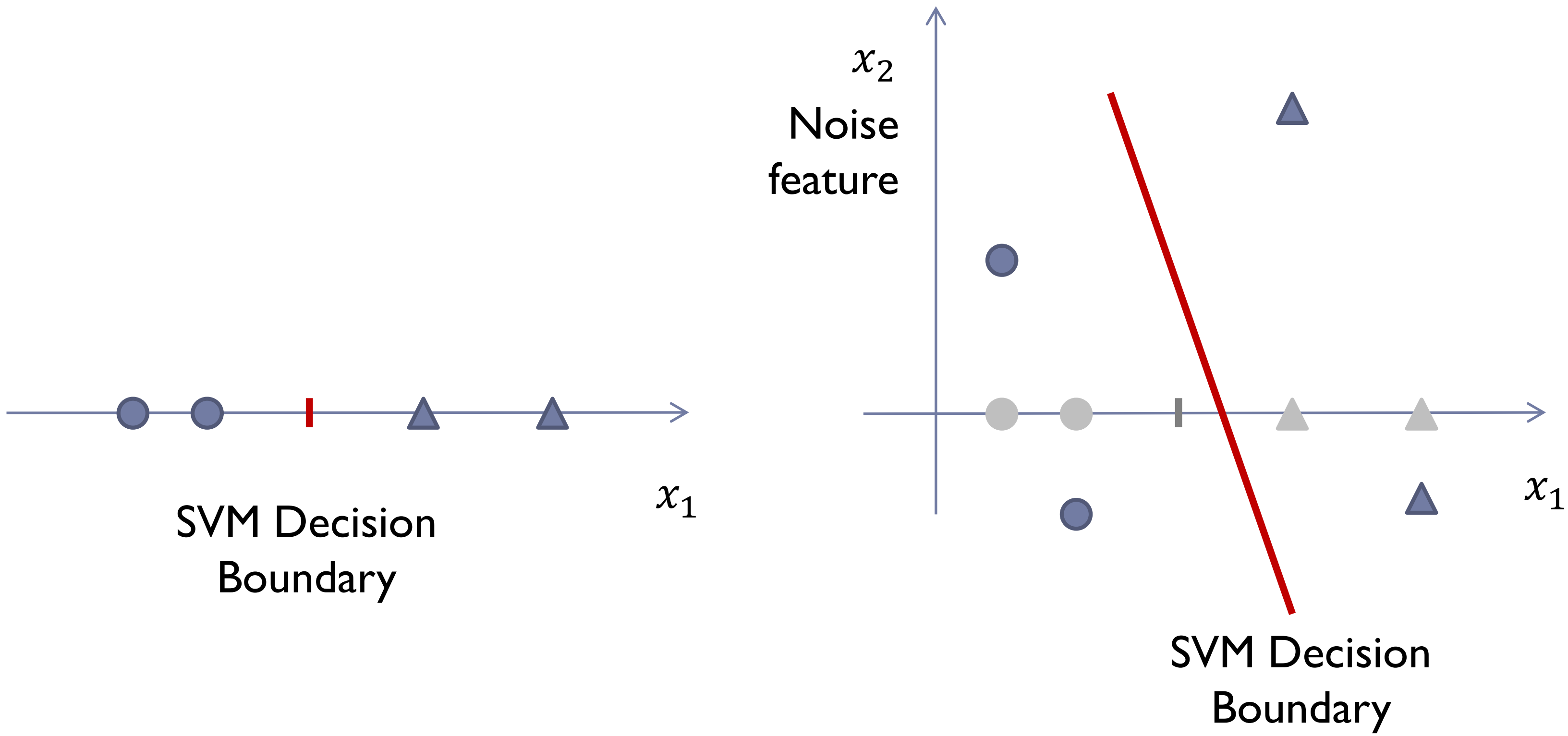
$$\boldsymbol{X} = \begin{bmatrix} x_1^{(1)} & \cdots & x_d^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(N)} & \cdots & x_d^{(N)} \end{bmatrix}, \; Y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix} \Longrightarrow \boxed{\text{Feature Selection}} \Longrightarrow i_1, i_2, \ldots, i_{d'}$$

The selected features

Supervised feature selection: Given a labeled set of data points, select a subset of features for data representation

Noise (or irrelevant) features

- Eliminating irrelevant features can decrease the classification error on test data



Some definitions

- One categorization of feature selection methods:
  - **Univariate method**: considers one variable (feature) at a time.
  - **Multivariate method**: considers subsets of features together.

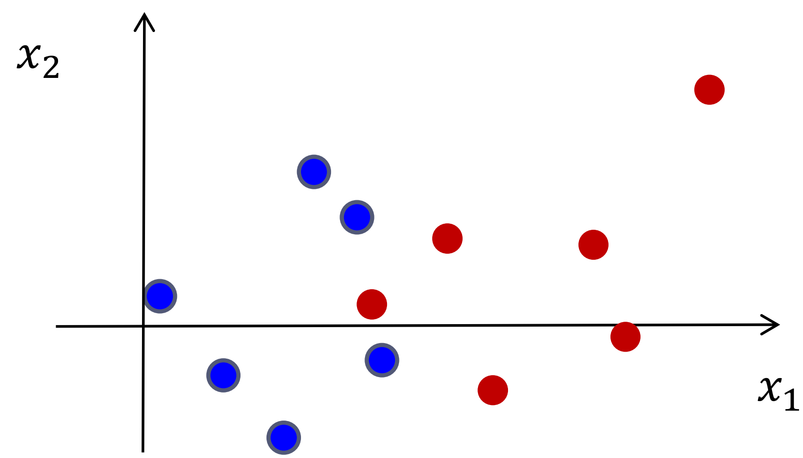
- Another categorization:
  - **Filter method**: ranks features or feature subsets independent of the classifier as a preprocessing step.
  - **Wrapper method**: uses a classifier to evaluate the score of features or feature subsets.
  - **Embedded method**: Feature selection is done during the training of a classifier
    - E.g., Adding a regularization term $\|\boldsymbol{w}\|_1$ in the cost function of linear classifiers

Filter: univariate

- Univariate filter method
  - Score each feature k based on the k-th column of the data matrix and the label vector
    - Relevance of the feature to predict labels: Can the feature discriminate the patterns of different classes?
  - Rank features according to their score values and select the ones with the highest scores.
    - How do you decide how many features k to choose? e.g., using cross validation to select among the possible values of k
- Advantage: computational and statistical scalability

Pearson Correlation Criteria

$$R(k) = \frac{cov(X_k, Y)}{\sqrt{var(X_k)}\sqrt{var(Y)}} \approx \frac{\sum_{i=1}^{N}\left(x_k^{(i)} - \overline{x_k}\right)\left(y^{(i)} - \bar{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_k^{(i)} - \overline{x_k}\right)^2 \sum_{i=1}^{N}(y^{(i)} - \bar{y})^2}}$$



$R(1) \gg R(2)$

Univariate Mutual Information

- Independence:

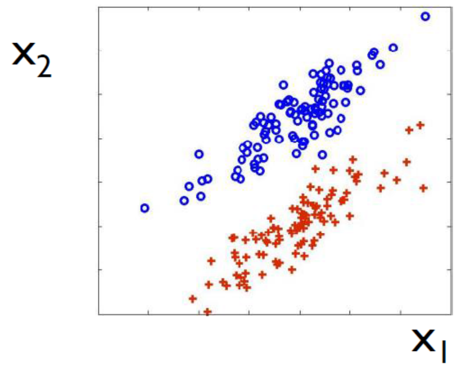
$$P(X,Y) = P(X)P(Y)$$

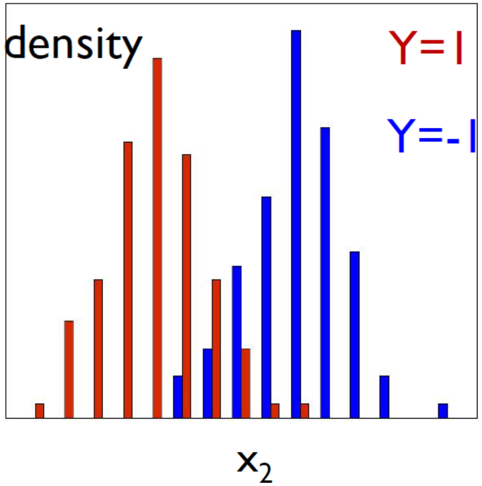
- Mutual information as a measure of dependence:

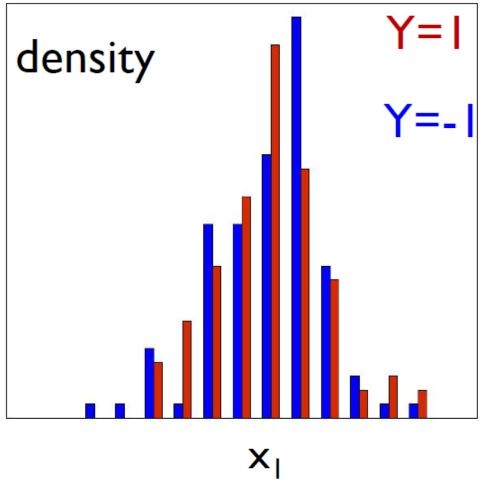
$$MI(X,Y) = E_{X,Y}\left[\log \frac{P(X,Y)}{P(X)P(Y)}\right]$$

- Score of X_k based on MI with Y:
  - $I(k) = MI(X_k, Y)$

Example







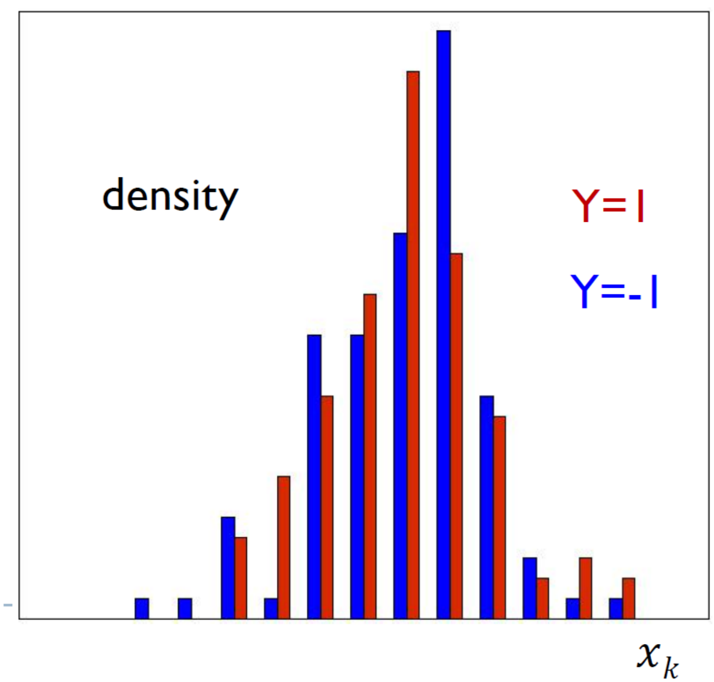
Irrelevance

- X_k: random variable corresponding to the k-th component of input feature vectors
- Y: random variable corresponding to the labels

- Irrelevance feature X_k to predict Y ($C = 2$):
  - $P(X_k|Y = 1) = P(X_k|Y = -1)$

Using KL divergence to find a distance between $P(X_k|Y = 1)$ and $P(X_k|Y = -1)$:

$$d(k) = D(P(X_k|Y = 1)||P(X_k|Y = -1)) + D(P(X_k|Y = -1)||P(X_k|Y = 1))$$



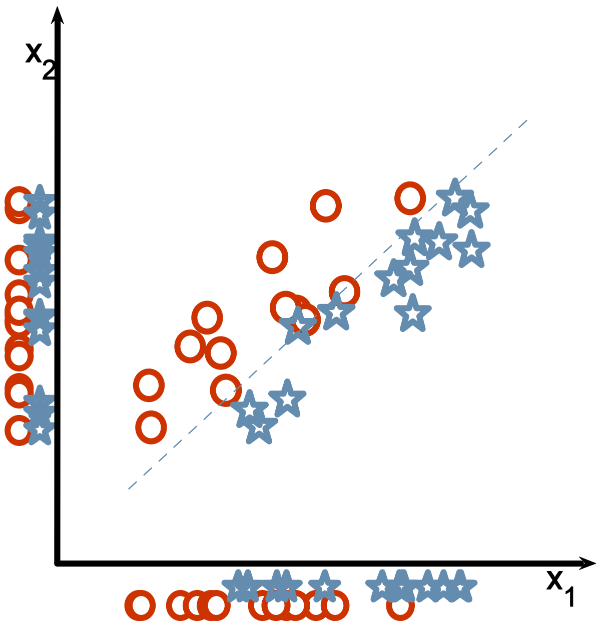
Filter – univariate: Disadvantage

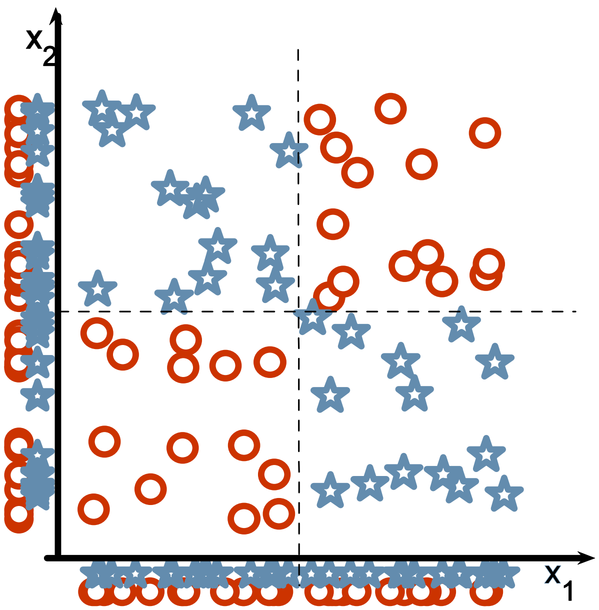
- Redundant subset: Same performance could possibly be achieved with a smaller subset of complementary variables that does not contain redundant features.

- What is the relation between redundancy and correlation:
  - Are highly correlated features necessarily redundant?
  - What about completely correlated ones?

Univariate methods: Failure

- Samples where univariate feature analysis and scoring fails:



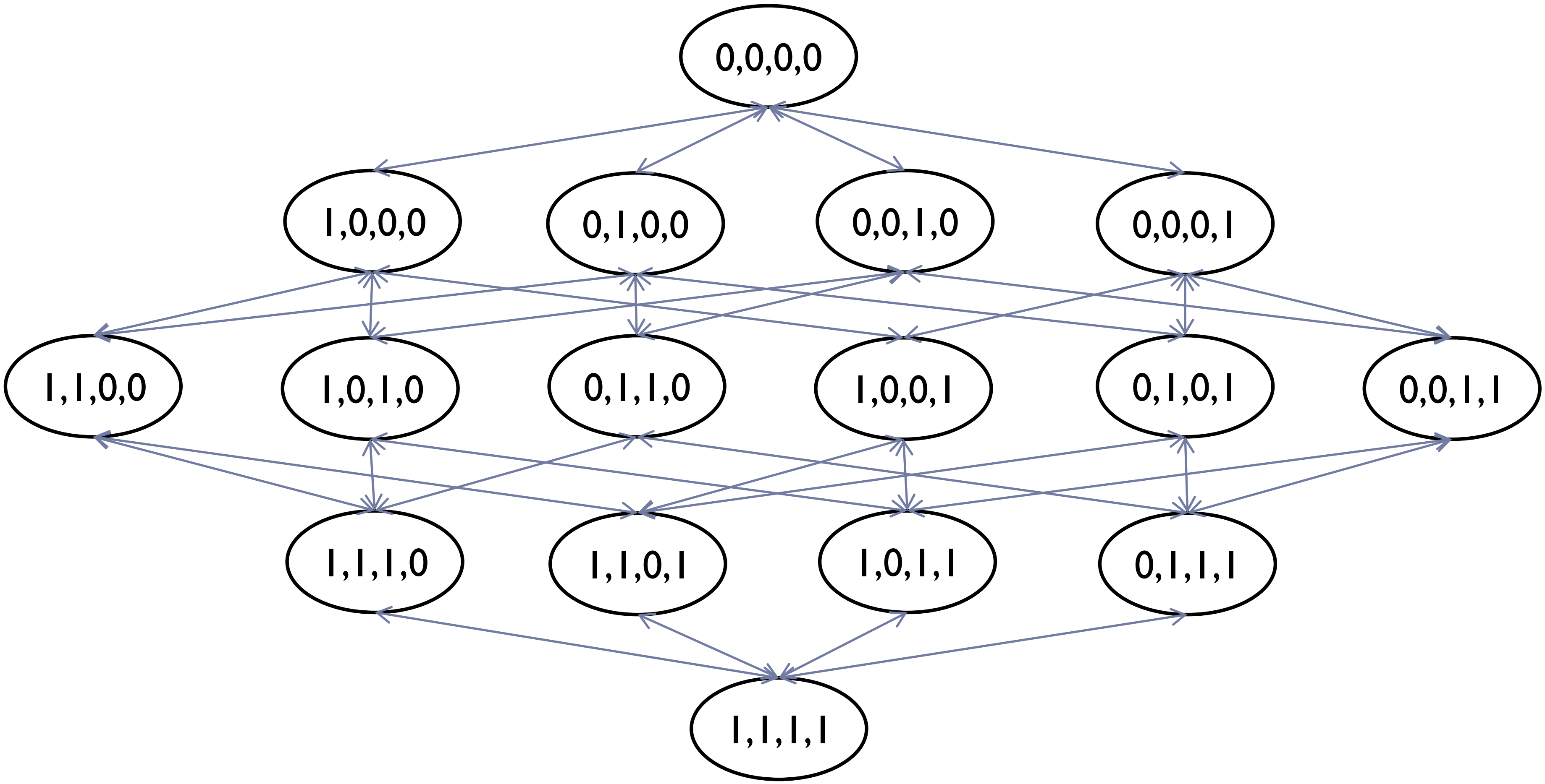


[Guyon-Elisseeff, JMLR 2004; Springer 2006]

Multi-variate feature selection

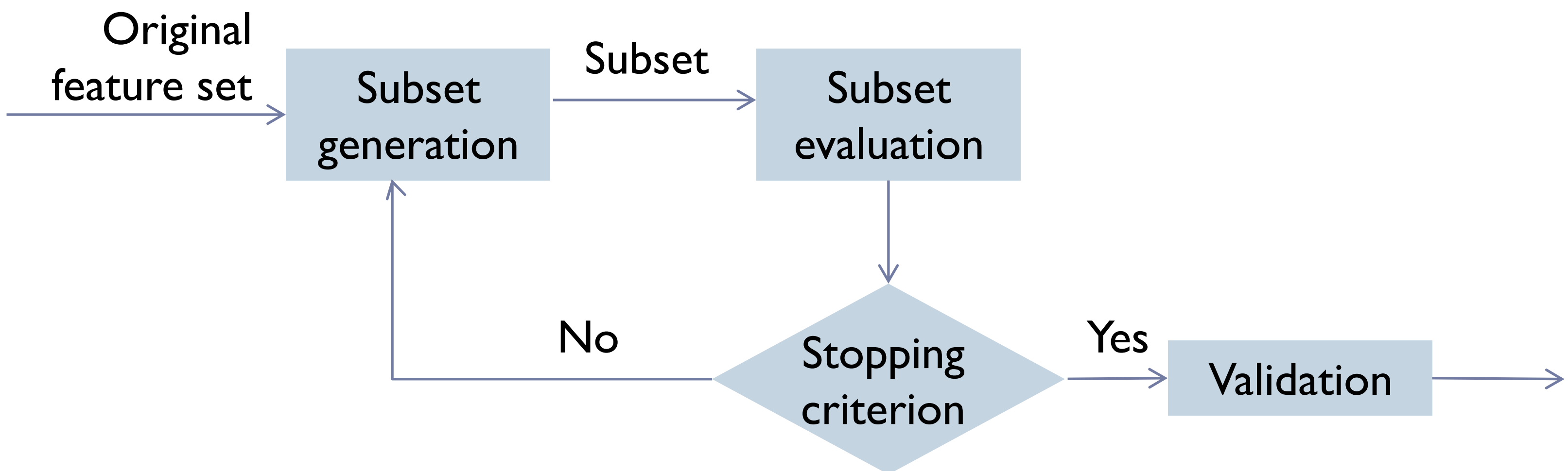
- Search in the space of all possible combinations of features.
  - all feature subsets: For d features, 2^d possible subsets.
  - high computational and statistical complexity.
- Wrappers use the classifier performance to evaluate the feature subset utilized in the classifier.
  - Training 2^d classifiers is infeasible for large d.
  - Most wrapper algorithms use a heuristic search.
- Filters use an evaluation function that is cheaper to compute than the performance of the classifier
  - e.g. correlation coefficient

Search space for feature selection ($d = 4$)



[Kohavi-John, 1997]

Multivariate methods: General procedure



Subset Generation: select a candidate feature subset for evaluation

Subset Evaluation: compute the score (relevancy value) of the subset

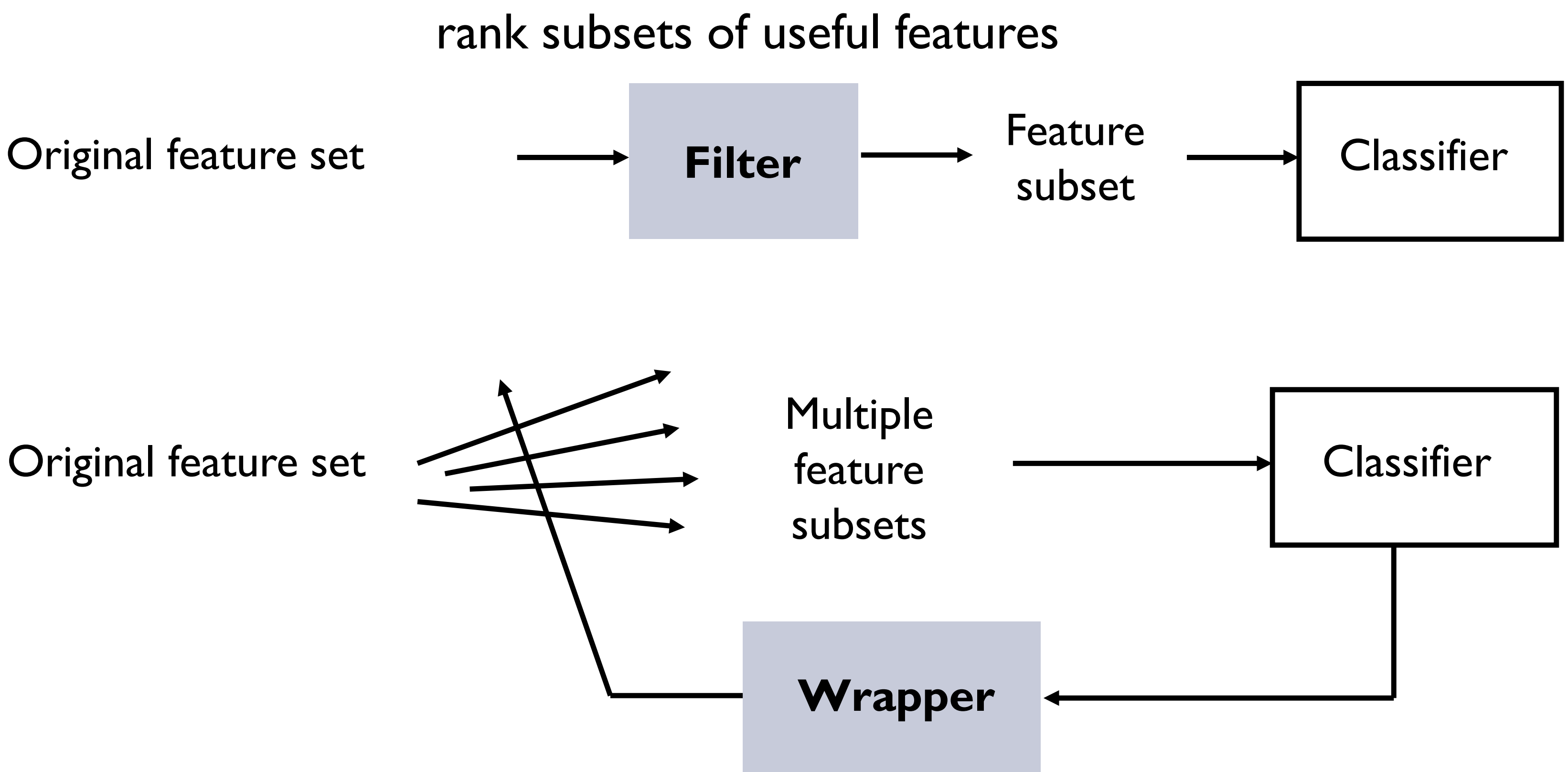
Stopping criterion: when stopping the search in the space of feature subsets

Validation: verify that the selected subset is valid

Stopping criteria

- Predefined number of features is selected
- Predefined number of iterations is reached
- Addition (or deletion) of any feature does not result in a better subset
- An optimal subset (according to the evaluation criterion) is obtained.

Filters vs. wrappers



take classifier into account to rank feature subsets
(e.g., using cross validation to evaluate features)

Wrapper methods: Performance assessment

- For each feature subset, train classifier on training data and assess its performance using evaluation techniques like cross-validation

Filter methods: Evaluation criteria

- Distance (Eulidean distance)
  - Class separability: Features supporting instances of the same class to be closer in terms of distance than those from different classes
- Information (Information Gain)
  - Select S1 if IG(S1,Y)>IG(S2,Y)
- Dependency (correlation coefficient)
  - good feature subsets contain features highly correlated with the class, yet uncorrelated with each other
- Consistency (min-features bias)
  - Selects features that guarantee no inconsistency in data
    - inconsistent instances have the same feature vector but different class labels
  - Prefers smaller subset with consistency (min-feature)

	f_1	f_2	class
instance 1	a	b	c1
instance 2	a	b	c2

inconsistent

Subset selection or generation

- Search direction
  - Forward
  - Backward
  - Random

- Search strategies
  - **Exhaustive - Complete**
    - Branch & Bound
    - Best first
  - **Heuristic**
    - Sequential forward selection
    - Sequential backward elimination
    - Plus-l Minus-r Selection
    - Bidirectional Search
    - Sequential floating Selection
  - **Non-deterministic**
    - Simulated annealing
    - Genetic algorithm

Search strategies

- Complete: Examine all combinations of feature subset
  - Optimal subset is achievable
  - Too expensive if d is large
- Heuristic: Selection is directed under certain guidelines
  - Incremental generation of subsets
  - Smaller search space and thus faster search
  - May miss out feature sets of high importance
- Non-deterministic or random: No predefined way to select feature candidate (i.e., probabilistic approach)
  - Optimal subset depends on the number of trials
  - Need more user-defined parameters

Feature Selection: Summary

- Most univariate methods are filters and most wrappers are multivariate.

- No feature selection method is universally better than others:
  - wide variety of variable types, data distributions, and classifiers.

- Match the method complexity to the ratio d/N:
  - univariate feature selection may work better than multivariate.

References

- I. Guyon and A. Elisseeff, An Introduction to Variable and Feature Selection, JMLR, vol. 3, pp. 1157-1182, 2003.
- S. Theodoridis and K. Koutroumbas, Pattern Recognition, 4th edition, 2008. [Chapter 5]
- H. Liu and L. Yu, Feature Selection for Data Mining, 2002.

Filters vs. Wrappers

- Filters
  - ☑ **Fast execution**: evaluation function computation is faster than a classifier training
  - ☑ **Generality**: Evaluate intrinsic properties of the data, rather than their interactions with a particular classifier ("good" for a larger family of classifiers)
  - ☒ **Tendency to select large subsets**: Their objective functions are generally monotonic (so tending to select the full feature set).
    - a cutoff is required on the number of features

- Wrappers
  - ☒ **Slow execution**: must train a classifier for each feature subset (or several trainings if cross-validation is used)
  - ☒ **Lack of generality**: the solution lacks generality since it is tied to the bias of the classifier used in the evaluation function.
  - ☑ **Ability to generalize**: Since they typically use cross-validation measures to evaluate classification accuracy, they have a mechanism to avoid overfitting.
  - ☑ **Accuracy**: Generally achieve better recognition rates than filters since they find a proper feature set for the intended classifier.