UVA CS 6316/4501 – Fall 2016 Machine Learning

Lecture 7: Feature Selection

Dr. Yanjun Qi

University of Virginia

Department of
Computer Science

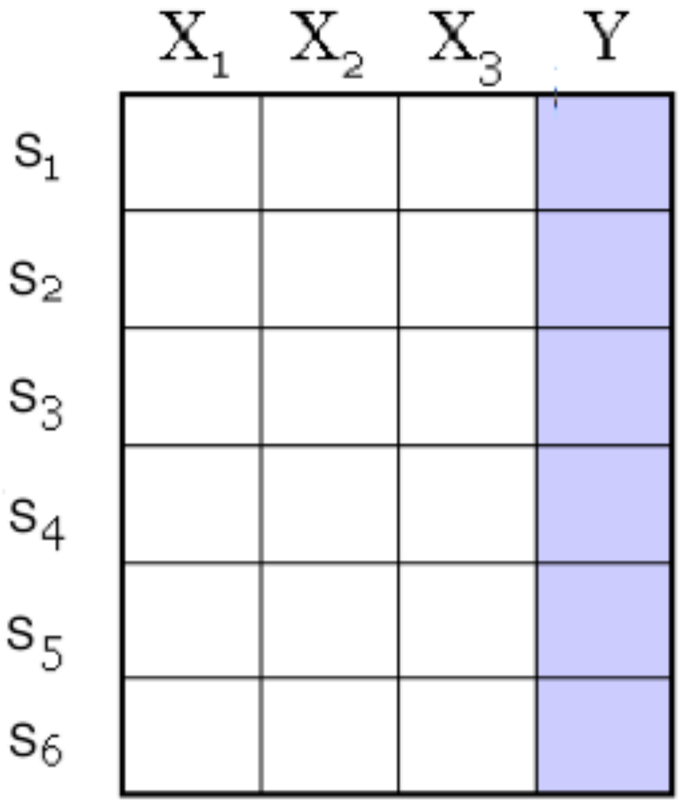
Where are we ? ➔ Five major sections of this course

- ❑ 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



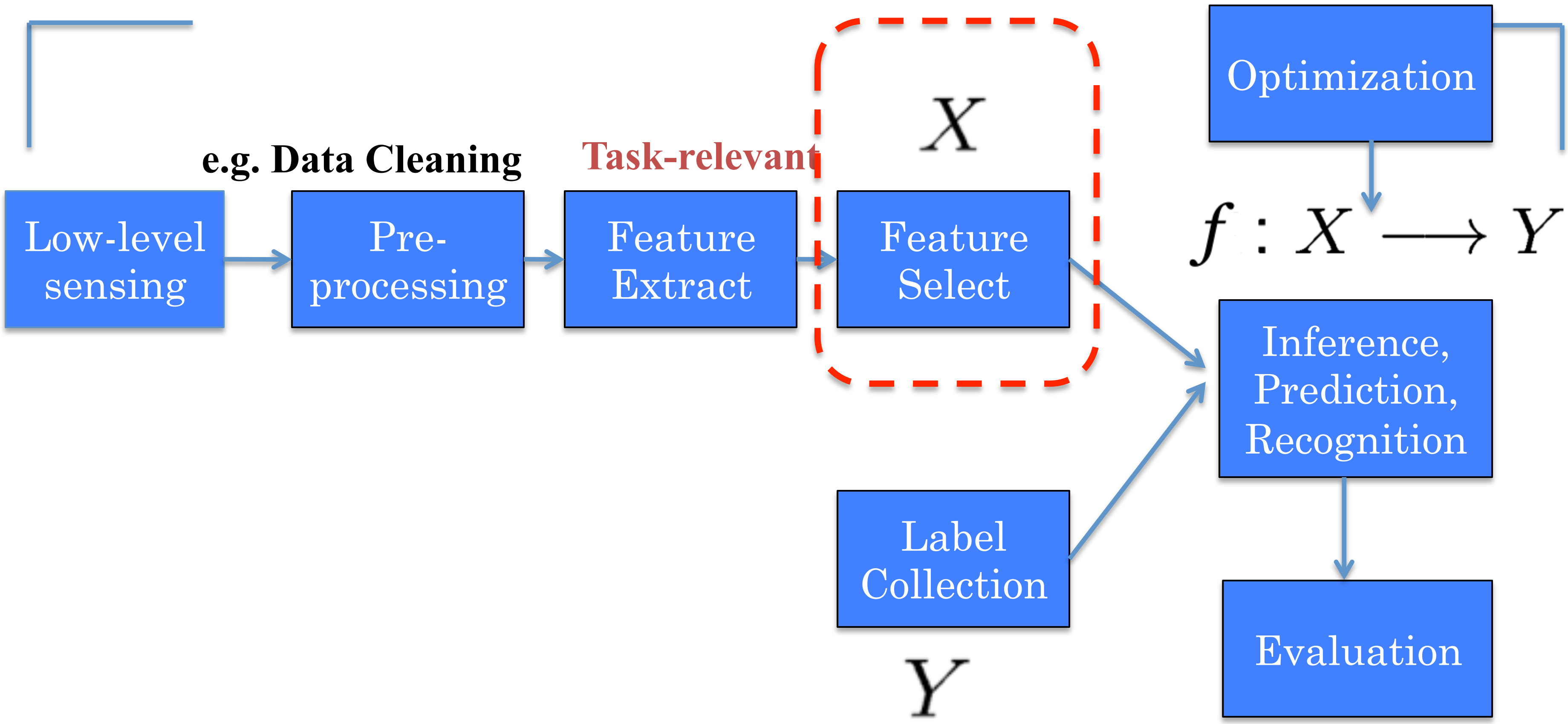
$$f : \boxed{X} \longrightarrow \boxed{Y}$$

- **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

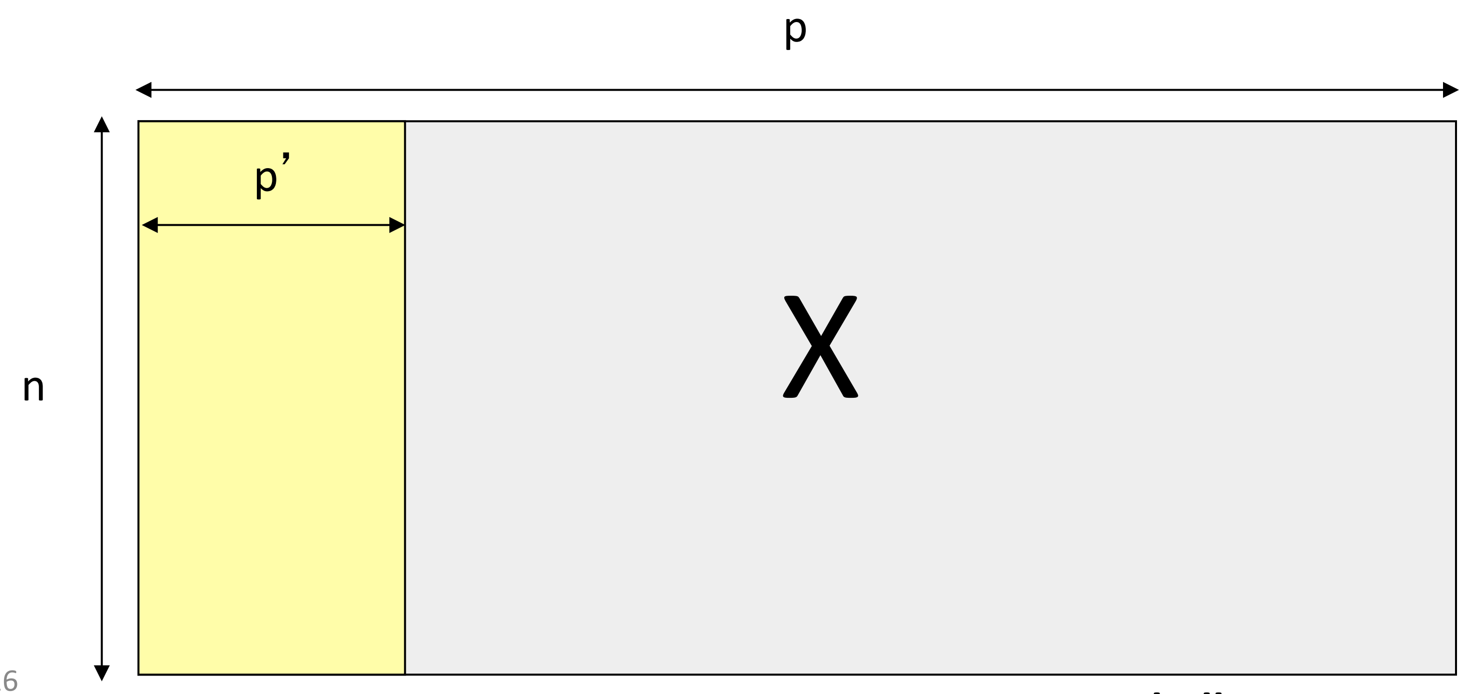
- Feature Selection (supervised)
  - Filtering approach
  - Wrapper approach
  - Embedded methods

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 machines.



From Dr. **Isabelle Guyon**

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

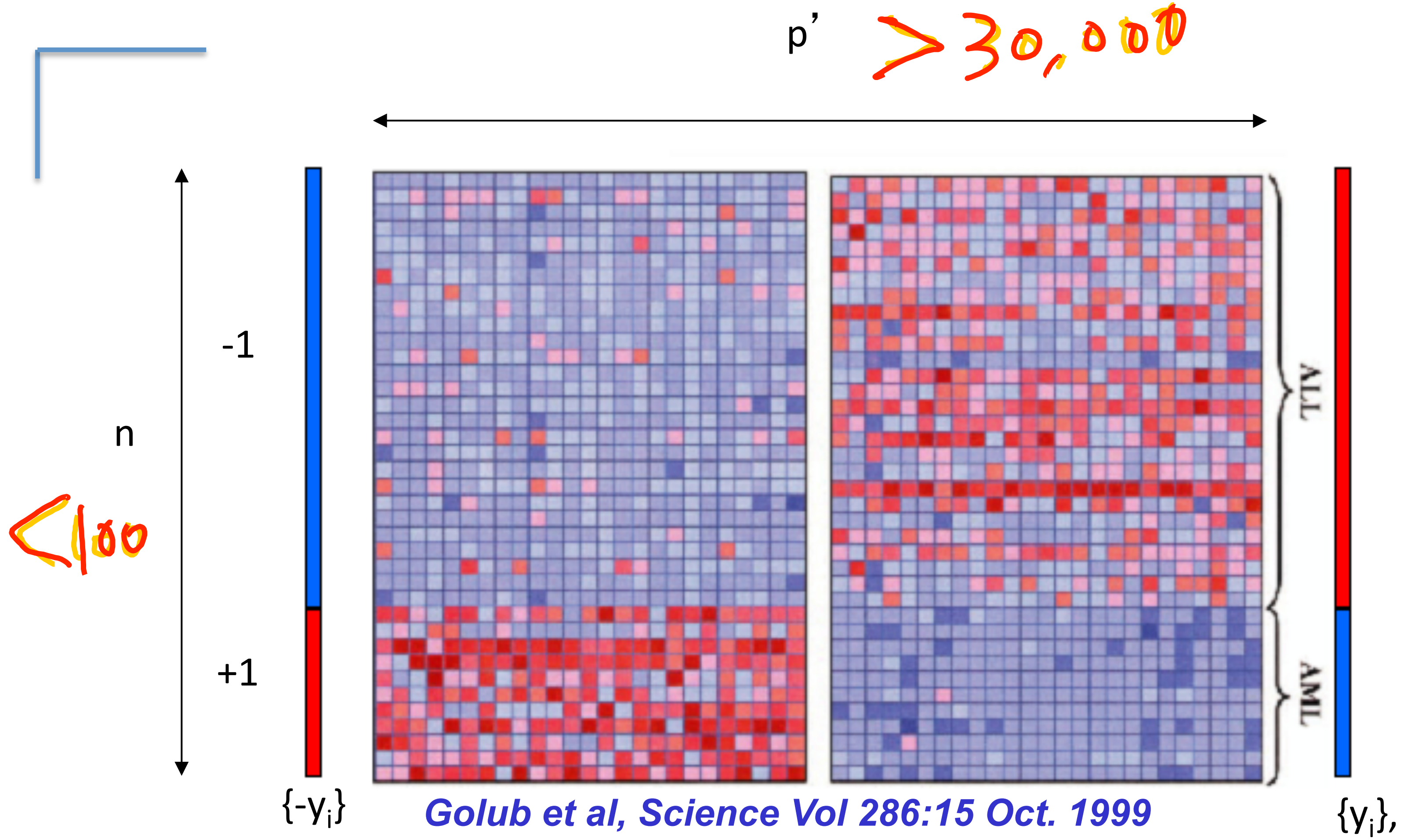
IV. Features

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)

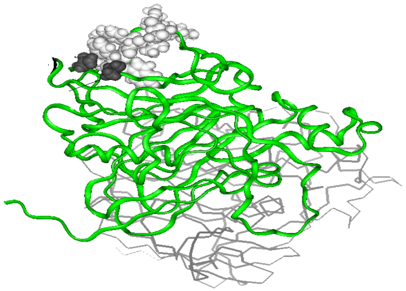
$n \approx 1700$ / $p > 30{,}000$

e.g., Leukemia Diagnosis



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

e.g., QSAR: Drug Screening



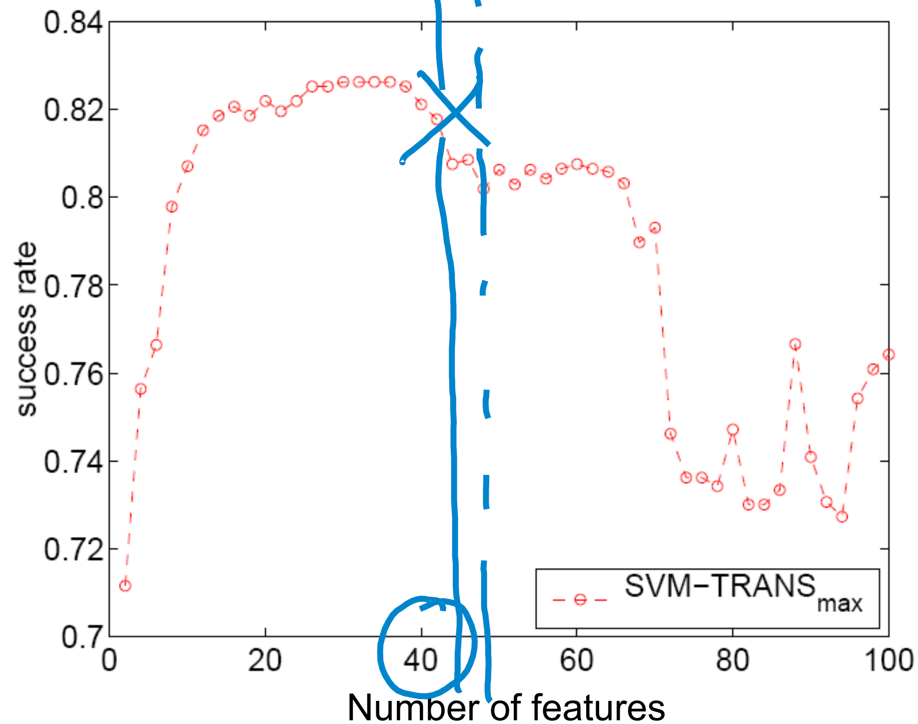
n

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.

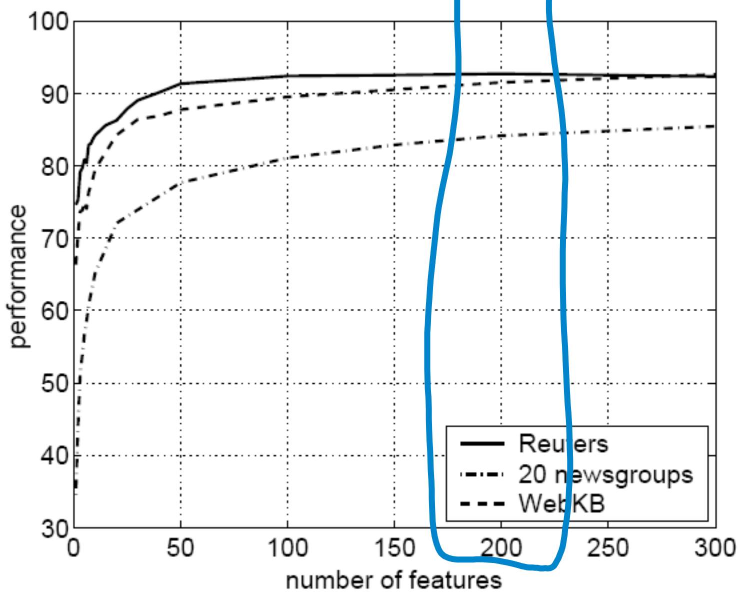
p

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



Weston et al, Bioinformatics, 2002

e.g., Text Categorization with feature Filtering



n

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.

(I) Filtering

– Filtering approach:

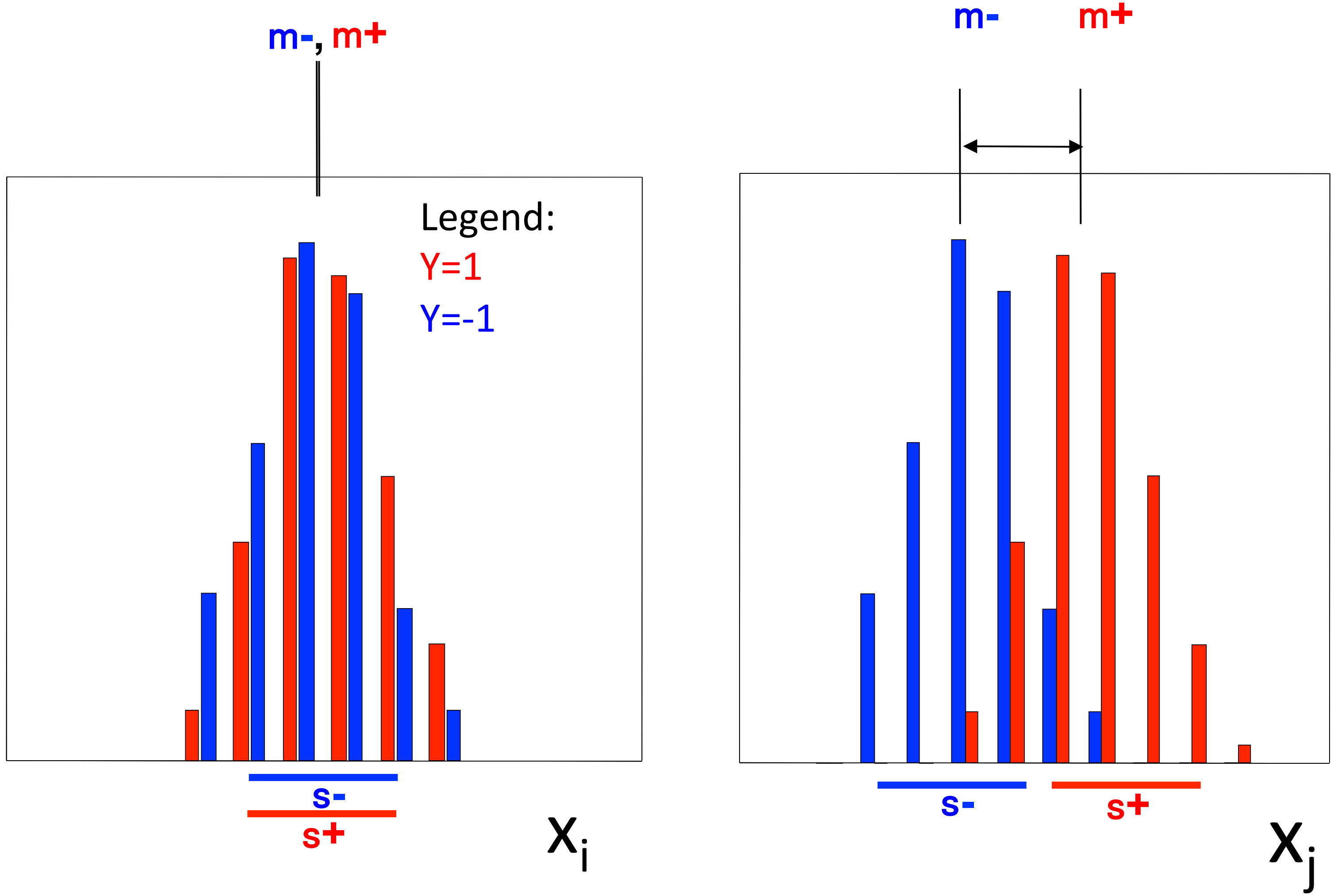
ranks features or feature subsets independently of the predictor.

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

(I) Filtering : univariate filtering approach, e.g. T-test

- Issue: determine the relevance of a given single feature.

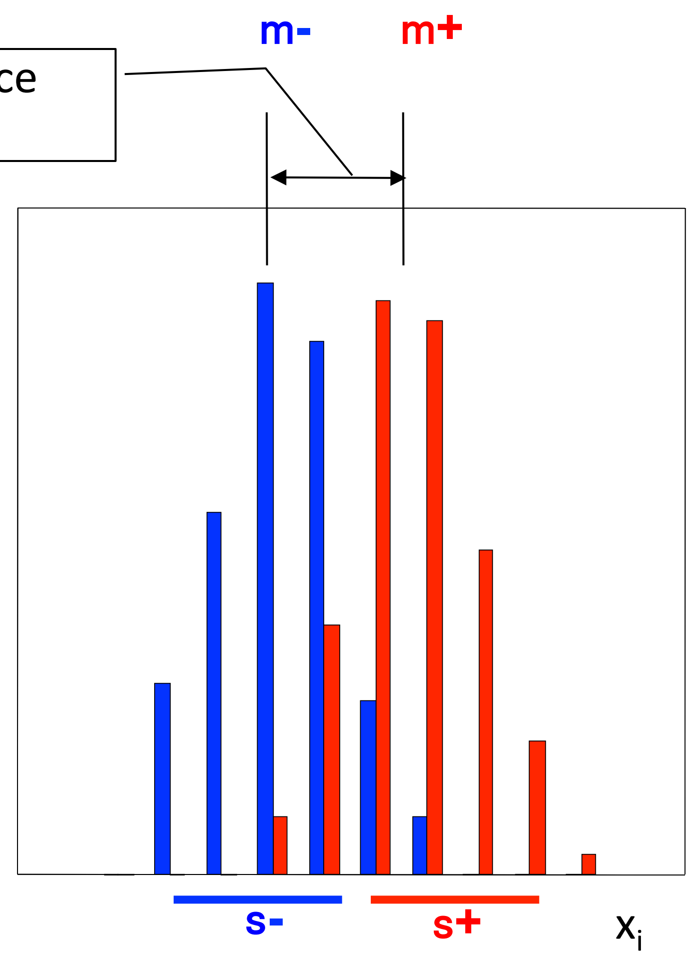
Density
$P(X_i \mid Y=-1)$
$P(X_i \mid Y=1)$



(I) Filtering : univariate filtering approach , e.g. T-test

T-test

Is this distance significant?



- 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

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

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

From Dr. **Isabelle Guyon**

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

- Pearson correlation coefficient

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

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

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

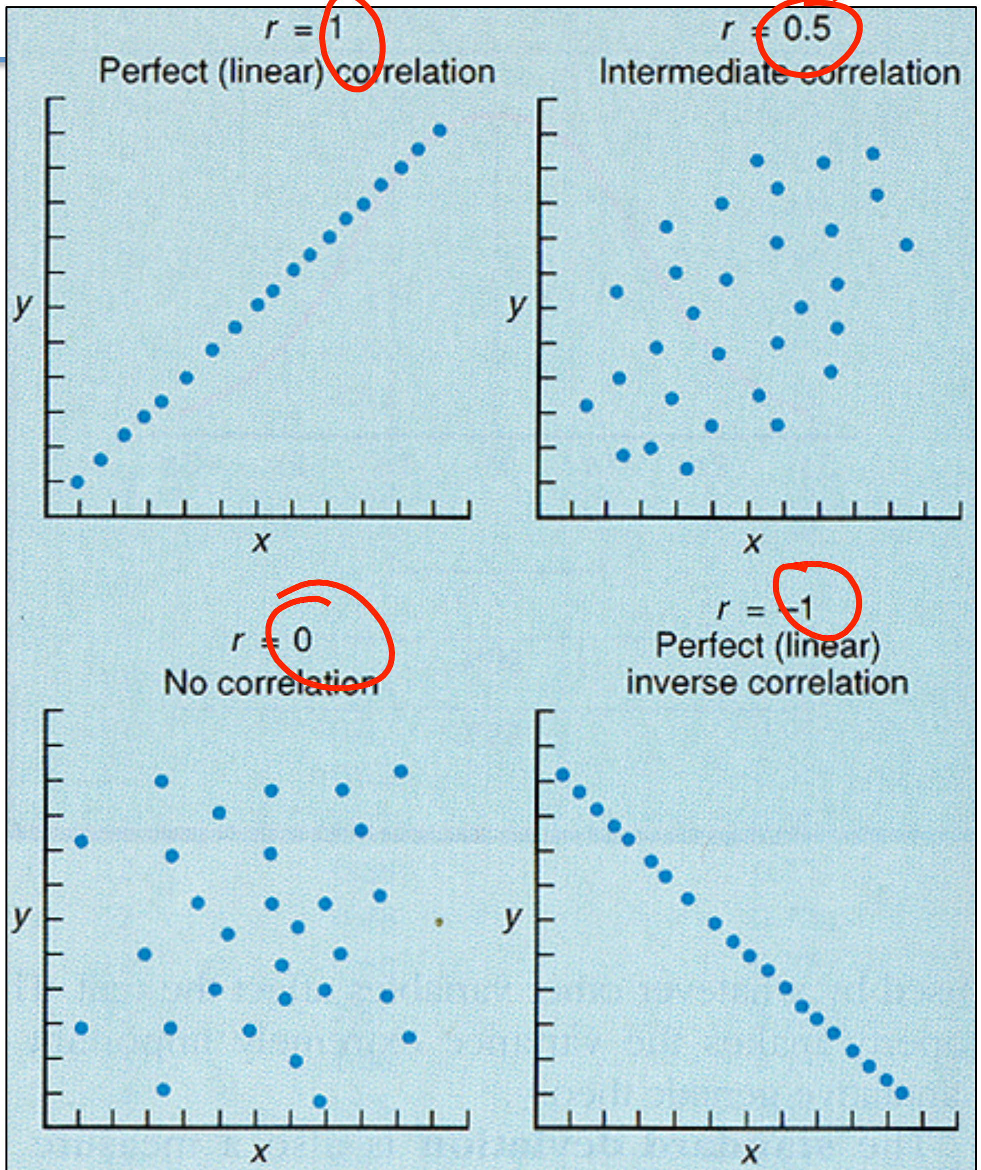
Correlation is unit independent

- 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**.

- 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 variable and target THOUGH

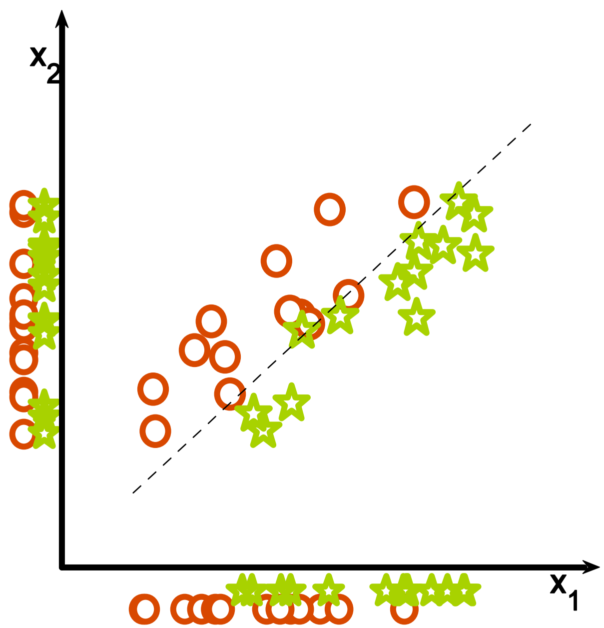
➔ E.g. Mutual information filter to get nonlinear dependencies

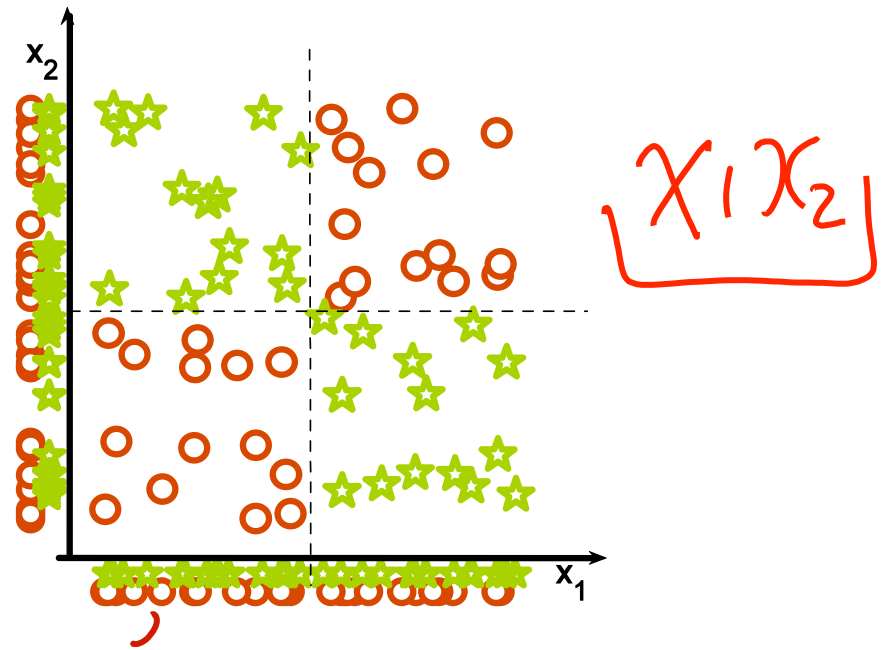
(I) Filtering : univariate filtering, (many other criteria)

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.
χ^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.

(I) Filtering : multivariate approach

Univariate selection may fail





Guyon-Elisseeff, JMLR 2004; Springer 2006

multivariate approach

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

many possible features

- words
- 2gram
- 3grams
- ⋮
- k grams

good, not, boring, ...

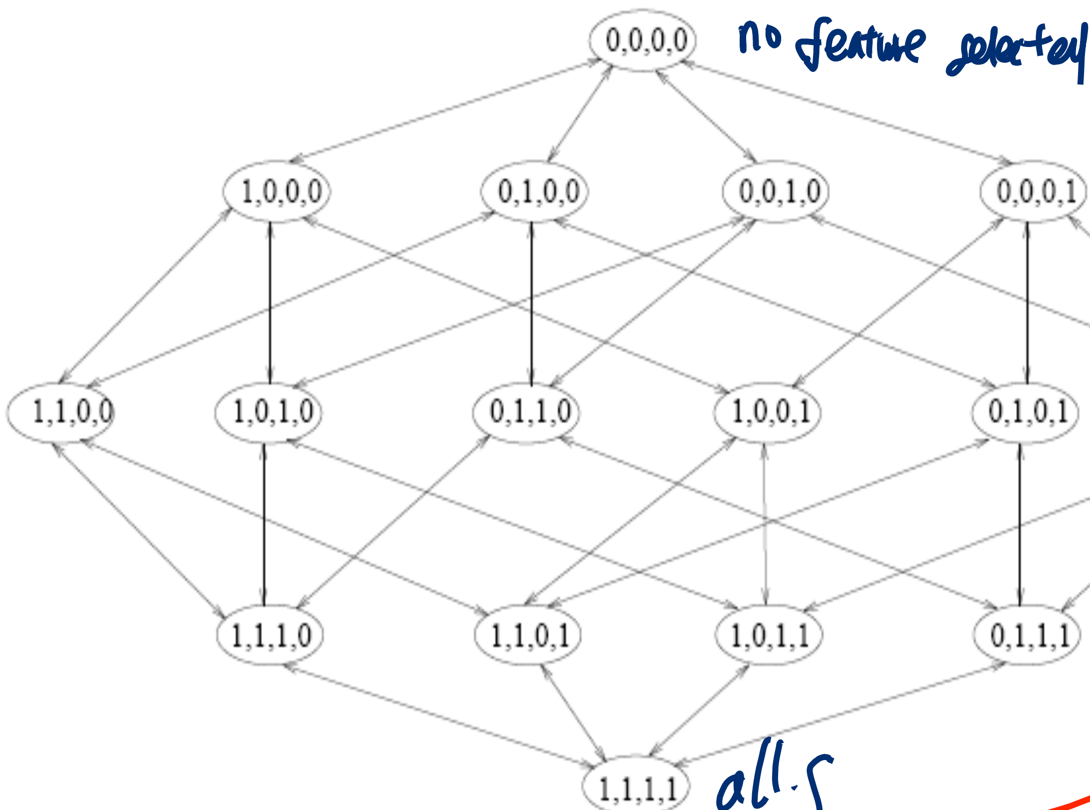
not good, not boring, ...

p features

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

$p \times 1$ Vector

Feature Selection: search strategies



p features, 2^p possible feature subsets!

(I) Filtering : Feature Subset Selection

- You need:
  - a measure for assessing the goodness of a feature subset (scoring function)

    2^p

  - a strategy to search the space of possible feature subsets

- Finding a minimal optimal feature set for an arbitrary target concept is NP-hard

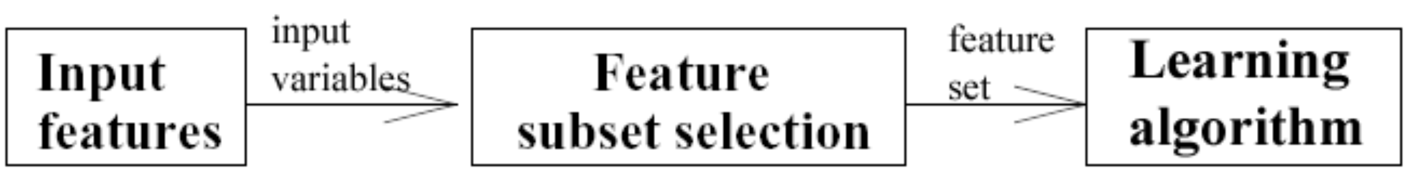
  => Good heuristics are needed!

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

(I) Filtering : Feature Subset Selection

Filter Methods

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



(I) Filtering : Feature Subset Selection

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)
- sometimes used as a preprocessing step for other methods

(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
- One needs to define:
  - (a). how to search the space of all possible variable subsets ?
  - (b). how to assess the prediction performance of a learner ?

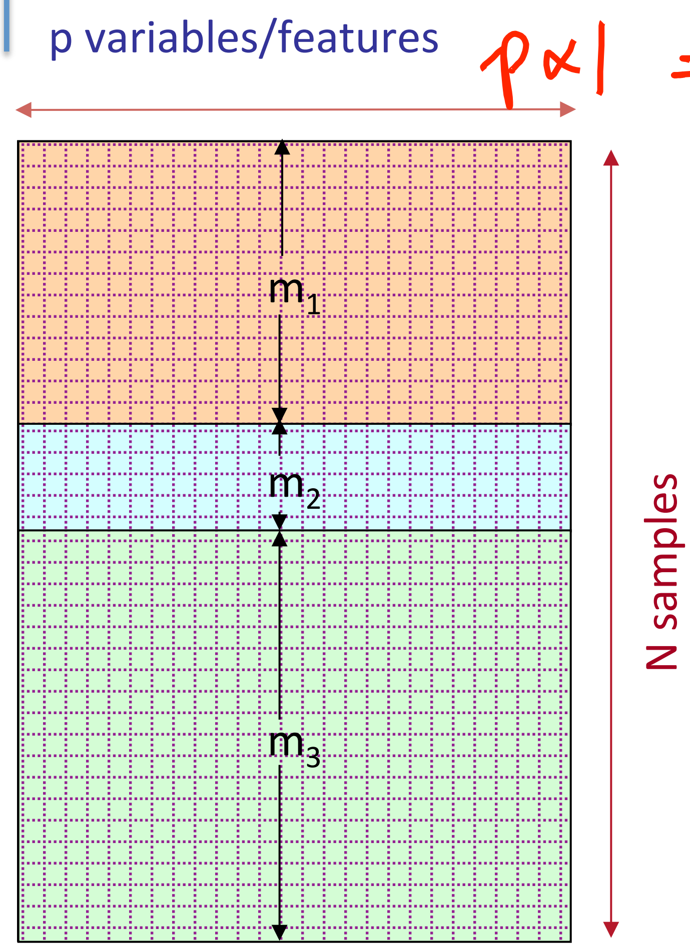
(2) Wrapper : **Feature Subset**

- 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 the space of all feature subsets ?

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

$p' \times 1 \Rightarrow \beta^* / \theta_i^* = \arg\min J(\beta / \theta_i)$ on training

$p \times 1 \Rightarrow \theta^* = \arg\min MSE(\theta_i)$ on validation



Split data into 3 sets:

training, validation, and test set.

2^P

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

$[0, 1, 0, 1, \ldots]$ p

θ_1

2) Select the feature subset, which performs best on validation data.

θ_2

- ~~Repeat and average if you want to reduce variance (cross-validation).~~

θ_m

3) Test on test data.

$[m \leq 2^P]$

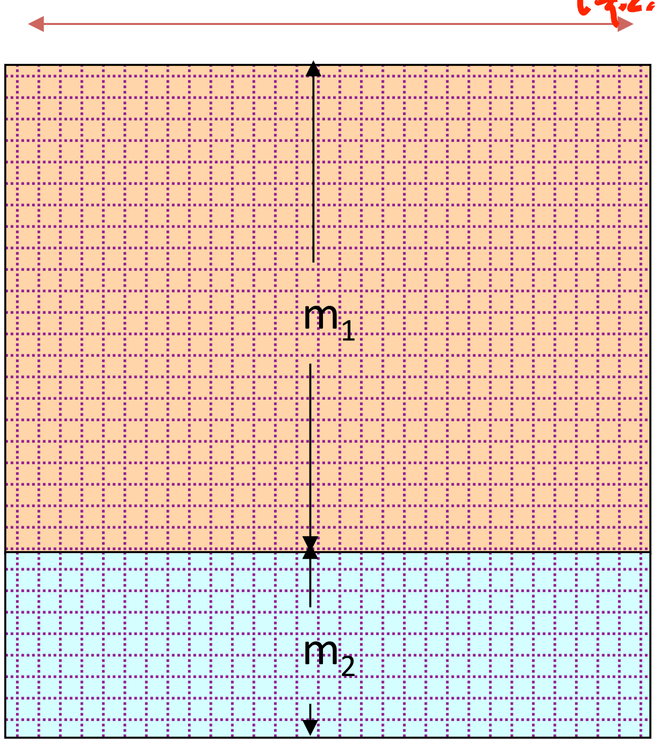
Danger of over-fitting with intensive search!
From Dr. **Isabelle Guyon**

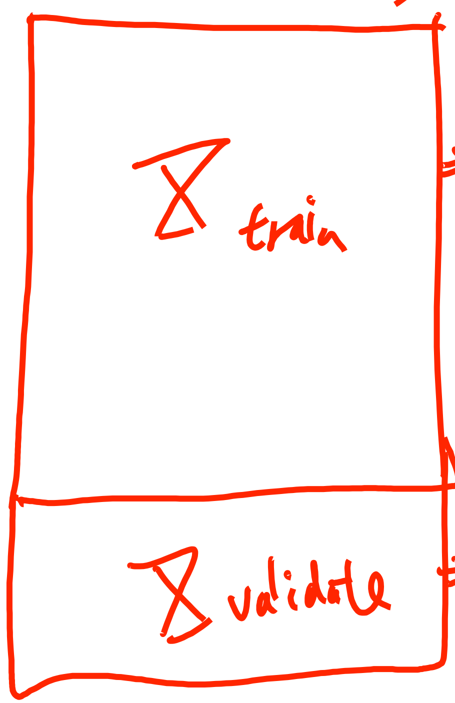
(a). Assessment: How to access a particular feature subset

$\Theta_i = [0, 1, \ldots, 1, 0]$

p variables/features

those with $[\Theta_i]_t \neq 0$, $t \in \{1, 2, \ldots, p\}$ → p'





① $\beta^*/\Theta_i = \underset{\beta/\Theta_i}{\arg\min}\, J(\beta/\Theta_i)$

e.g. linear regression training phase

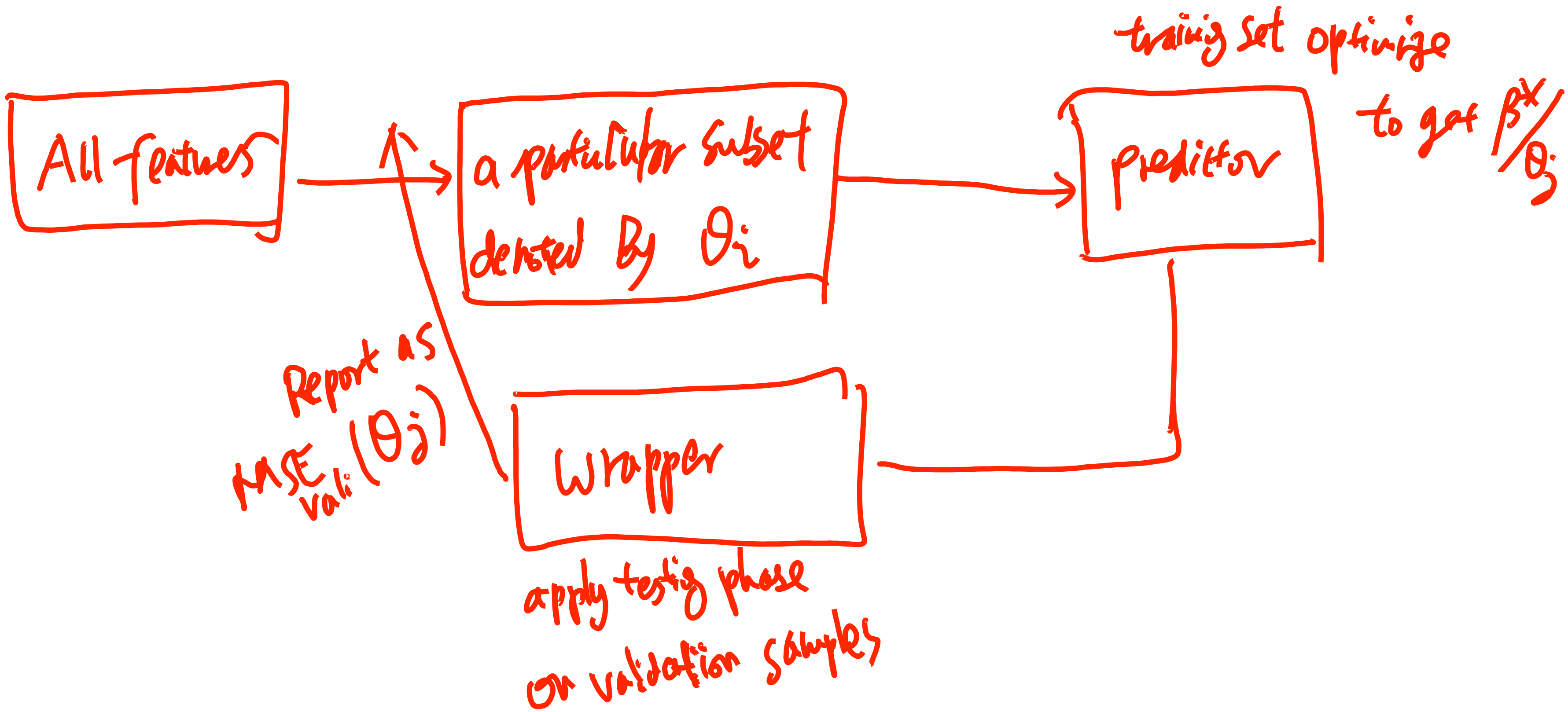
② assessment by using β^*/Θ_i on vali. samples

e.g. MSE validate samples

⇓

In short, $MSE_{vali}(\Theta_i)$

(a). Assessment: How to access a particular feature subset



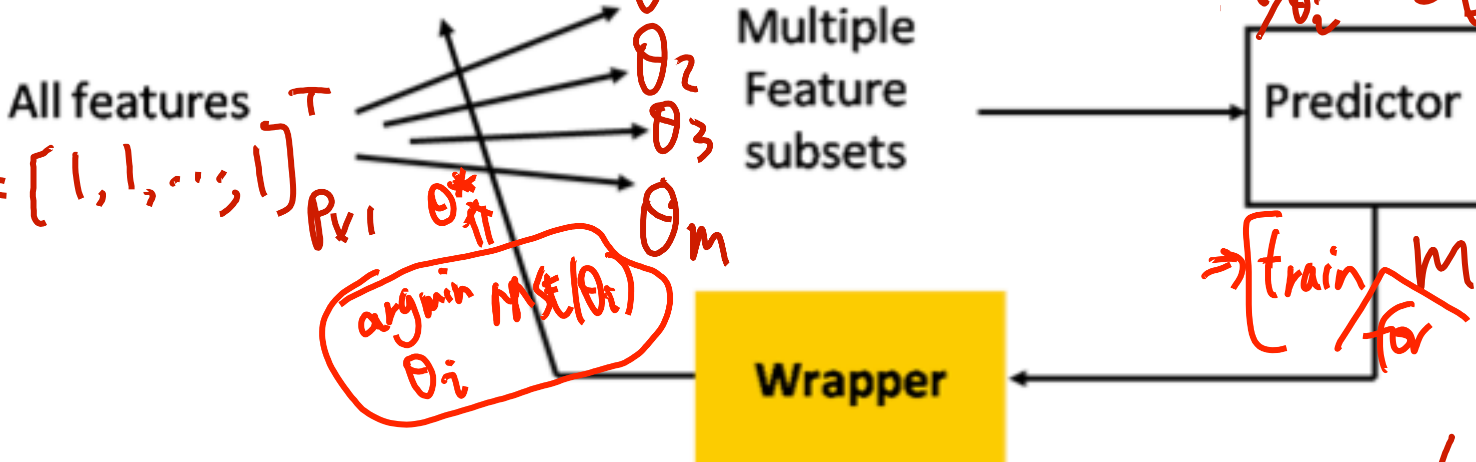
(a). Assessment: How to access multiple candidates of feature subsets

$m \le 2^p$

Wrapper Methods

$\theta_1 = [0, 1, \dots, 0]^T$

$\beta^*_{/\theta_i} = \underset{\beta/\theta_i}{\arg\min} (J(\beta | \theta_i))$



$\theta = [1, 1, \dots, 1]^T_{p \times 1}$

θ_2, θ_3, θ_m

$\theta^* \Uparrow \underset{\theta_i}{\arg\min}\, MSE(\theta_i)$

$i = 1, \dots, m$

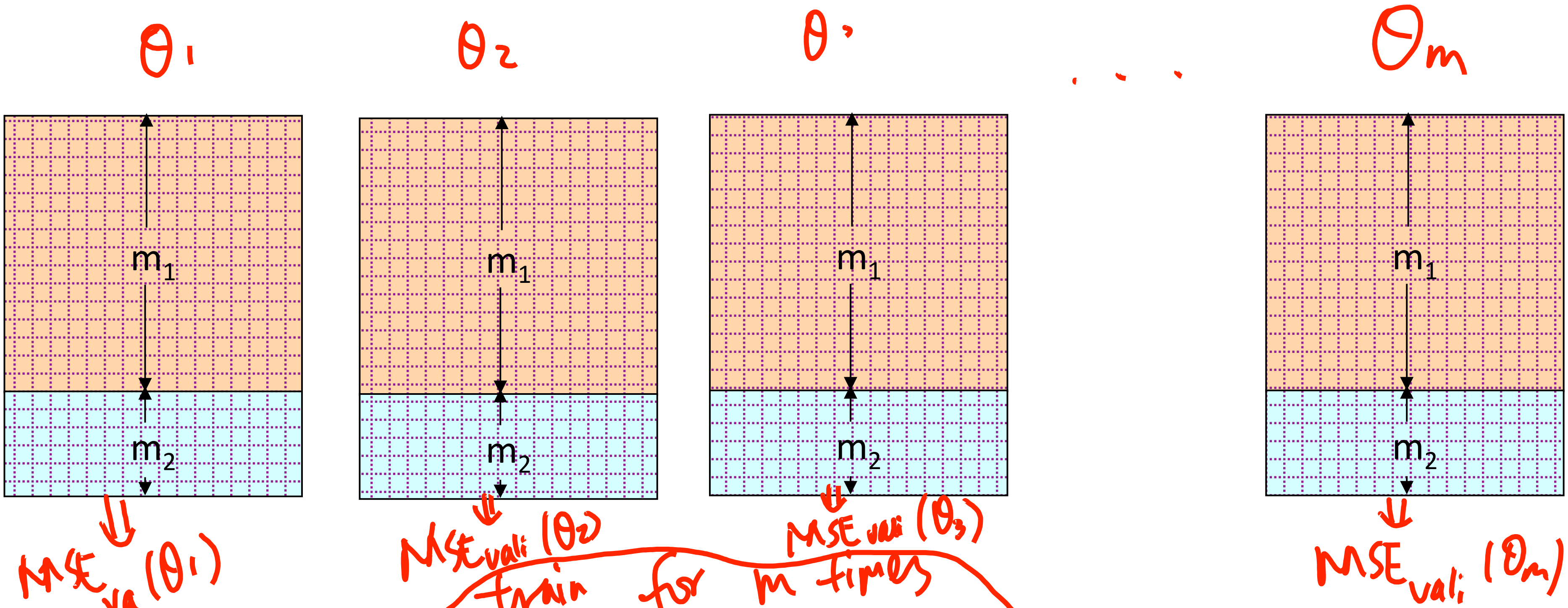
[train m times] for

Scoring func $= MSE(\beta^*_{/\theta_i}, \text{test})$

[test on vali set for m times]

in short $MSE_{vali}(\theta_i)$ $i = 1, 2, \dots, m$

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



train for m times

test for m times

Wrapper feature selection / three set of labeled samples

① $\theta^*_{p\times 1} = [0, 1, 0, 0, \ldots, 1]^T \Rightarrow$ Validation

to get best θ^*

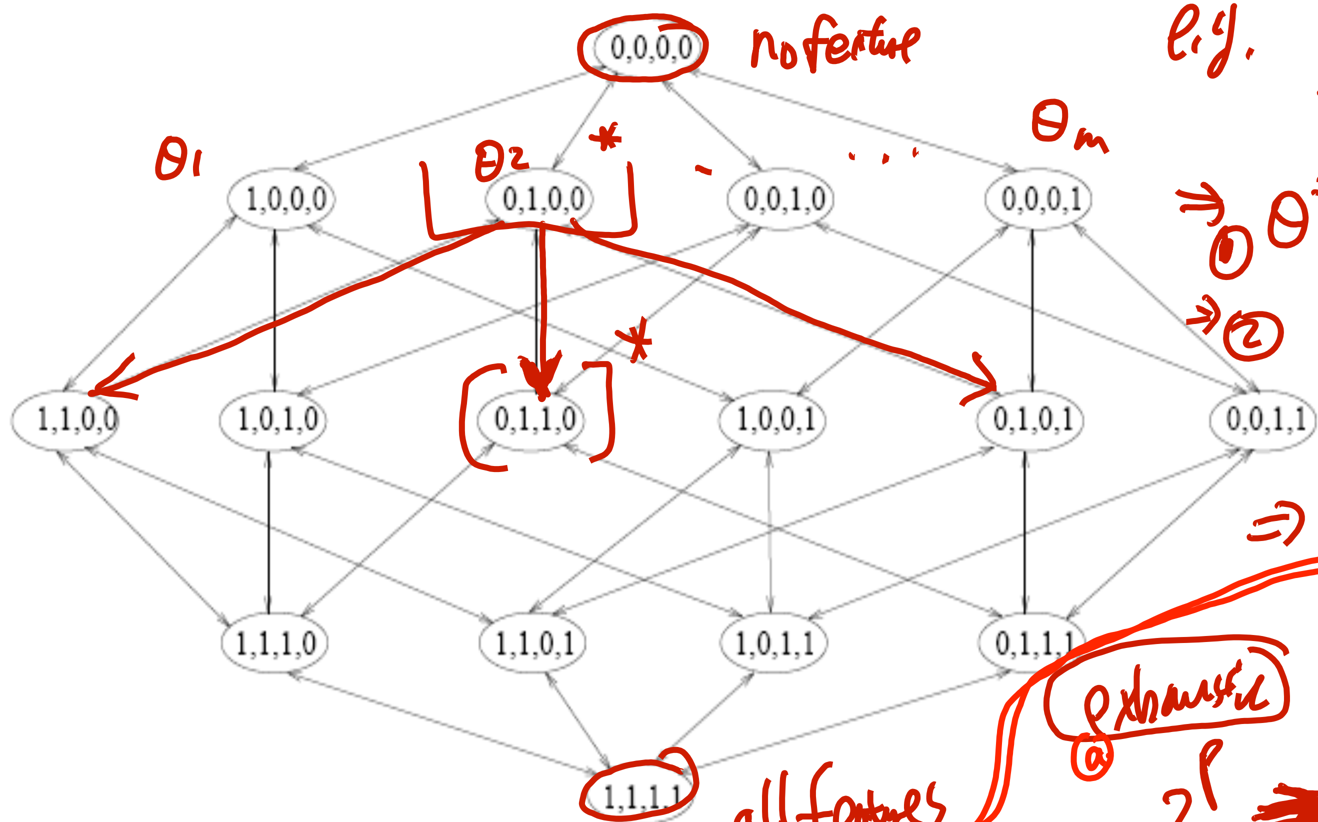
② $\beta^*_{p'\times 1}\big|_{\theta_i} = \underset{\beta}{\operatorname{argmin}}\, J(\beta \mid \theta_i^*) \Rightarrow$ training

for each θ_i, get best β^* / θ_i

③ $\beta^*_{p'\times 1} \mid \theta^*_{p\times 1} \Rightarrow$ testing

obtain / check the generalization performance of Best feature subset / Best β.

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



e.g. greedy forward selection

⇒ ① θ_1^* → for level 1

⇒ ② θ_2^* ⇒ for level 2

⇒ ⋮

ⓐ exhaustive 2^p ⇒ ≪ ⓑ greedy ⇓ $p+(p-1)+\ldots$ ⇓ $\frac{p(p+1)}{2}$ ⇓ 10

$2^4=16$ ⇒

m ⇒ ???

(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: even more search strategies for selecting feature subset

0

$\{\} \to \{1\} \to \{2\} \to \dots$ $\{p\} \to \{p-1\} \to \{p-2\}$

- **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**

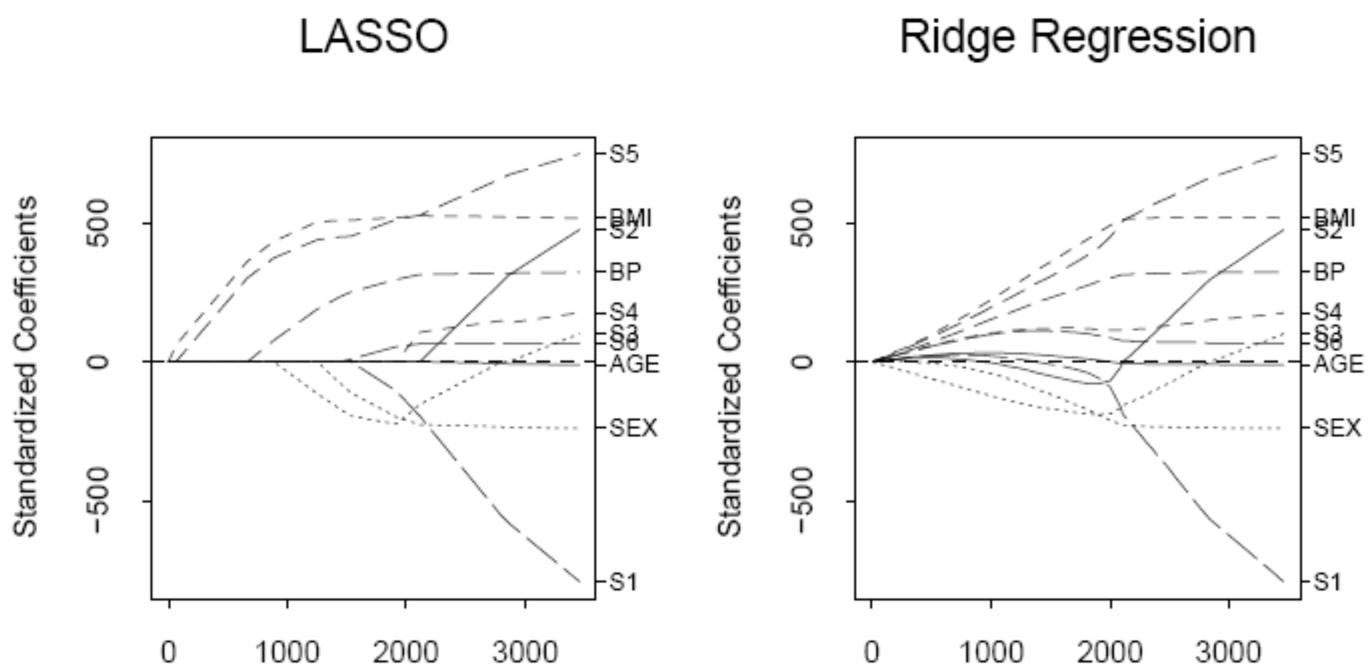
(3) Embedded

– Embedding approach:

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

(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)



sum(|beta|)   sum(|beta|)

From ESL book

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

Embedded Methods

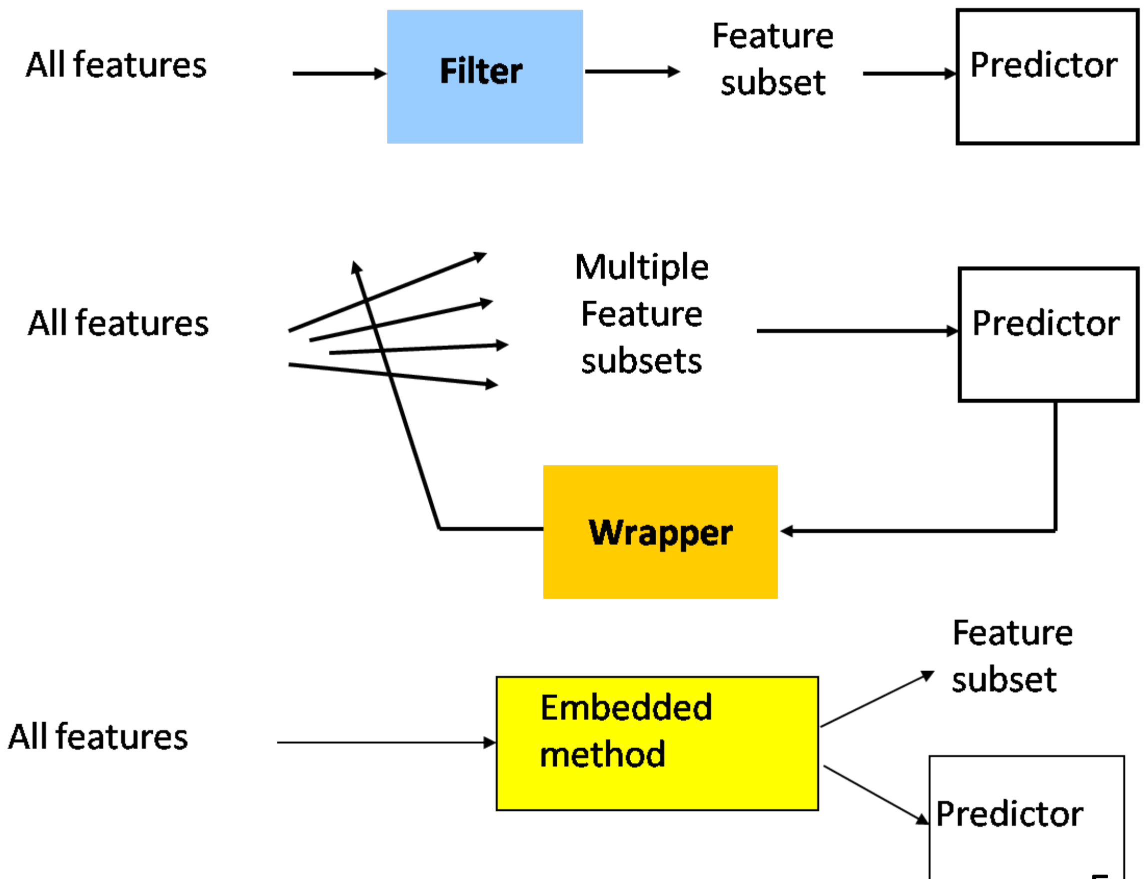
- Specific to a given learning machine!

- Performs variable selection (implicitly) in the process of training

- Just train a (single) model

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**

Vs. Dimensionality Reduction (Later)

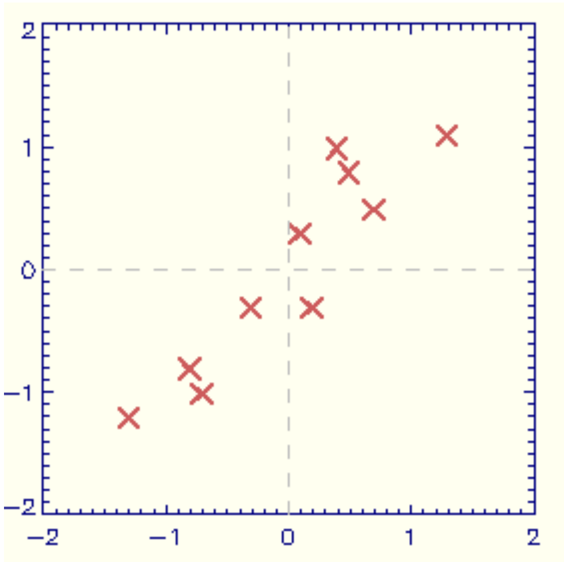
In the presence of many of features, select the most relevant subset of (weighted) combinations of features.

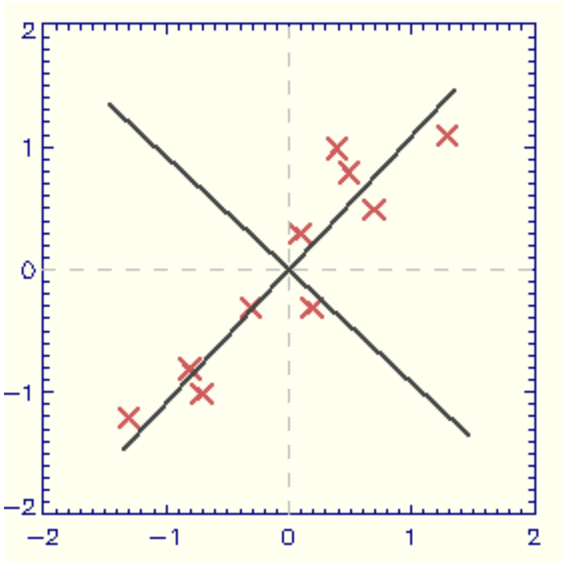
Feature Selection: $X_1,\ldots,X_p \rightarrow X_{k1},\ldots,X_{kp'}$

Dimensionality Reduction: $X_1,\ldots,X_m \rightarrow f_1(X_1,\ldots,X_m),\ldots,f_p(X_1,\ldots,X_m)$

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.

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**