# UVA CS 4501: Machine Learning

## Lecture 7: Feature Selection

Dr. Yanjun Qi

University of Virginia

Department of
Computer Science

# Where are we ? ➔ Five major sections of this course

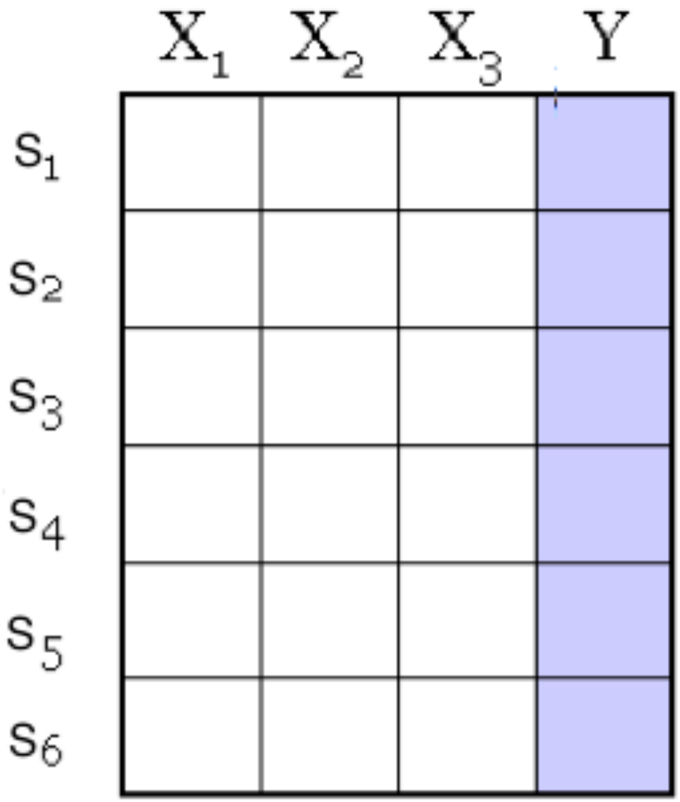
- ❑ Regression (supervised)
- ❑ Classification (supervised)
- ❑ Unsupervised models
- ❑ Learning theory
- ❑ Graphical models

# Today →

## Regression (supervised)

- ❑ Four ways to train / perform optimization for linear regression models
  - ❑ Normal Equation
  - ❑ Gradient Descent (GD)
  - ❑ Stochastic GD
  - ❑ Newton's method
- ❑ Supervised regression models
  - ❑ Linear regression (LR)
  - ❑ LR with non-linear basis functions
  - ❑ Locally weighted LR
  - ❑ LR with Regularizations
- ❑ Feature selection

# A labeled Dataset




$$f : \boxed{X} \longrightarrow \boxed{Y}$$

- **Data**/*points/instances/examples/samples/records*: [ rows ]
- **Features**/*attributes/dimensions/independent variables/covariates/predictors/regressors*: [ columns, except the last]
- **Target**/*outcome/response/label/dependent variable*: special column to be predicted [ last column ]

# Today

- Feature Selection (supervised)
  - Filtering approach
  - Wrapper approach
  - Embedded methods

# Feature Selection → Simpler models

- Because:
  - Simpler to use (lower computational complexity)
  - Easier to train (needs less examples)
  - Less sensitive to noise
  - Easier to explain (more interpretable)
  - Generalizes better (lower variance - Occam's razor)
    --- More in future lectures!!!

# Occam's razor: law of parsimony

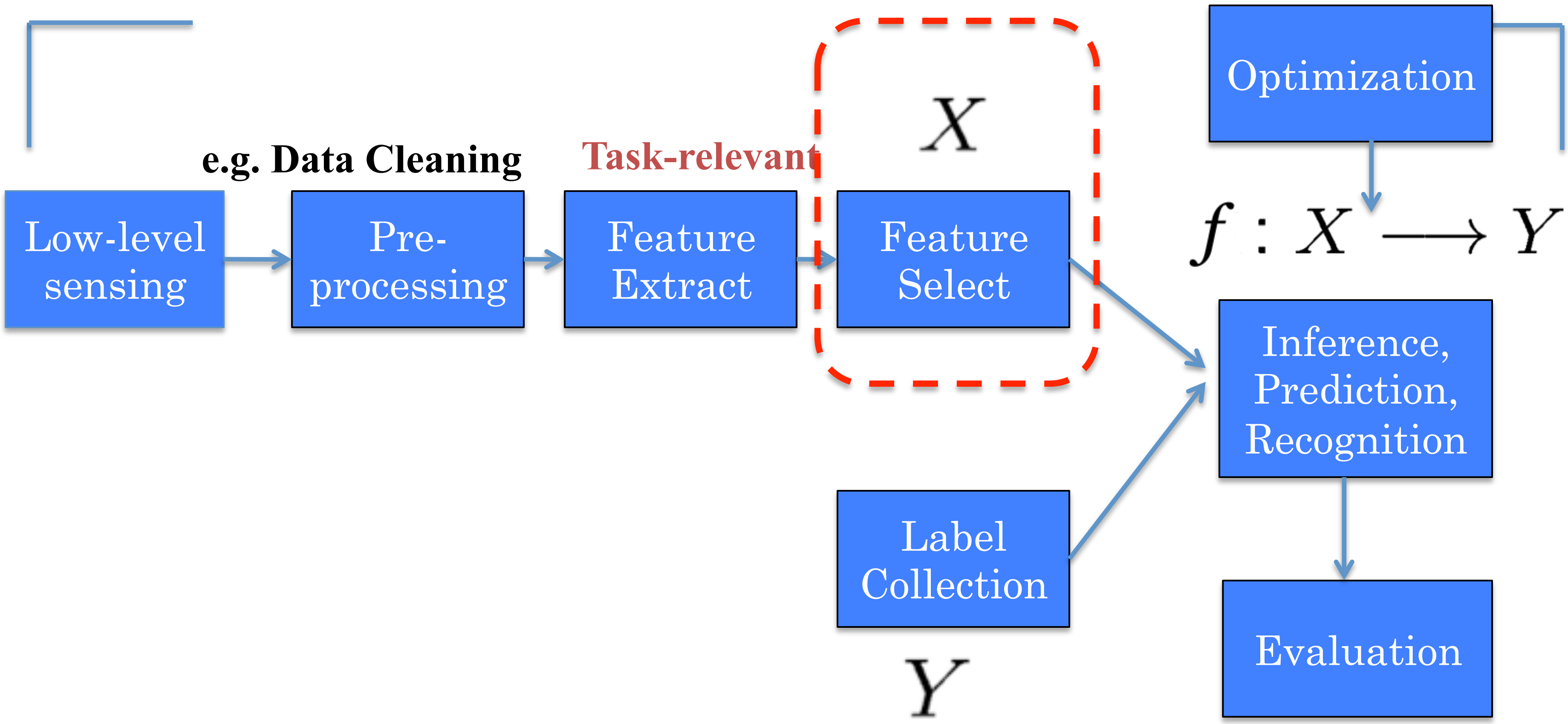
image at: [ww.butterflyeffect.ca/.../ OccamsRazor.htmlRemove frame](ww.butterflyeffect.ca/.../OccamsRazor.htmlRemove frame)


**The principle of Occam's razor**

states that the explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference to any observable predictions of the theory

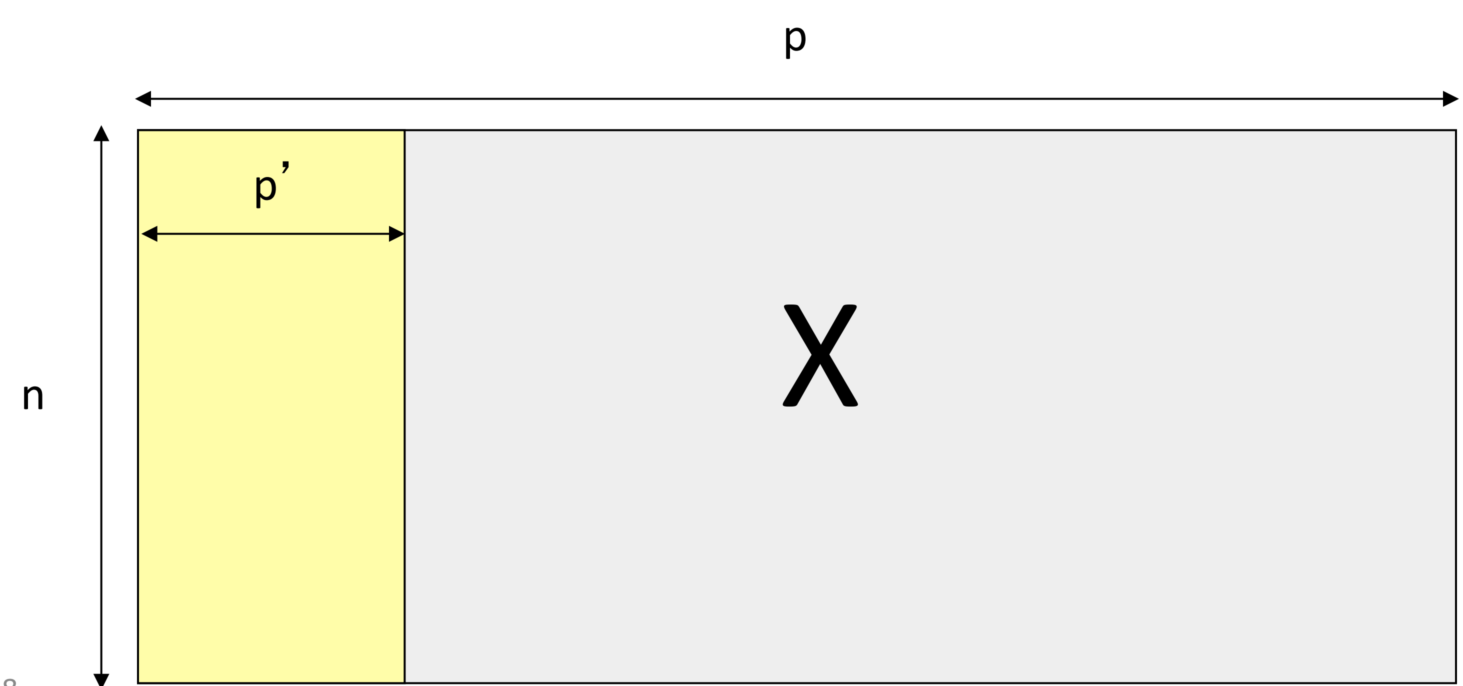
parsimony: extreme unwillingness to spend money or use resources.

# A Typical Machine Learning Pipeline



# Feature Selection

- **Thousands to millions of low level features**: select the most relevant ones to build **better, faster, and easier to understand** learning models.




From Dr. **Isabelle Guyon**

# e.g., Movie Reviews and Revenues: An Experiment in Text Regression, Proceedings of HLT '10 (1.7k n / >3k features)

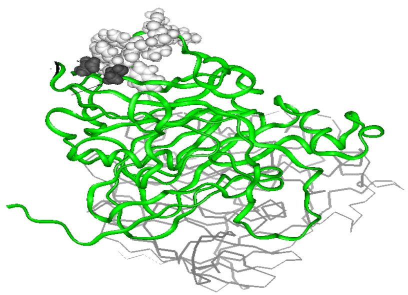
## IV. Features

e.g. counts of a ngram in the text

| | |
|---|---|
| I | Lexical n-grams (1,2,3) |
| II | Part-of-speech n-grams (1,2,3) |
| III | Dependency relations (nsubj,advmod,... ) |
| **Meta** | U.S. origin, running time, budget (log), # of opening screens, genre, MPAA rating, holiday release (summer, Christmas, Memorial day,... ), star power (Oscar winners, high-grossing actors) |

$n \approx 1700$ / $p > 30{,}000$

# e.g., QSAR: Drug Screening



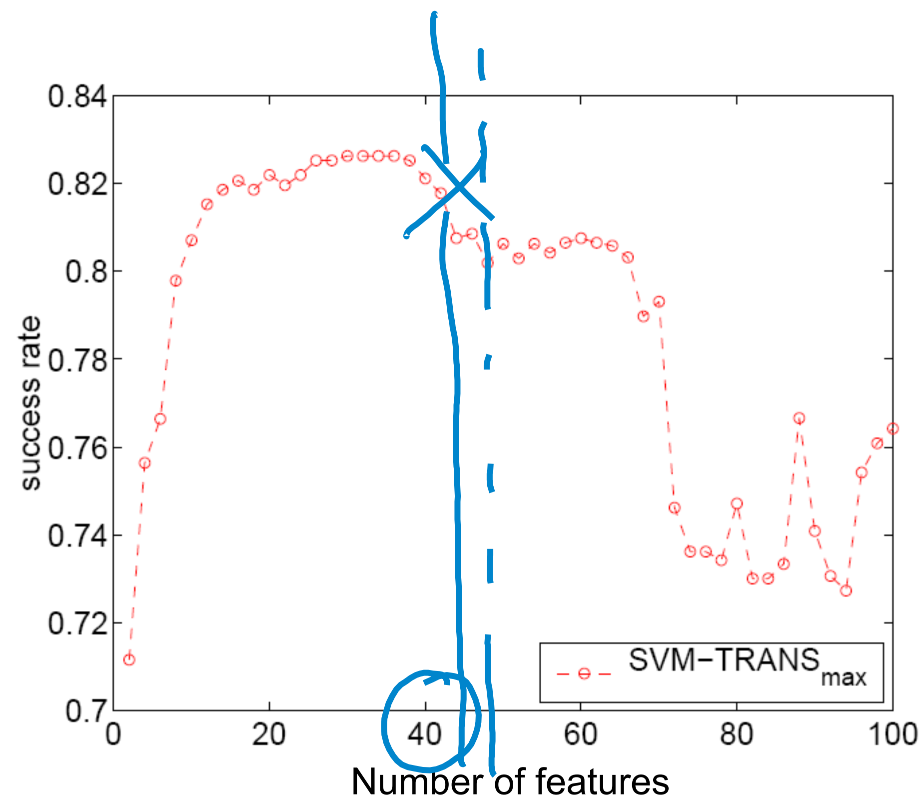
n

**Binding to Thrombin (DuPont Pharmaceuticals)**

- 2543 compounds tested for their ability to bind to a target site on thrombin, a key receptor in blood clotting; 192 “active” (bind well); the rest “inactive”. Training set (1909 compounds) more depleted in active compounds.

p

- **139,351 binary features**, which describe three-dimensional properties of the molecule.




***Weston et al, Bioinformatics, 2002***

# e.g., Leukemia Diagnosis

p’ >30,000

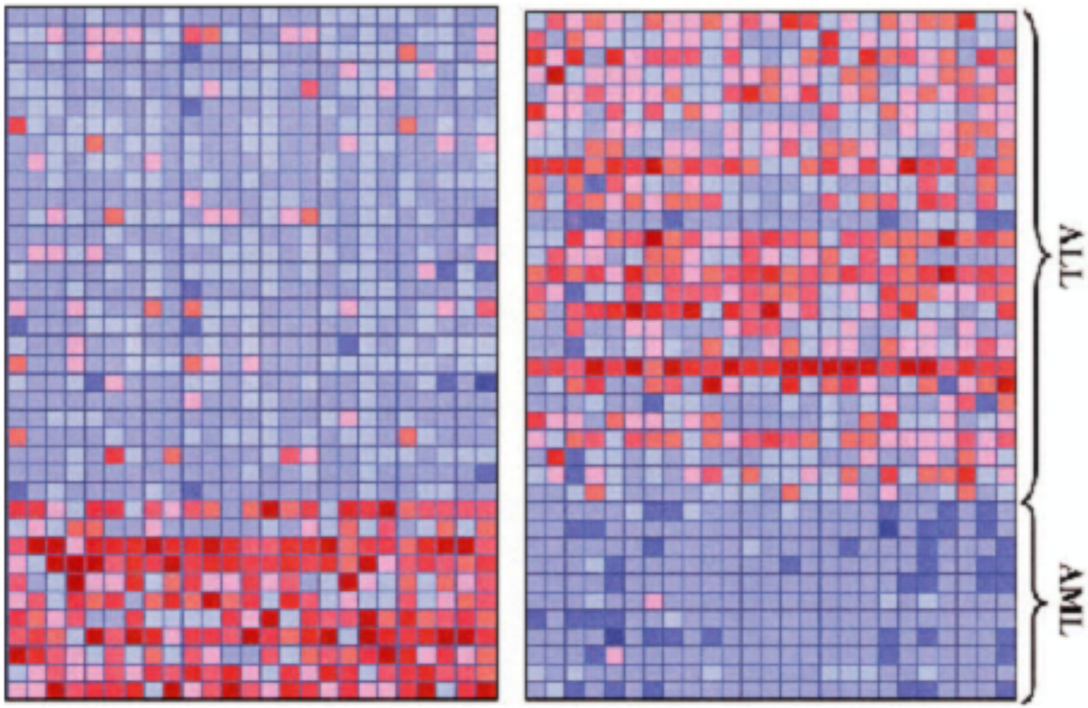
n

<100

-1

+1

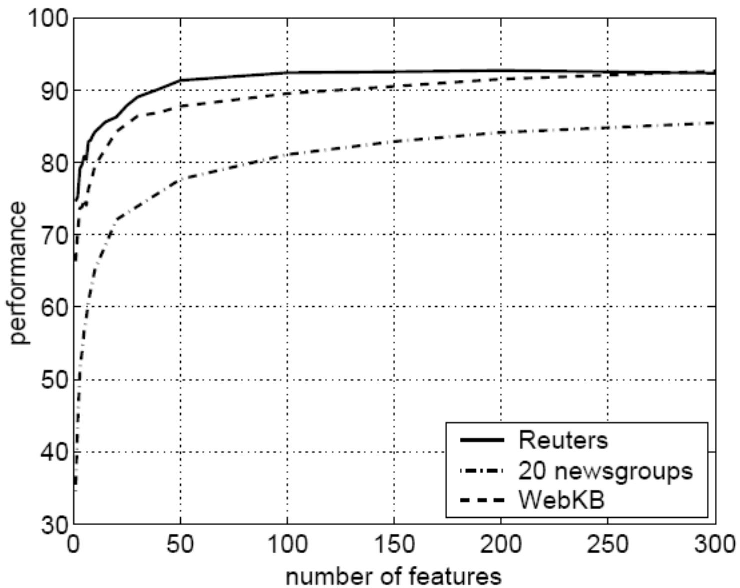
$\{-y_i\}$




Golub et al, Science Vol 286:15 Oct. 1999

$\{y_i\}$,

# e.g., Text Categorization with feature Filtering




**Reuters**: 21578 news wire, 114 semantic categories.

**20 newsgroups**: 19997 articles, 20 categories.

**WebKB**: 8282 web pages, 7 categories.

Bag-of-words: >100,000 features.

Top 3 words of some output Y categories:

- **Alt.atheism**: atheism, atheists, morality
- **Comp.graphics**: image, jpeg, graphics
- **Sci.space**: space, nasa, orbit
- **Soc.religion.christian**: god, church, sin
- **Talk.politics.mideast**: israel, armenian, turkish
- **Talk.religion.misc**: jesus, god, jehovah

***Bekkerman et al, JMLR, 2003***

# Summary: Feature Selection

– Filtering approach:

ranks features or feature subsets independently of the predictor.

- …using univariate methods: consider one variable at a time
- …using multivariate methods: consider more than one variables at a time

– Wrapper approach:

uses a predictor to assess (many) features or feature subsets.

– Embedding approach:

uses a predictor to build a (single) model with a subset of features that are internally selected.

# Nomenclature

- **Univariate method**: considers one variable (feature) at a time.
- **Multivariate method:** considers subsets of variables (features) together.
- **Filter method:** ranks features or feature subsets independently of the predictor.
- **Wrapper method:** uses a predictor to assess features or feature subsets.

# Today

- ❑ Feature Selection
  - ✓ General Introduction
  - ✓ Filtering
  - ✓ Wrapper
  - ✓ Embedded Method

# (I) Filtering

– Filtering approach:

ranks features or feature subsets independently of the predictor.

- …using univariate methods: consider one variable at a time
- …using multivariate methods: consider more than one variables at a time

# (I) Filtering: Univariate: e.g., Pearson Correlation

- Pearson correlation coefficient

$$r(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\times\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

- Measuring the linear **correlation** between two variables: x and y,
- giving a value between +1 and −1 inclusive, where 1 is total positive **correlation**, 0 is no **correlation**, and −1 is total negative **correlation**.

$$\text{where } \bar{x}=\tfrac{1}{n}\sum_{i=1}^{n}x_i \text{ and } \bar{y}=\tfrac{1}{n}\sum_{i=1}^{n}y_i.$$

$$|r(x,y)|\leq 1$$

Correlation is unit independent

# (I) Filtering: Univariate: e.g., Pearson Correlation

- Pearson correlation coefficient

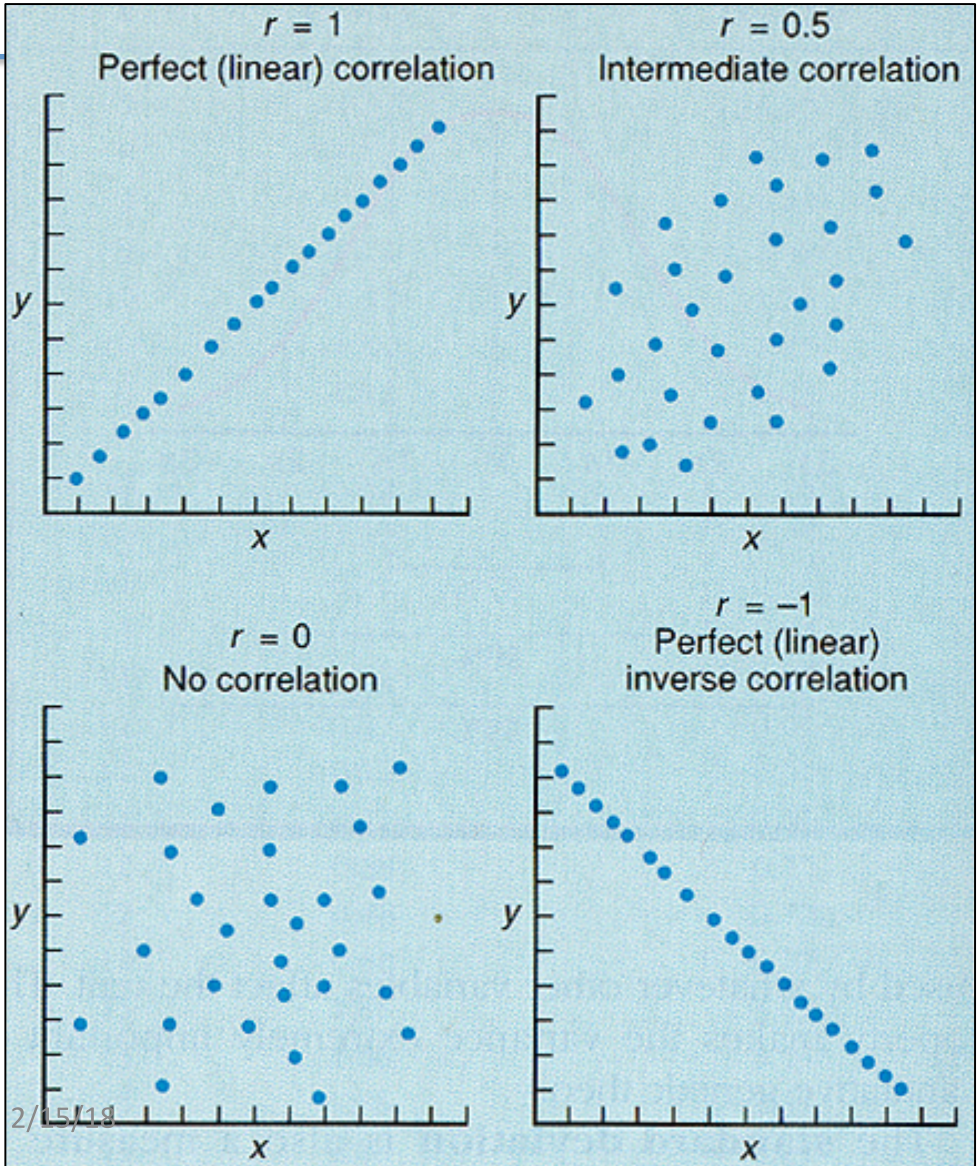
$$r(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\times\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$|r(x,y)|\le 1$$

where $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ and $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$.

- Special case: cosine distance $s(x,y)=\frac{\vec{x}\cdot\vec{y}}{|\vec{x}|\cdot|\vec{y}|}$

# (I) Filtering: Univariate: e.g., Pearson Correlation

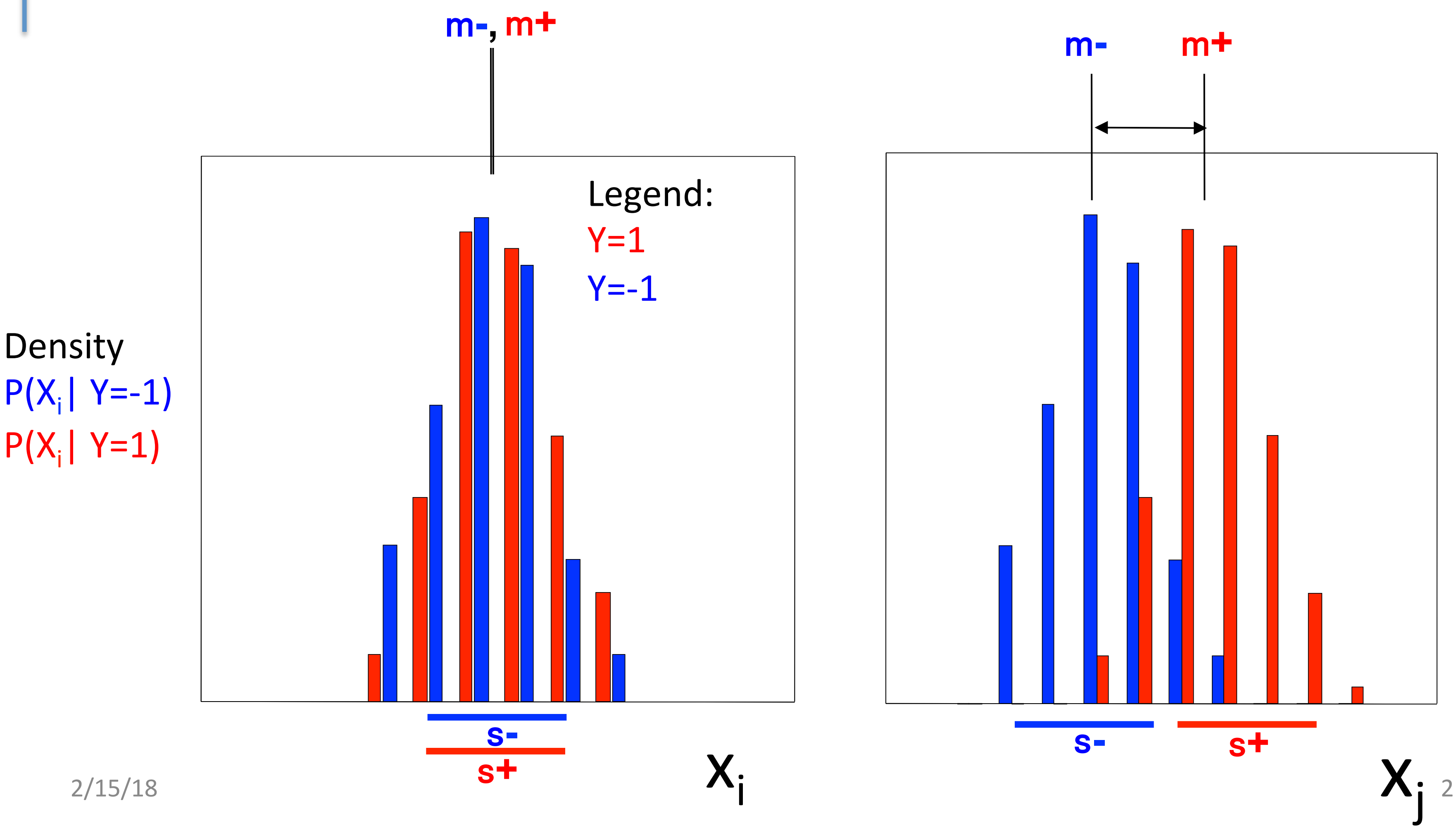



- can only detect linear dependencies between two variables
- (e.g. between one feature vs. target)

# (I) Filtering: univariate filtering

## e.g. T-test

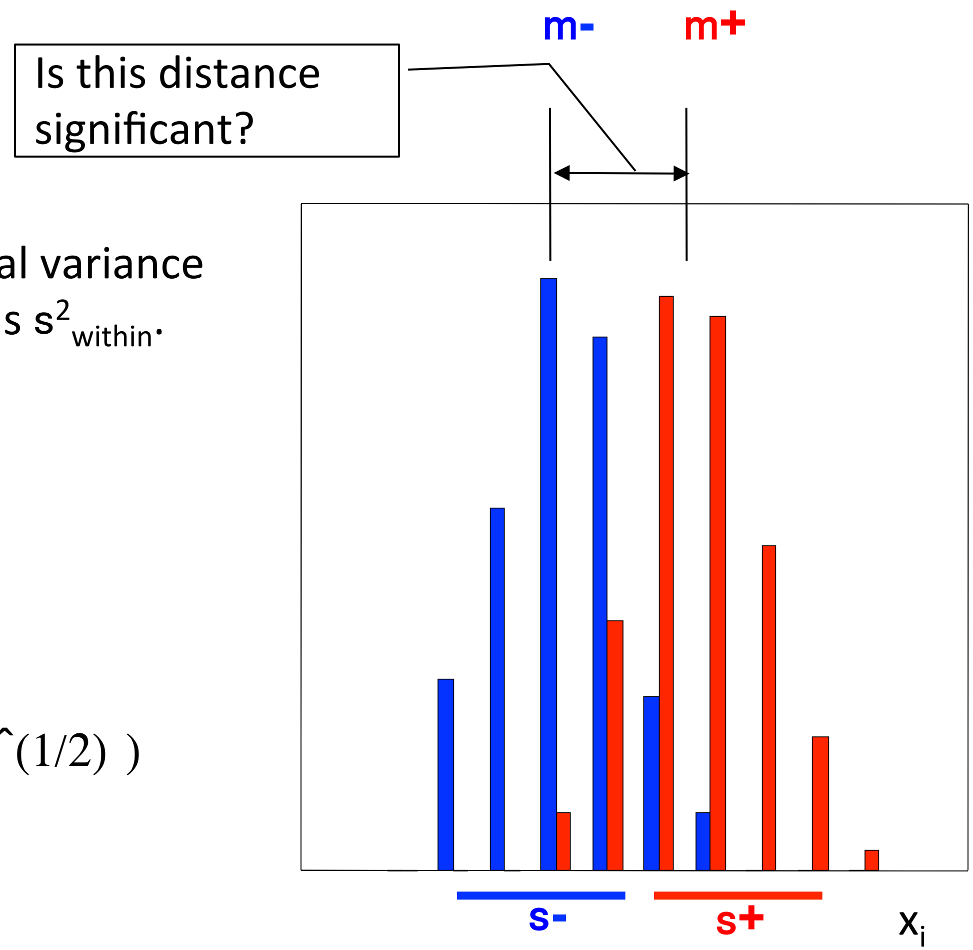
- Goal: determine the relevance of a given single feature for two classes of samples.



# (I) Filtering: univariate filtering

## e.g. T-test

### T-test




From Dr. **Isabelle Guyon**

- Normally distributed classes, equal variance $s^2$ unknown; estimated from data as $s^2_{within}$.
- Null hypothesis $H_0$: m+ = m-
- T statistic:

If $H_0$ is true, then

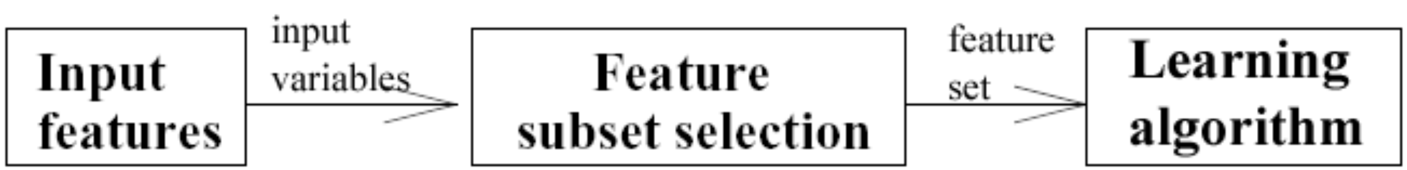
$$t = (m+ - m-)/(s_{within}(1/|m^+| + 1/|m^-|)^{(1/2)})$$

$$\sim \text{Student}(m^+ + m^- - 2 \text{ d.f.})$$

# (I) Filtering : multi-variate: Feature Subset Selection

## Filter Methods

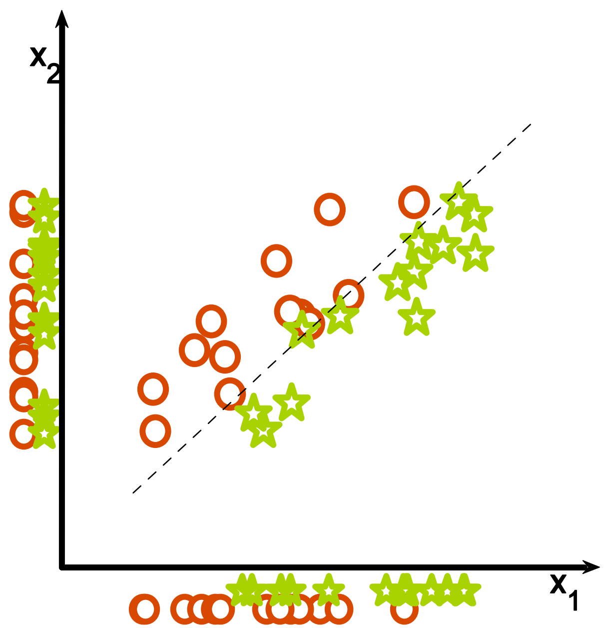
- Select subsets of variables as a pre-processing step, independently of the used classifier!!



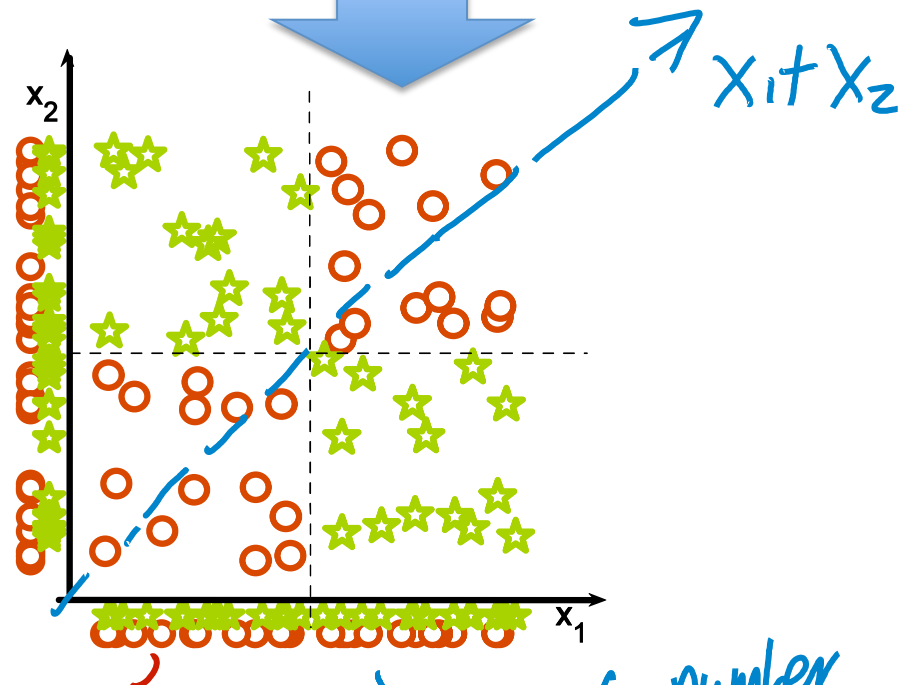

- E.g. Group correlation
- E.g. Information theoretic filtering methods such as Markov blanket

# (I) Filtering : multi-variate: Feature Subset Selection

Binary Classification Task




Univariate selection may fail




on x1 number line, two classes totally overlap!

Guyon-Elisseeff, JMLR 2004; Springer 2006

# (I) Filtering : multi-variate: Feature Subset Selection

e.g. amazon review text $\underbrace{x}$ → $y$ review score 1~5

many possible features

- words: good, not, boring, ....
- 2gram: not good, not boring, ....
- 3grams: very good, very very good, not very boring,
- ⋮
- k grams ......

# (I) Filtering : multi-variate: Feature Subset Selection

- You need:
  - a measure for assessing the goodness of a feature subset (scoring function)
  - a strategy to search the space of possible feature subsets
- Finding a minimal optimal feature set for an arbitrary target concept is NP-hard

  => Good heuristics are needed!



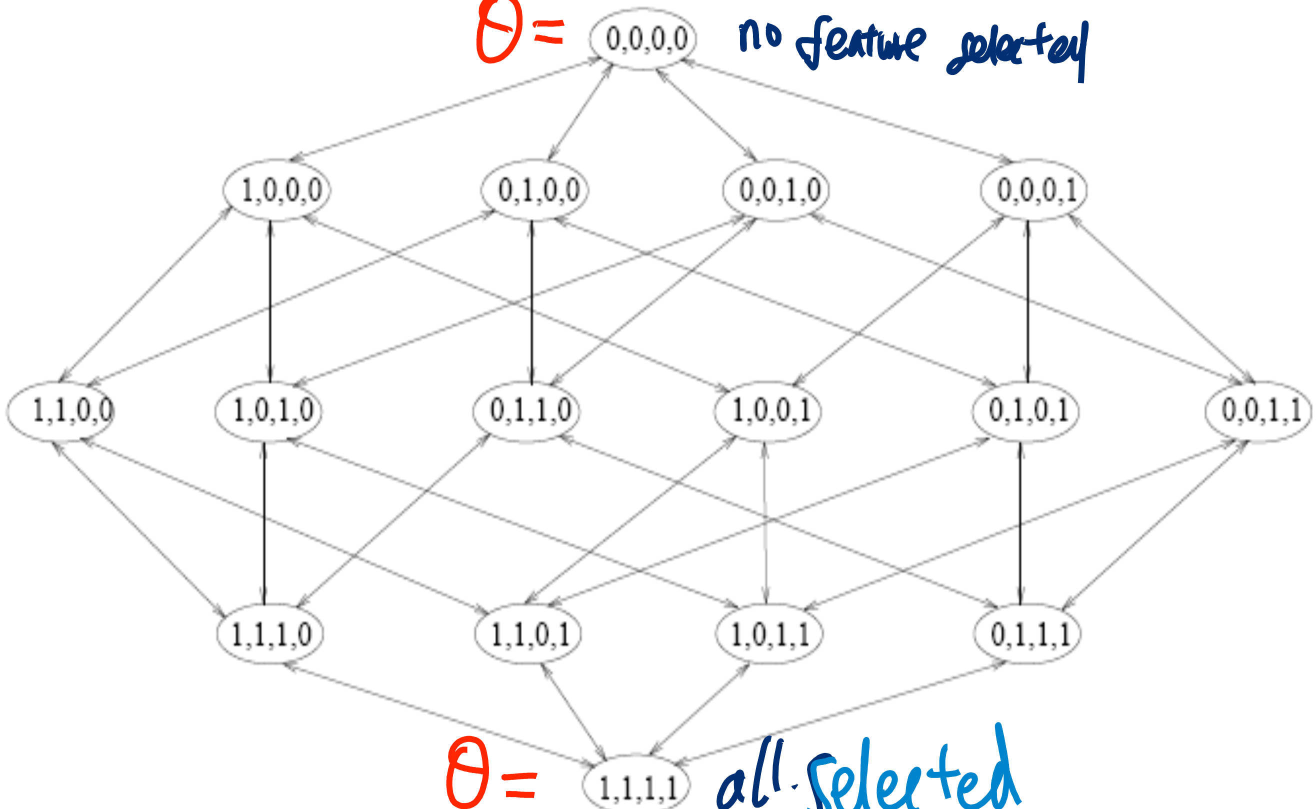


[9] E. Amaldi, V. Kann: The approximability of minimizing nonzero variables and unsatisfied relations in linear systems. (1997)

each feature subset can be described by $\Rightarrow \theta = [0/1, 0/1, 0/1, \ldots, 0/1]^T$

$p \times 1$ Vector

$\theta =$ no feature selected




$\theta =$ all selected

p features, $\underline{2^p}$ possible feature subsets!

# (I) Filtering : Summary

## Filter Methods

- usually fast
- provide generic selection of features, not tuned by given learner (universal)
- this is also often criticised (feature set not optimized for used learner)
- Often used as a preprocessing step for other methods

# (I) Filtering : (many other choices)

| Method | | X | | | Y | | | Comments |
|---|---|---|---|---|---|---|---|---|
| Name | Formula | B | M | C | B | M | C | |
| Bayesian accuracy | Eq. 3.1 | + | s | | + | s | | Theoretically the golden standard, rescaled Bayesian relevance Eq. 3.2. |
| Balanced accuracy | Eq. 3.4 | + | s | | + | s | | Average of sensitivity and specificity; used for unbalanced dataset, same as AUC for binary targets. |
| Bi-normal separation | Eq. 3.5 | + | s | | + | s | | Used in information retrieval. |
| F-measure | Eq. 3.7 | + | s | | + | s | | Harmonic of recall and precision, popular in information retrieval. |
| Odds ratio | Eq. 3.6 | + | s | | + | s | | Popular in information retrieval. |
| Means separation | Eq. 3.10 | + | i | + | + | | | Based on two class means, related to Fisher's criterion. |
| T-statistics | Eq. 3.11 | + | i | + | + | | | Based also on the means separation. |
| Pearson correlation | Eq. 3.9 | + | i | + | + | i | + | Linear correlation, significance test Eq. 3.12, or a permutation test. |
| Group correlation | Eq. 3.13 | + | i | + | + | i | + | Pearson's coefficient for subset of features. |
| $\chi^2$ | Eq. 3.8 | + | s | | + | s | | Results depend on the number of samples $m$. |
| Relief | Eq. 3.15 | + | s | + | + | s | + | Family of methods, the formula is for a simplified version ReliefX, captures local correlations and feature interactions. |
| Separability Split Value | Eq. 3.41 | + | s | + | + | s | | Decision tree index. |
| Kolmogorov distance | Eq. 3.16 | + | s | + | + | s | + | Difference between joint and product probabilities. |
| Bayesian measure | Eq. 3.16 | + | s | + | + | s | + | Same as Vajda entropy Eq. 3.23 and Gini Eq. 3.39. |
| Kullback-Leibler divergence | Eq. 3.20 | + | s | + | + | s | + | Equivalent to mutual information. |
| Jeffreys-Matusita distance | Eq. 3.22 | + | s | + | + | s | + | Rarely used but worth trying. |
| Value Difference Metric | Eq. 3.22 | + | s | | + | s | | Used for symbolic data in similarity-based methods, and symbolic feature-feature correlations. |
| Mutual Information | Eq. 3.29 | + | s | + | + | s | + | Equivalent to information gain Eq. 3.30. |
| Information Gain Ratio | Eq. 3.32 | + | s | + | + | s | + | Information gain divided by feature entropy, stable evaluation. |
| Symmetrical Uncertainty | Eq. 3.35 | + | s | + | + | s | + | Low bias for multivalued features. |
| J-measure | Eq. 3.36 | + | s | + | + | s | + | Measures information provided by a logical rule. |
| Weight of evidence | Eq. 3.37 | + | s | + | + | s | + | So far rarely used. |
| MDL | Eq. 3.38 | + | s | | + | s | | Low bias for multivalued features. |

***Guyon-Elisseeff, JMLR 2004; Springer 2006***

# Today

- ❑ Feature Selection
  - ✓ General Introduction
  - ✓ Filtering
  - ✓ Wrapper
  - ✓ Embedded Method

# (2) Wrapper

– Wrapper approach:

uses a predictor to assess (many) features or feature subsets.

# (2) Wrapper : Feature Subset Selection

## Wrapper Methods

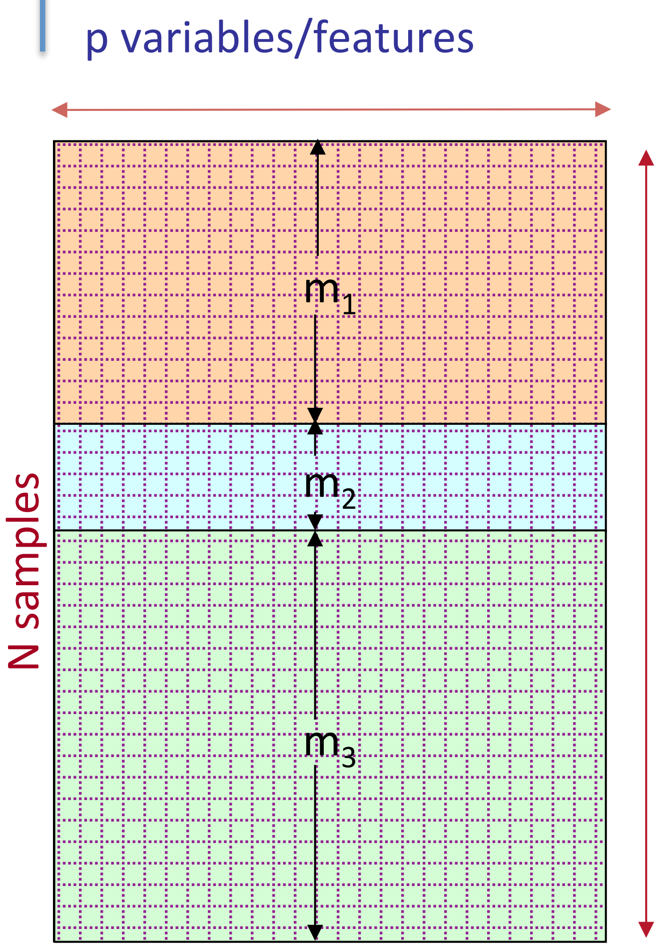
- Learner is considered a black-box
- Interface of the black-box is used to score subsets of variables according to the predictive power of the learner when using the subsets.

- Results vary for different learners

# (2) Wrapper : Feature Subset Selection

- Two major questions to answer:
  - (a). Assessment: How to asses performance of a learner that uses a particular feature subset ?
  - (b). Search: How to search in the space of all feature subsets ?

# (a). Assessment: **feature subset assessment (for wrapper approach)**




Split data into 3 sets:
training, validation, and test set.

1) For each feature subset, train predictor on training data.

2) Select the feature subset, which performs best on validation data.
   - Repeat and average if you want to reduce variance (cross-validation).

3) Test on test data.

**Danger of over-fitting** with intensive search!

From Dr. **Isabelle Guyon**

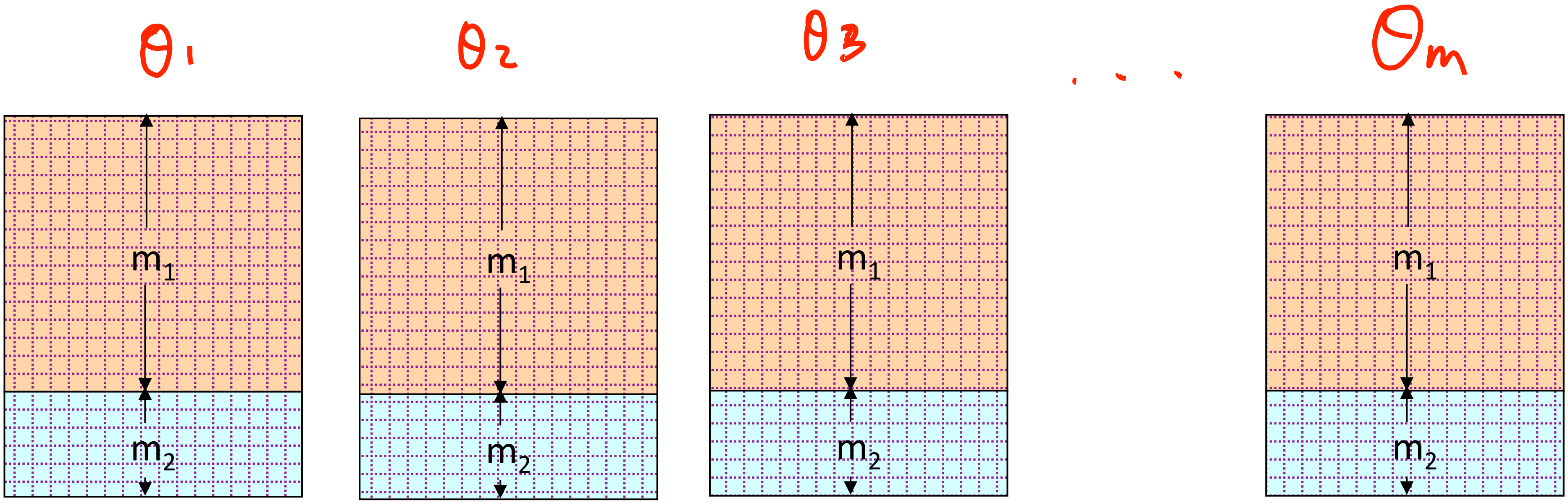
train data : $\underset{\beta_{\theta^{(j)}}}{\text{argmin}}\; J(\beta_{\theta^{(j)}}) \Rightarrow \beta^*_{\theta^{(j)}}$

validation data : $\underset{\{\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(m)}\}}{\text{argmin}}$ PredictLoss$(\beta^*_\theta) \Rightarrow \beta^*_{\theta^*}$

m usually $\ll 2^p$

test data: $\underbrace{\text{PredictLoss}(\beta^*_{\theta^*})}_{\text{to report}}$

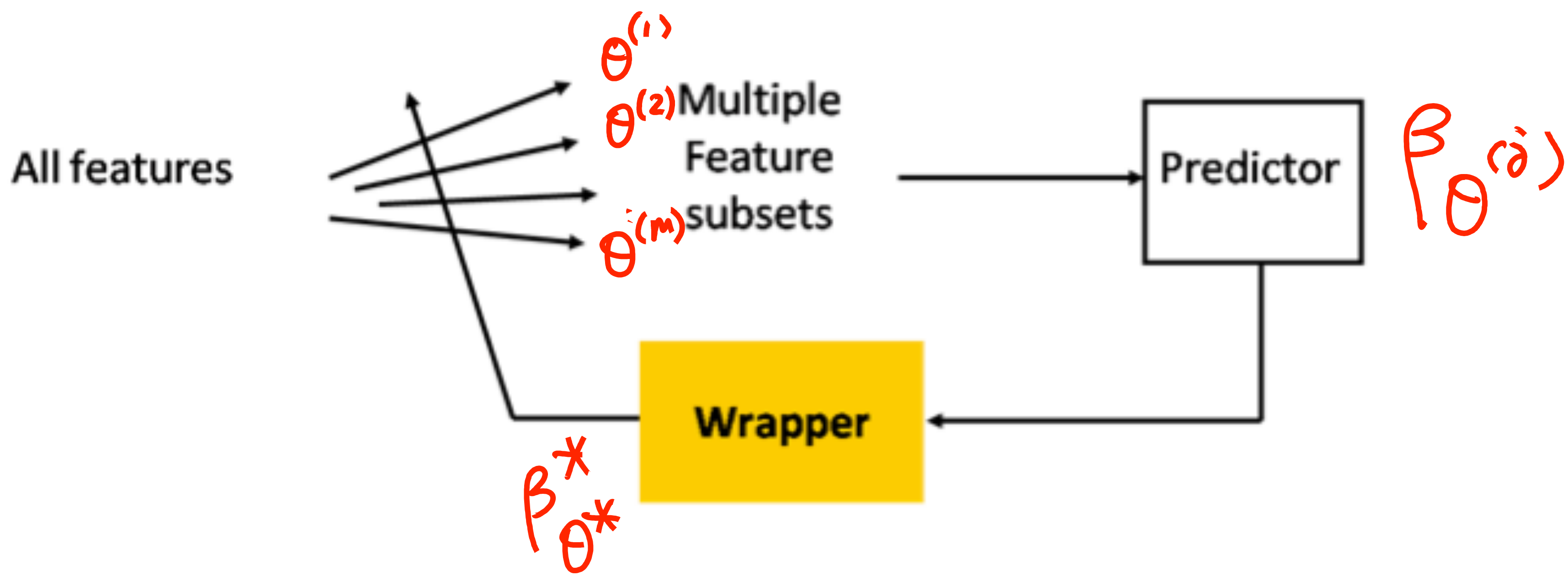
# (a). Assessment: How to access multiple candidates of feature subsets




train for m times on train fold
test for m times on validation fold

# (a). Assessment: **How to access multiple candidates of feature subsets**

## Wrapper Methods




Att: might run iteratively for T times

# (b). Search: How to search the space of all feature subsets ?

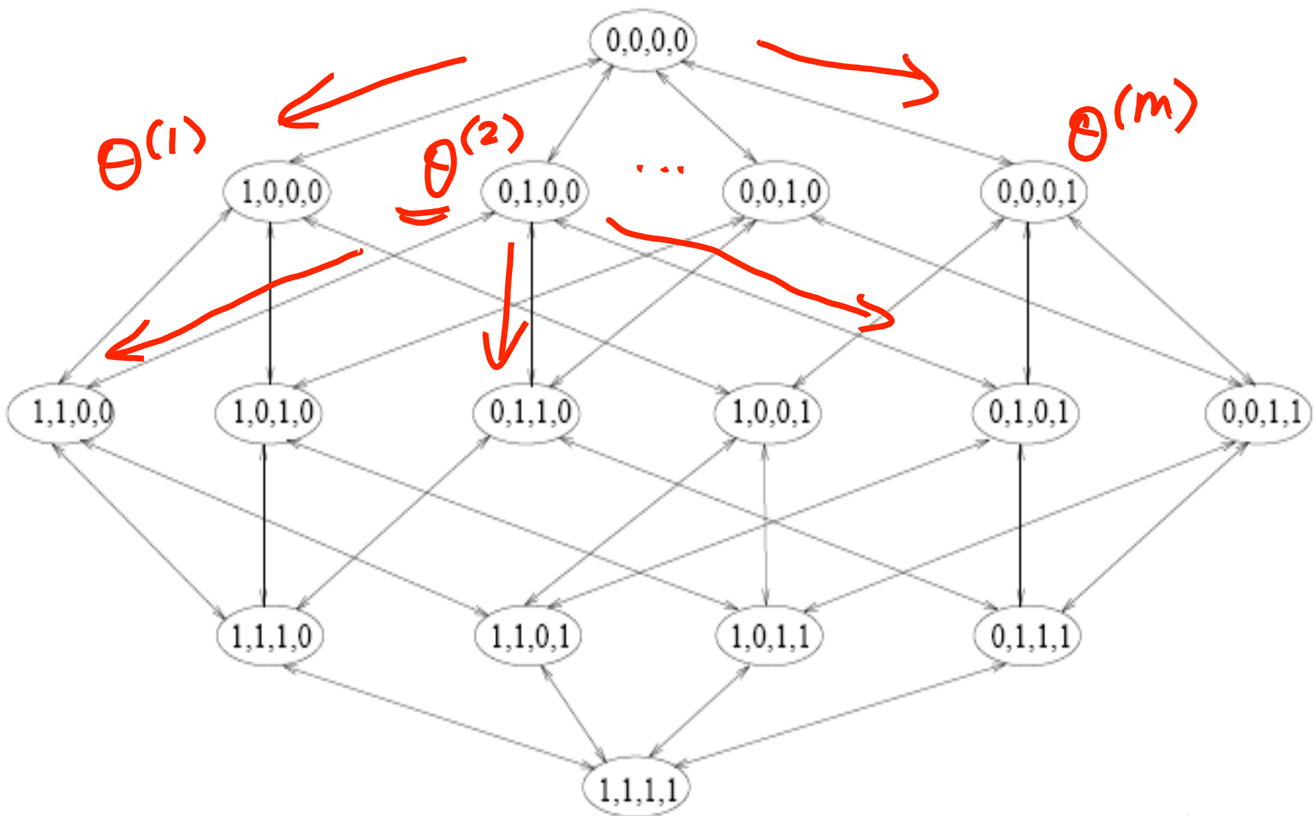
## Wrapper Methods

- The problem of finding the optimal subset is NP-hard!

- A wide range of heuristic search strategies can be used. Two different classes:
  - **Forward selection**
    (start with empty feature set and add features at each step)
  - **Backward elimination**
    (start with full feature set and discard features at each step)

- predictive power is usually measured on a validation set or by cross-validation
- By using the learner as a black box wrappers are universal and simple!
- Criticism: a large amount of computation is required.

# (b). Search: How to search the space of all feature subsets ?




Step 1:

Step2:

⋮

StepT

$$2^p \Rightarrow p(p-1)(p-2)\cdots(p-T)$$

# (b). Search: even more search strategies for selecting feature subset

- **Forward selection** or **backward elimination.**
- **Beam search:** keep k best path at each step.
- **GSFS:** generalized sequential forward selection – when (n-k) features are left try all subsets of g features. More trainings at each step, but fewer steps.
- **PTA(l,r):** plus l , take away r – at each step, run SFS l times then SBS r times.
- **Floating search**: One step of SFS (resp. SBS), then SBS (resp. SFS) as long as we find better subsets than those of the same size obtained so far.

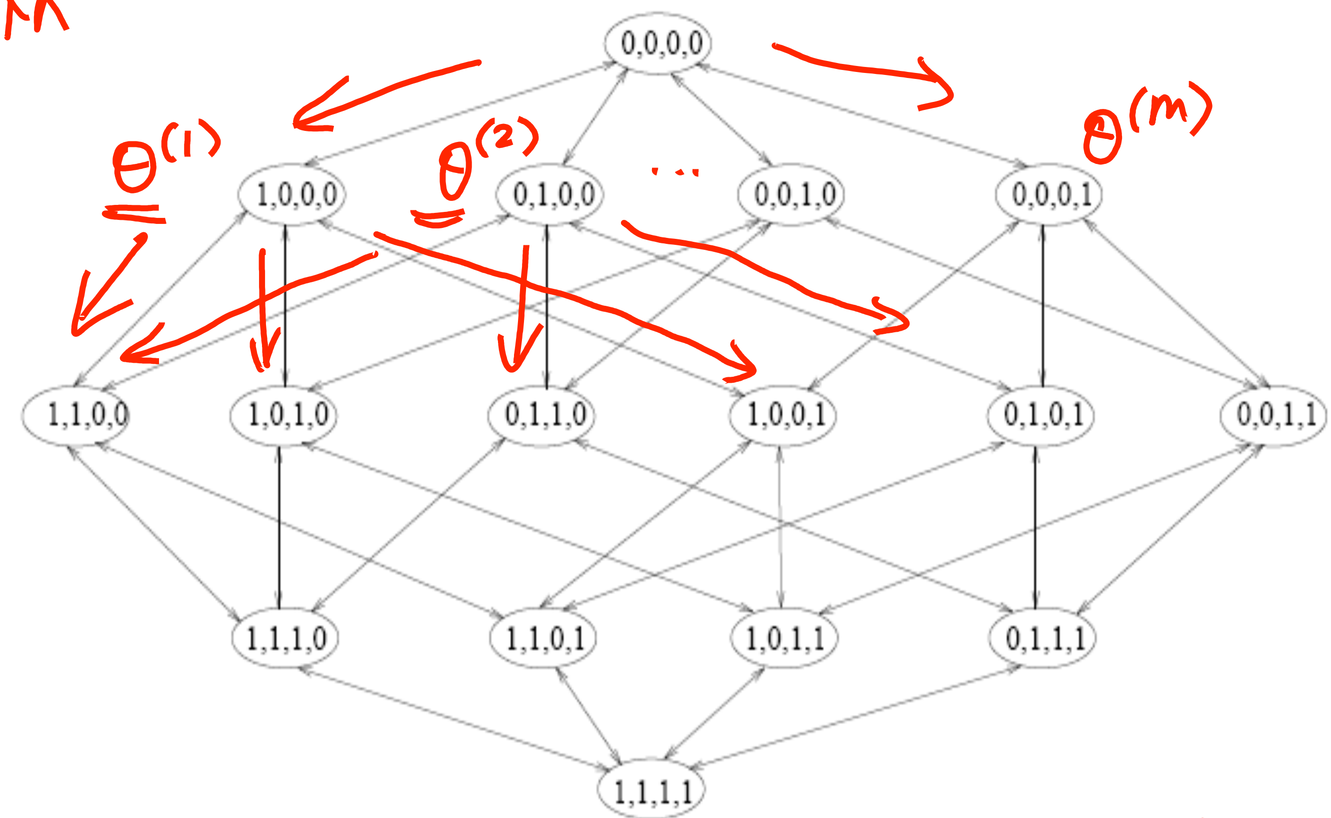
From Dr. **Isabelle Guyon**

# (b). Search: How to search the space of all feature subsets ?

e.g. BEAM search

[keep top k=2 path]

$\theta^{(1)}$ $\theta^{(2)}$ ... $\theta^{(m)}$




Step 1:

Step2:

⋮

Step T

$$2^p \Rightarrow p(p-1)(p-2)\cdots(p-T)$$

# Today

- Feature Selection
  - ✓ General Introduction
  - ✓ Filtering
  - ✓ Wrapper
  - ✓ Embedded Method

# (3) Embedded

– Embedding approach:

uses a predictor to build a (single) model with a subset of features that are internally selected.

# (3) Embedded: **Feature Subset Selection**

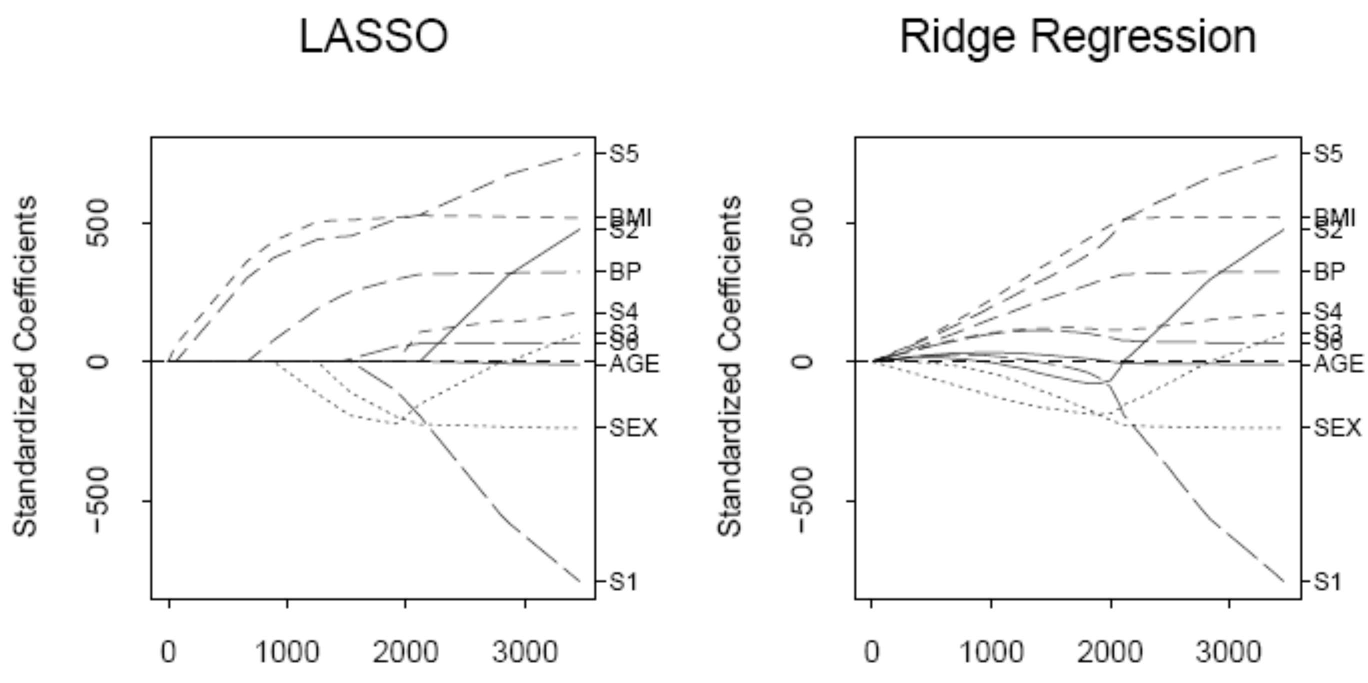
Embedded Methods

- Specific to a given learning machine!
- Performs variable selection (implicitly) in the process of training
- Just train a (single) model

# (3) Embedded: e.g. Feature Selection via Embedded Methods: e.g., $L_1$-regularization

$l_1$ penalty: $y \sim Model(X\beta) + \lambda \sum |\beta_i|$ (lasso)

$l_2$ penalty: $y \sim Model(X\beta) + \lambda \sum \beta_i^2$ (ridge regression)



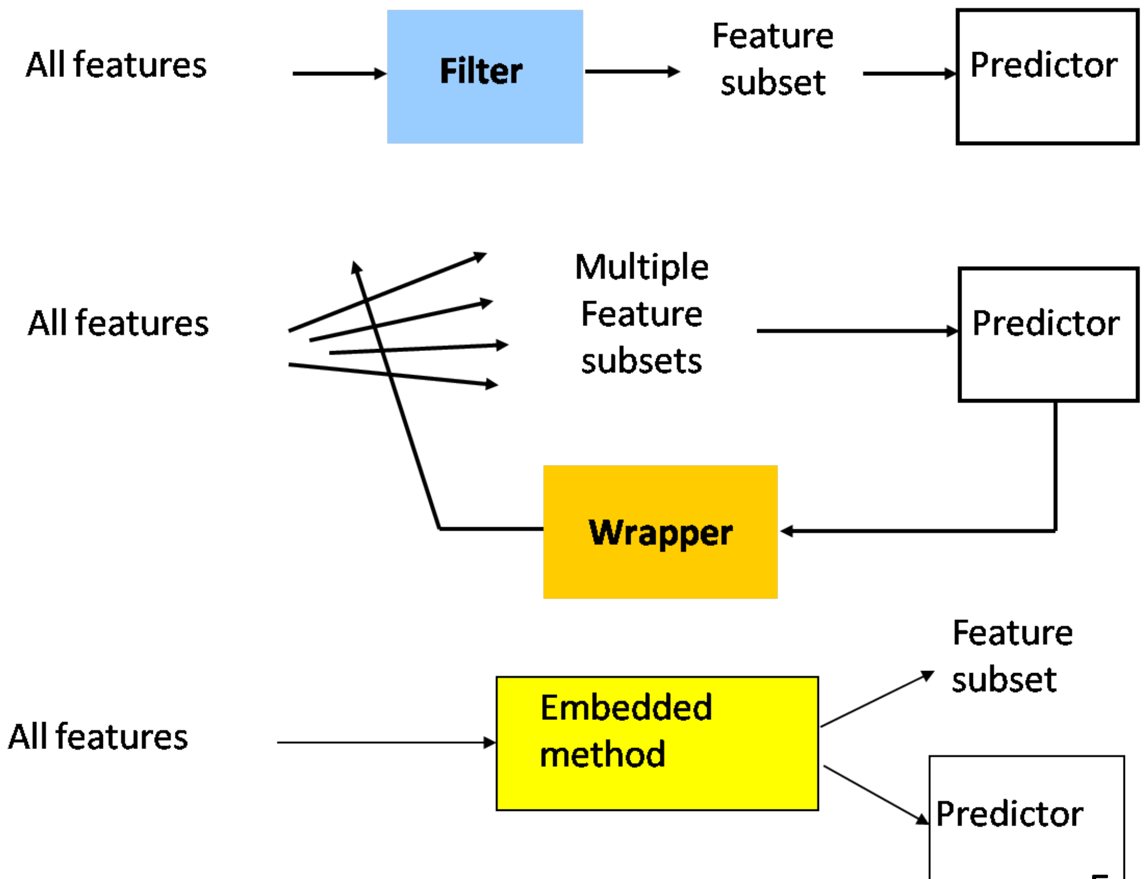





From ESL book

# Summary: filters vs. wrappers vs. embedding

- **Main goal**: rank subsets of useful features




From Dr. **Isabelle Guyon**

# In practice…

- **No method is universally better:**
  - **wide variety of types of variables, data distributions, learning machines, and objectives.**

- **Feature selection is not always necessary to achieve good performance.**

*NIPS 2003 and WCCI 2006 challenges :* http://clopinet.com/challenges



From Dr. **Isabelle Guyon**

# References


- Prof. Andrew Moore's slides
- Hastie, Trevor, et al. *The elements of statistical learning*. Vol. 2. No. 1. New York: Springer, 2009.
- Dr. **Isabelle Guyon's feature selection tutorials**

# Vs. Dimensionality Reduction (Later)

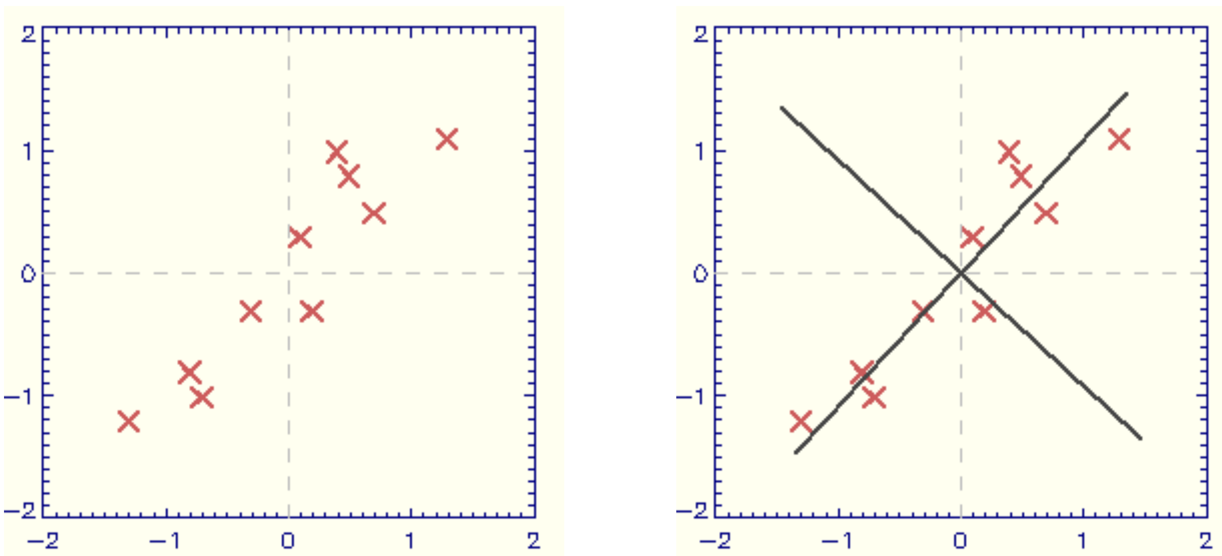
In the presence of many of features, select the most relevant subset of (weighted) combinations of features.

Feature Selection: $X_1,\ldots,X_p \rightarrow X_{k1},\ldots,X_{kp'}$

Dimensionality Reduction: $X_1,\ldots,X_m \rightarrow g_1(X_1,\ldots,X_m),\ldots,g_{p'}(X_1,\ldots,X_m)$

# Dimensionality Reduction: e.g., (Linear) Principal Components Analysis

- **PCA** finds a *linear* mapping of dataset X to a dataset X' of lower dimensionality. The variance of X that is remained in X' is maximal.




Dataset X is mapped to dataset X', here of the same dimensionality. The first dimension in X' (= the first principal component) is the direction of maximal variance. The second principal component is orthogonal to the first.