EECS 545: Machine Learning

Lecture 11. Feature selection

Honglak Lee

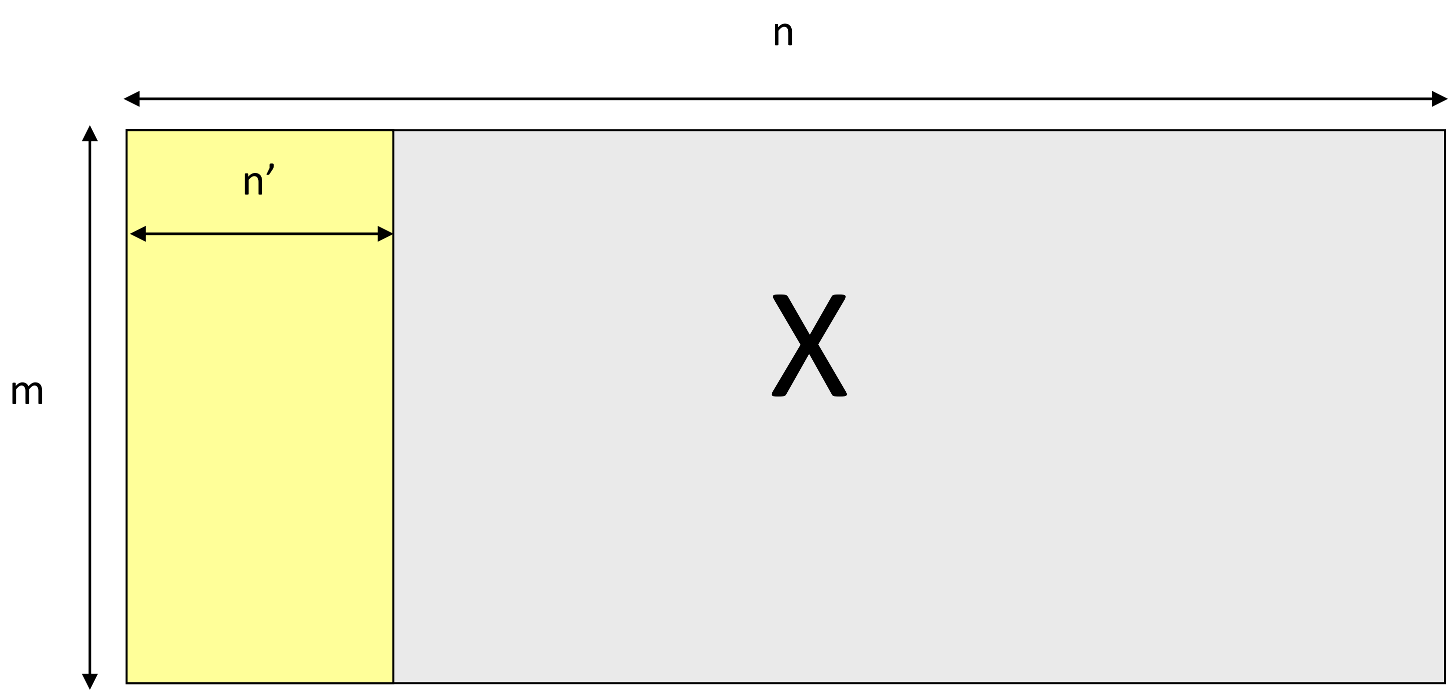
2/14/2011

Outline

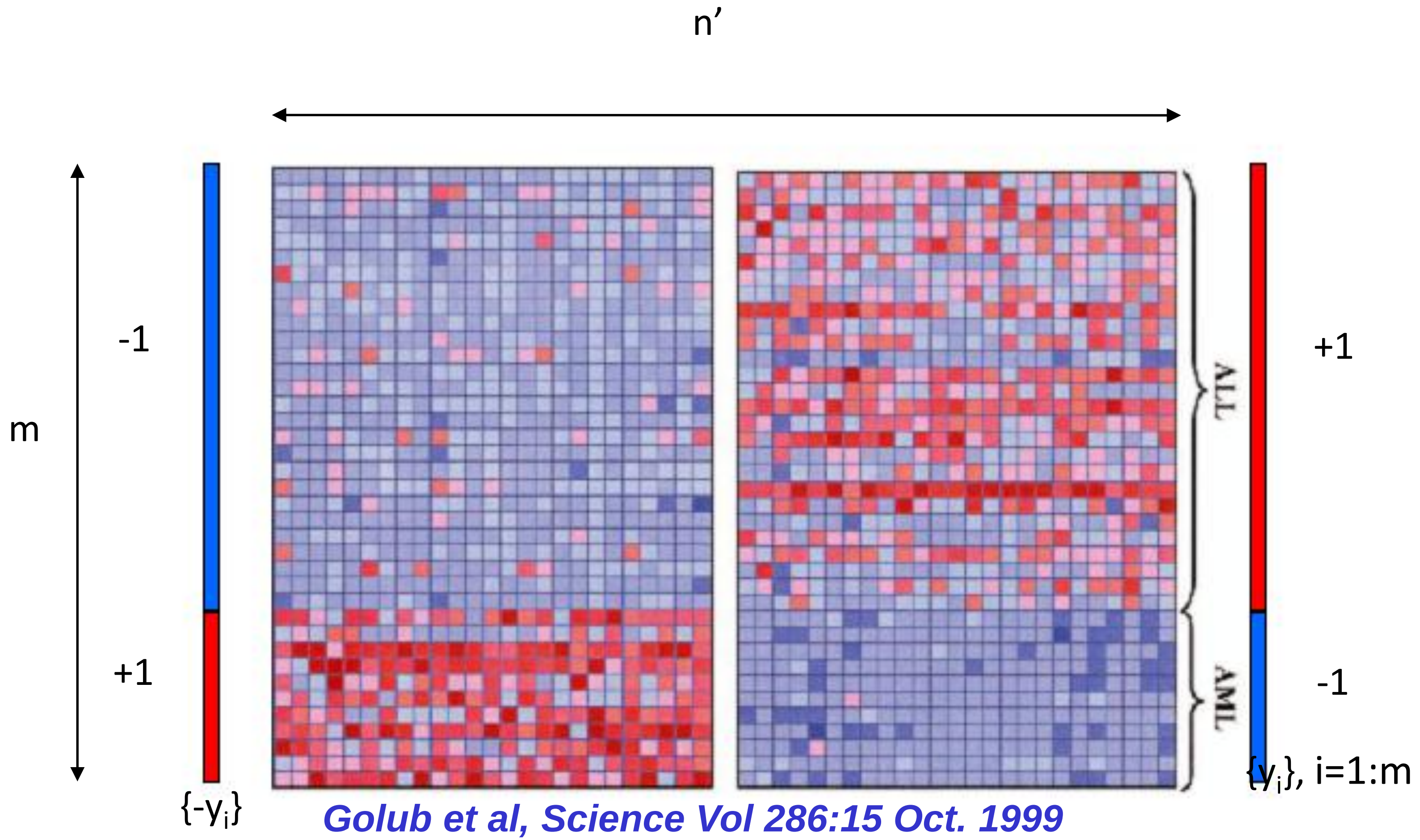
- Overview of feature selection
- Univariate method
- Filtering method
- Wrapper method
  - Forward feature selection
  - Backward feature selection
- Embedded method
  - L1 regularization

Feature Selection

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

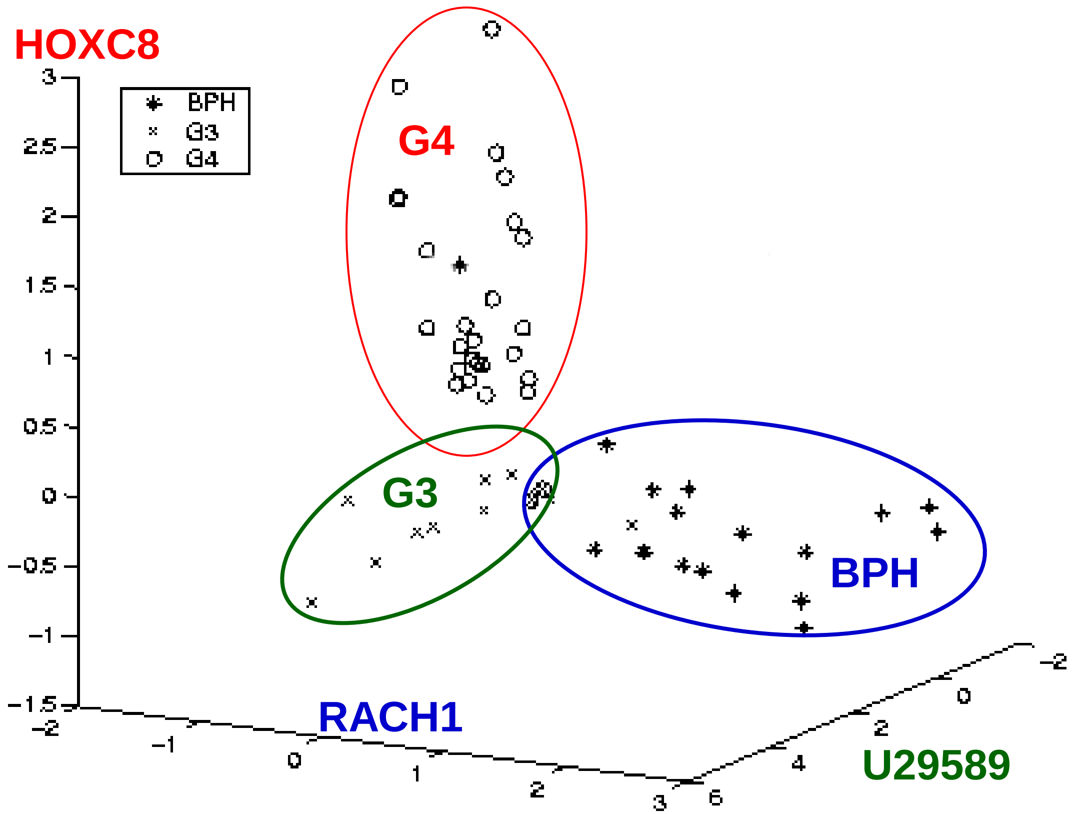


Leukemia Diagnosis



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

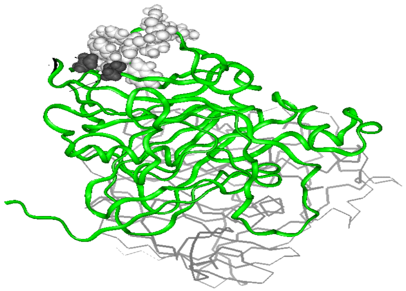
Prostate Cancer Genes



RFE SVM, ***Guyon-Weston, 2000. US patent 7,117,188***

Application to prostate cancer. ***Elisseeff-Weston, 2001***

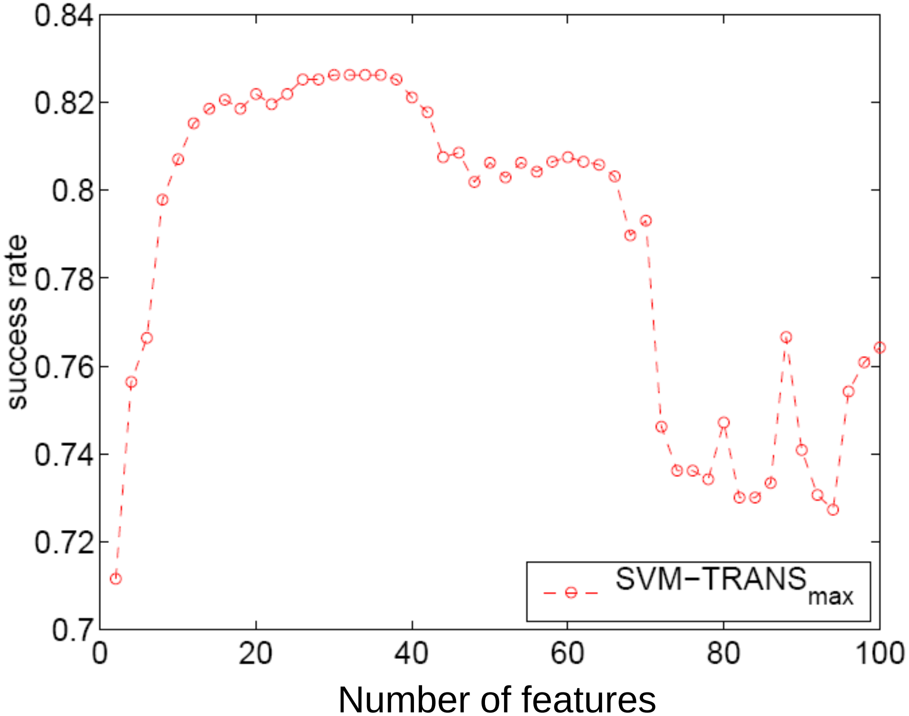
QSAR: Drug Screening



Binding to Thrombin (DuPont Pharmaceuticals)

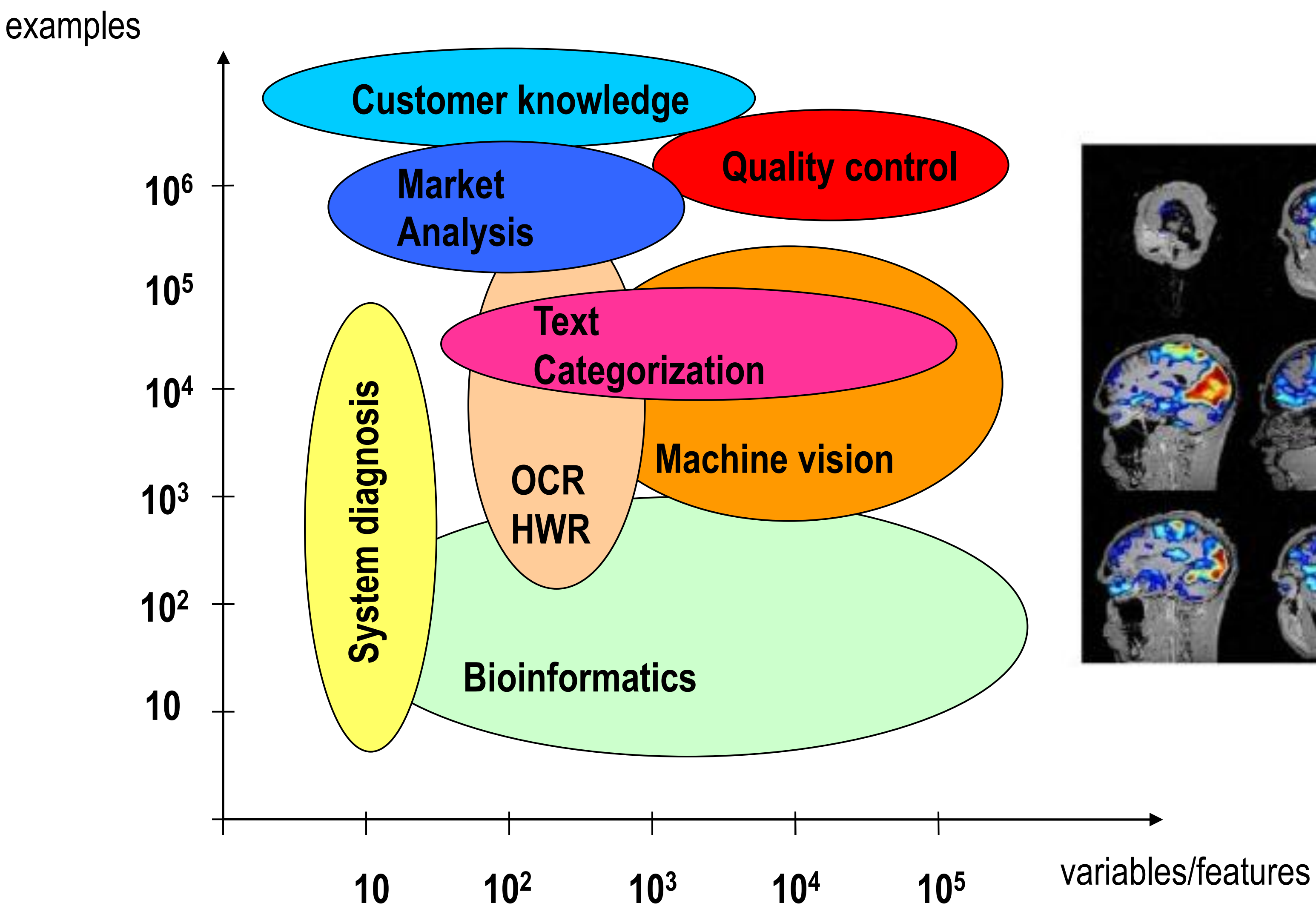
- 2543 compounds tested for their ability to bind to a target site on thrombin, a key receptor in blood clotting; 192 “active” (bind well); the rest “inactive”. Training set (1909 compounds) more depleted in active compounds.

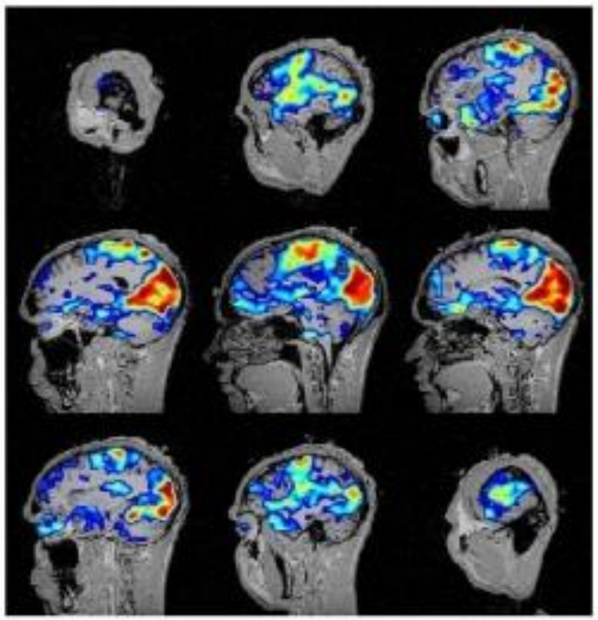
- 139,351 binary features, which describe three-dimensional properties of the molecule.



Weston et al, Bioinformatics, 2002

Applications





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 (classifier).
- **Wrapper method:** uses a classifier to assess features or feature subsets.

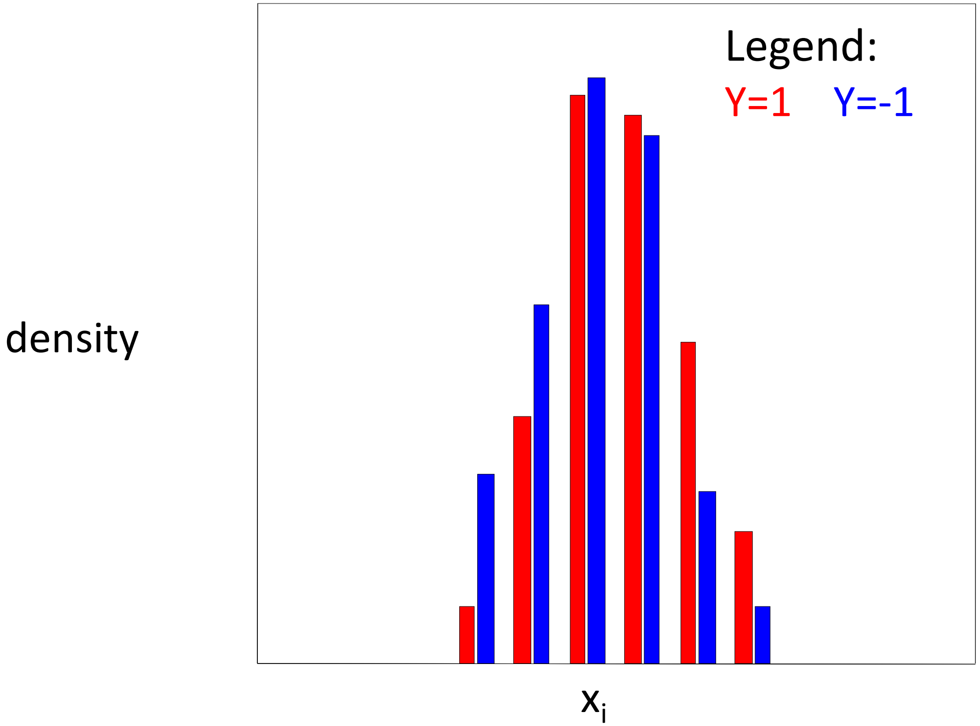
Univariate Filter Methods

Individual Feature Irrelevance

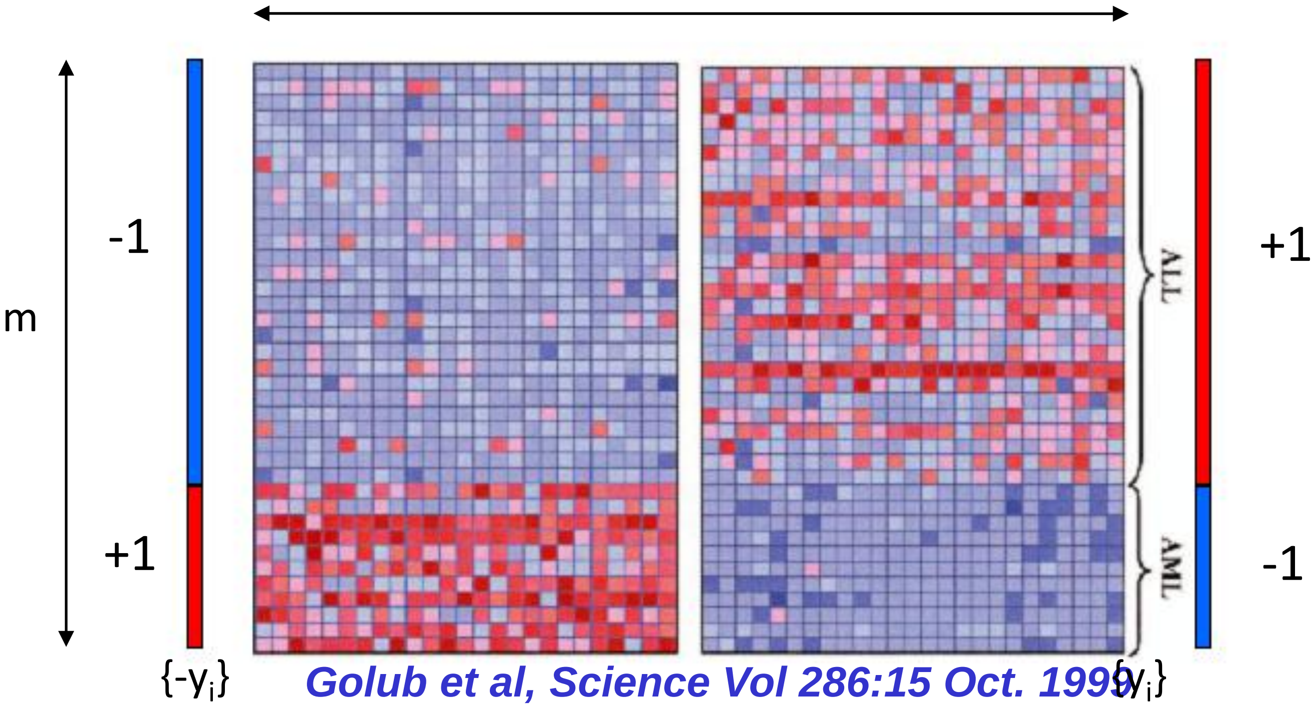
$$P(X_i, Y) = P(X_i)\,P(Y)$$

$$P(X_i \mid Y) = P(X_i)$$

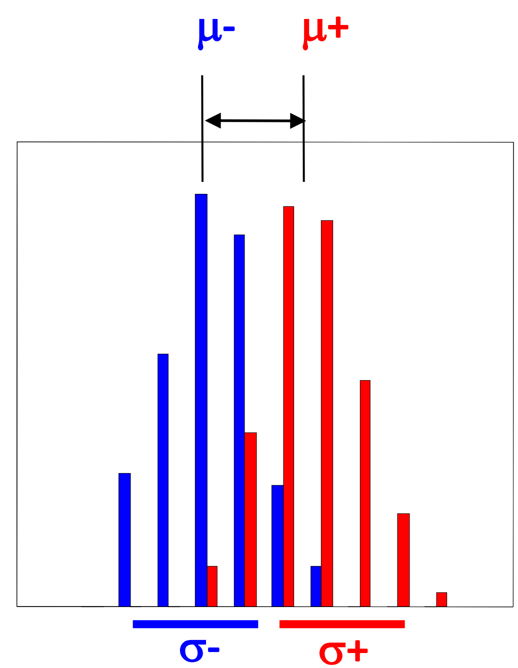
$$\color{red}{P(X_i \mid Y=1)} = \color{blue}{P(X_i \mid Y=-1)}$$



S2N



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



$$\text{S2N} = \frac{|\mu+ - \mu-|}{\sigma+ + \sigma-}$$

$$\text{S2N} \cong R \sim \mathbf{x} \bullet \mathbf{y}$$

after “standardization” $\mathbf{x} \leftarrow (\mathbf{x}-\mu_x)/\sigma_x$

Univariate Dependence

- Independence:

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

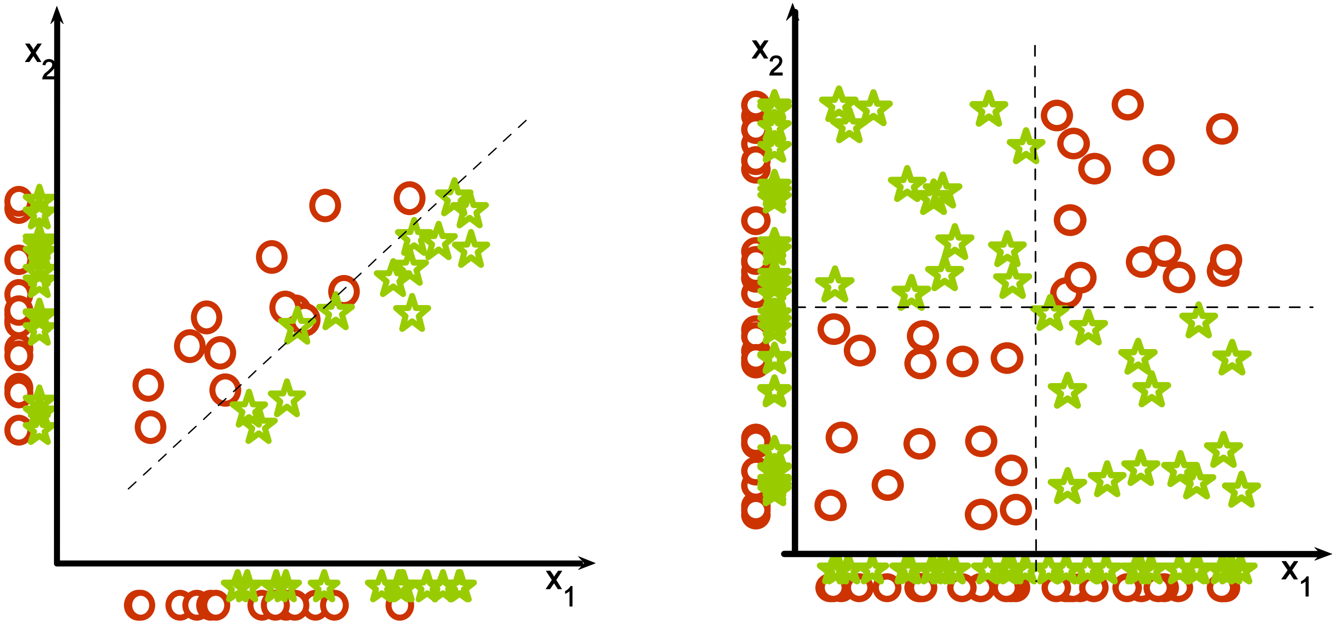
- Measure of dependence:

$$MI(X, Y) = \int P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} dX\,dY$$

$$= KL\big(\,P(X,Y)\;||\;P(X)P(Y)\,\big)$$

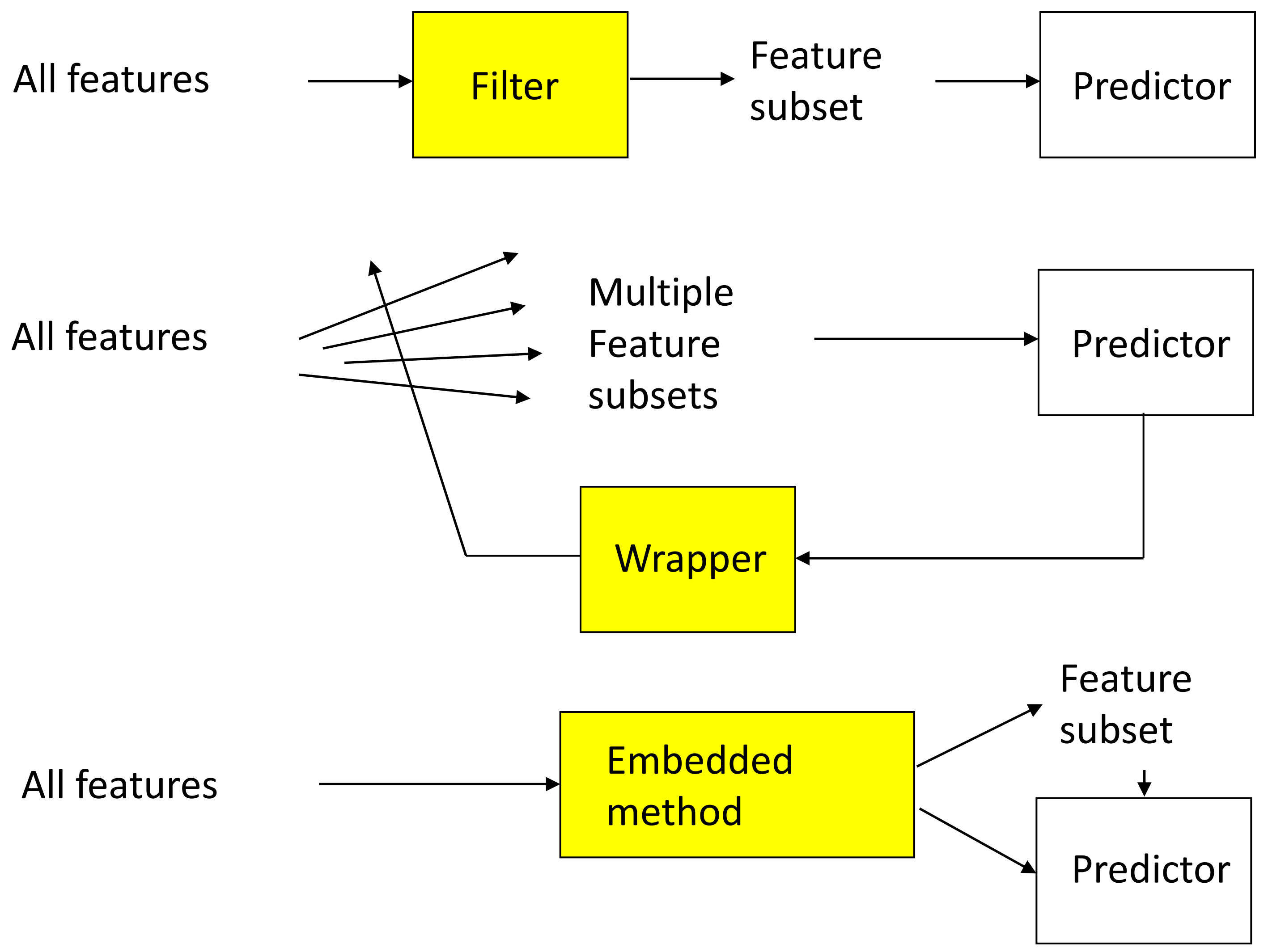
Multivariate Methods

Univariate selection may fail



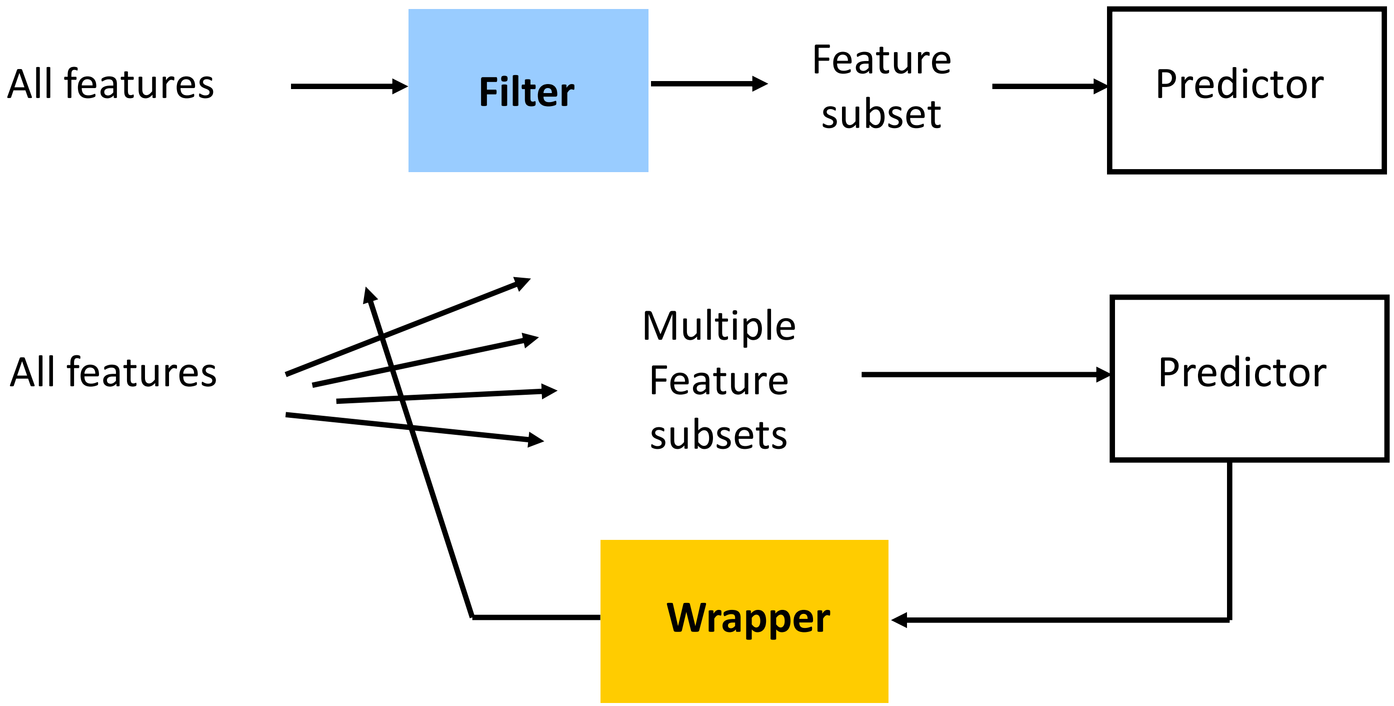
Guyon-Elisseeff, JMLR 2004; Springer 2006

Filters,Wrappers, and Embedded methods



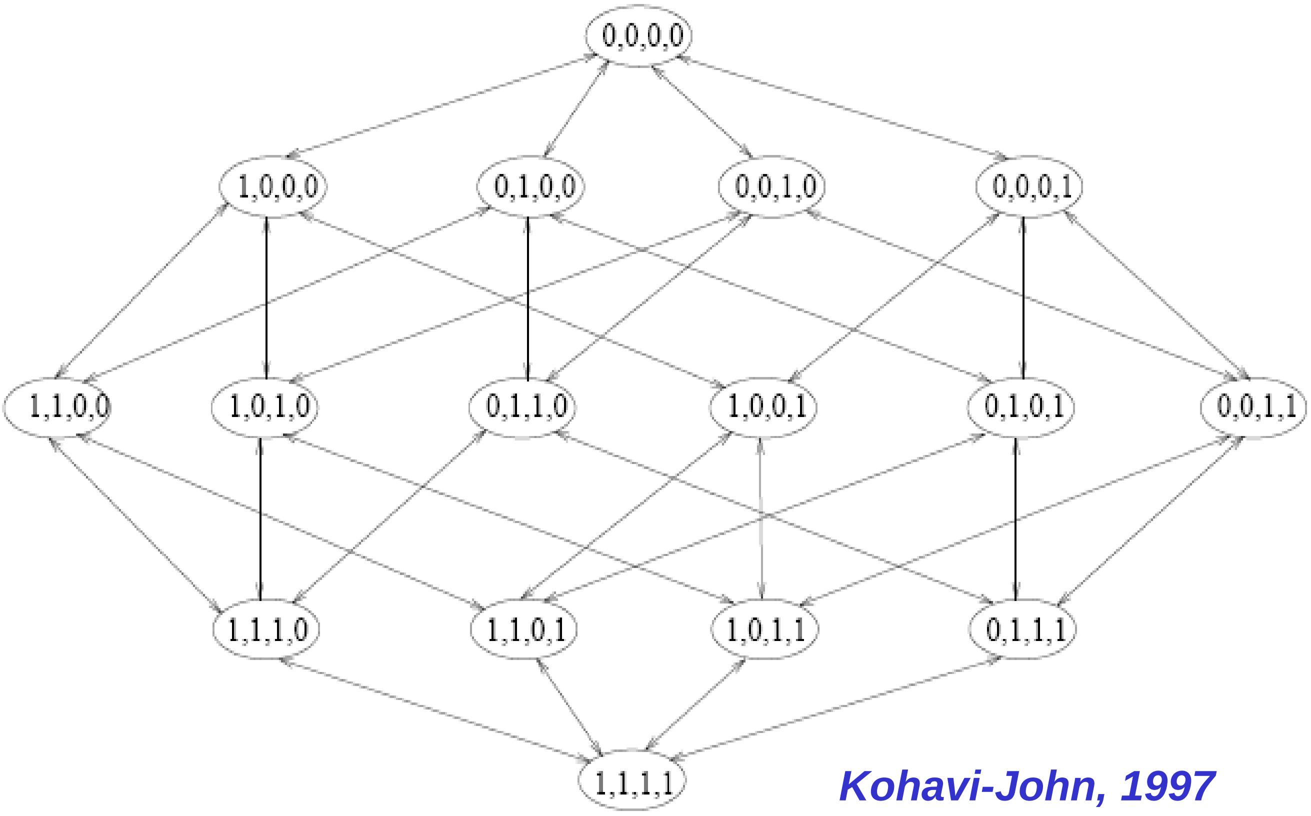
Filters vs. Wrappers

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



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

Wrappers for feature selection



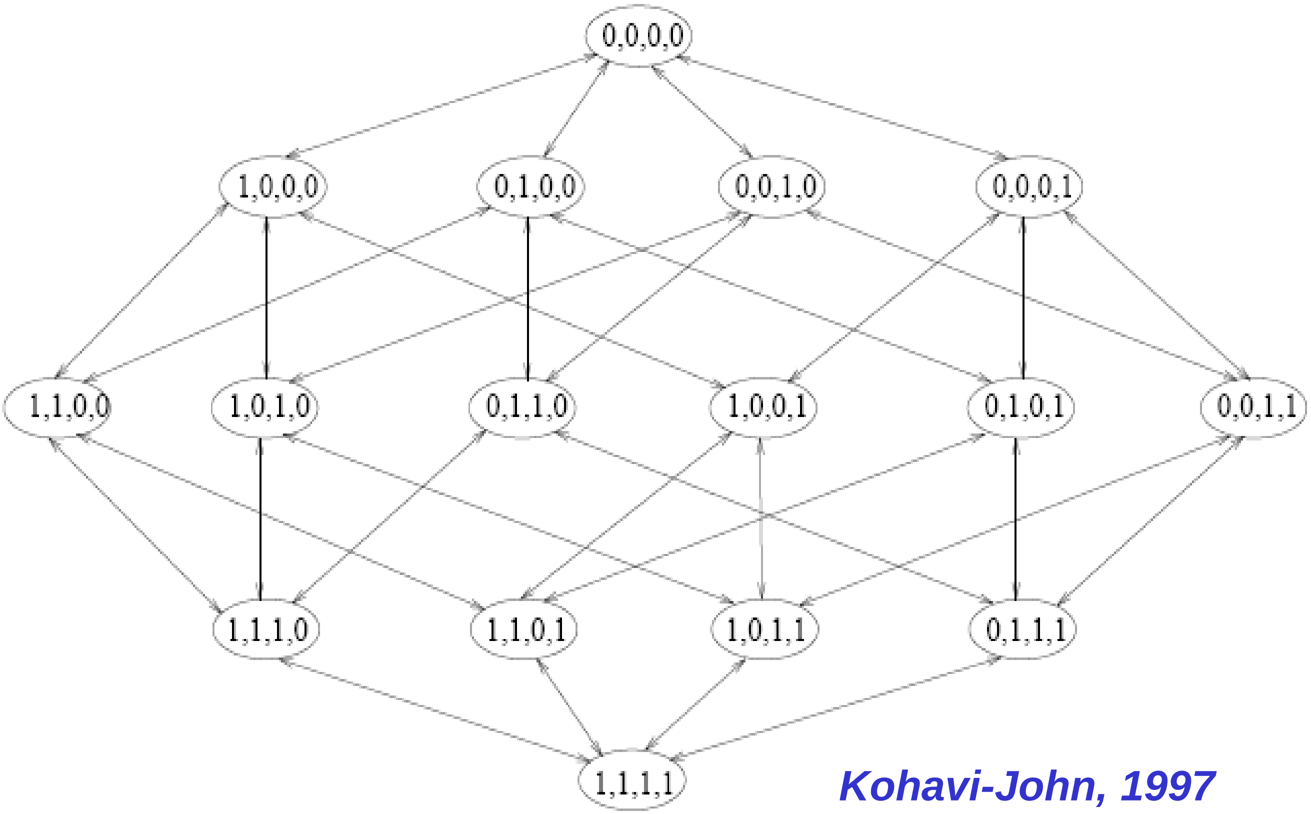
Kohavi-John, 1997

N features, 2^N possible feature subsets!

Search Strategies

- **Exhaustive search**
- **Greedy search:**
  - forward selection
  - backward elimination
- **Beam search:** keep k best path at each step.

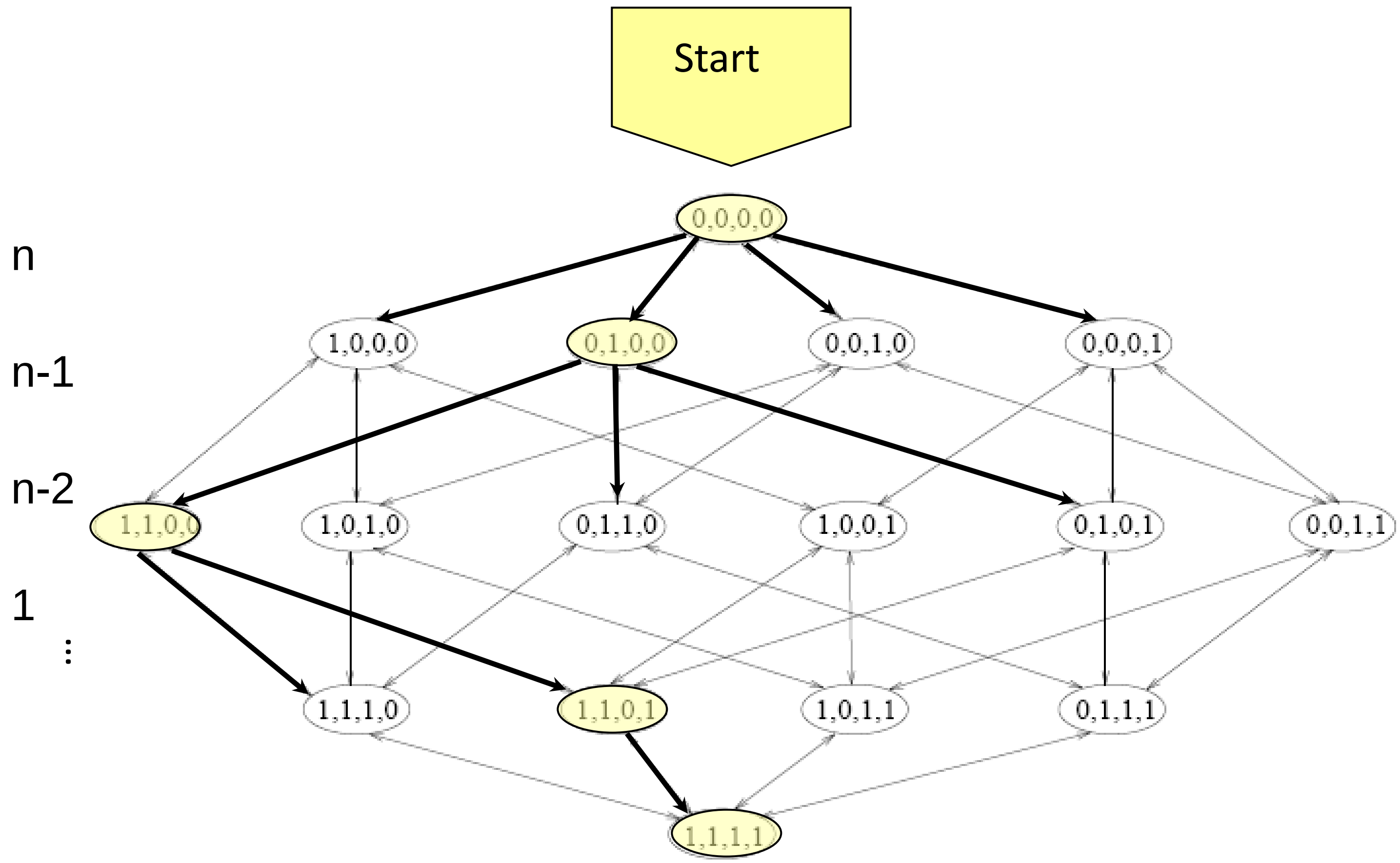
Multivariate FS is complex



Kohavi-John, 1997

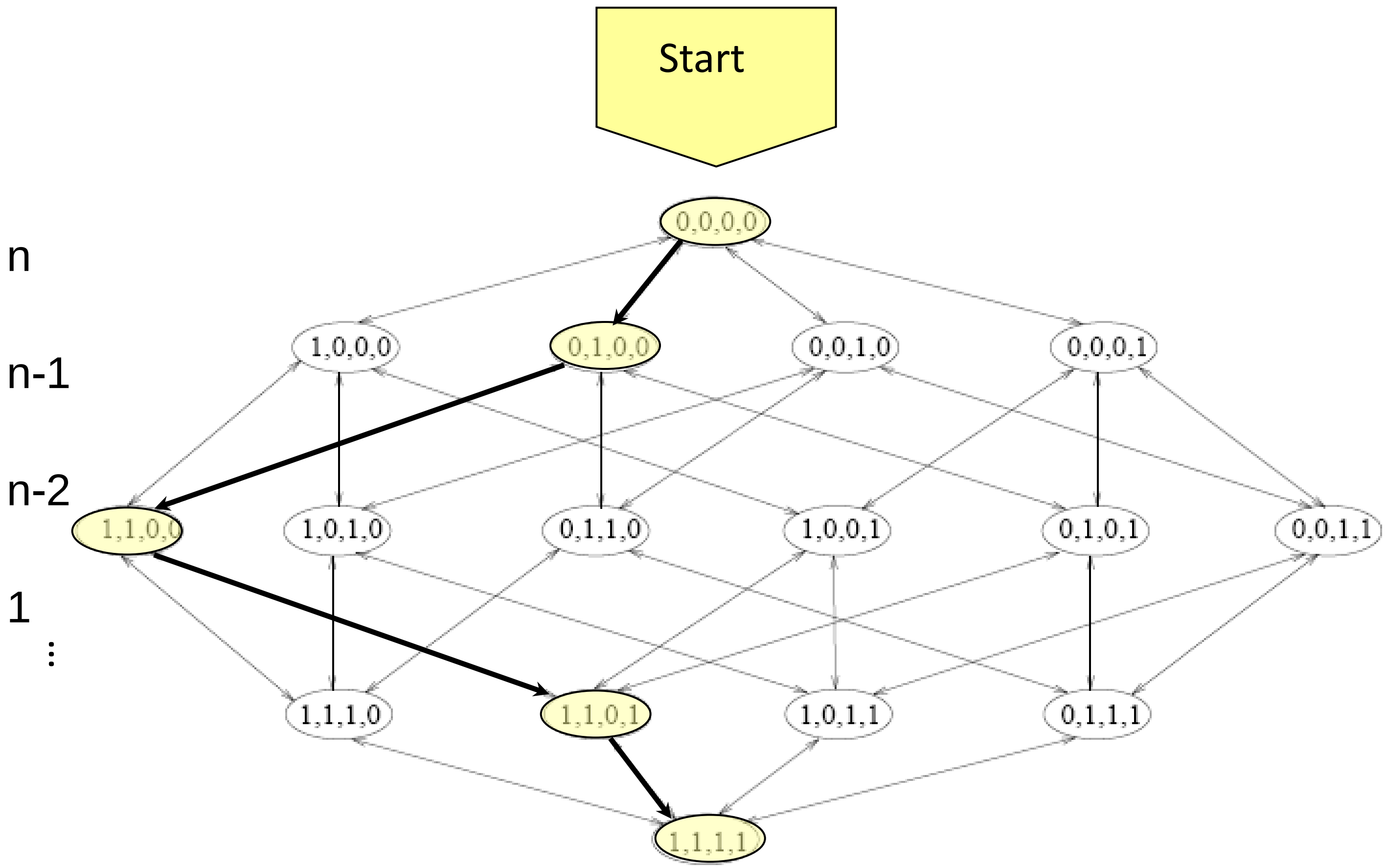
N features, 2^N possible feature subsets!

Forward Selection (wrapper)



Also referred to as SFS: Sequential Forward Selection

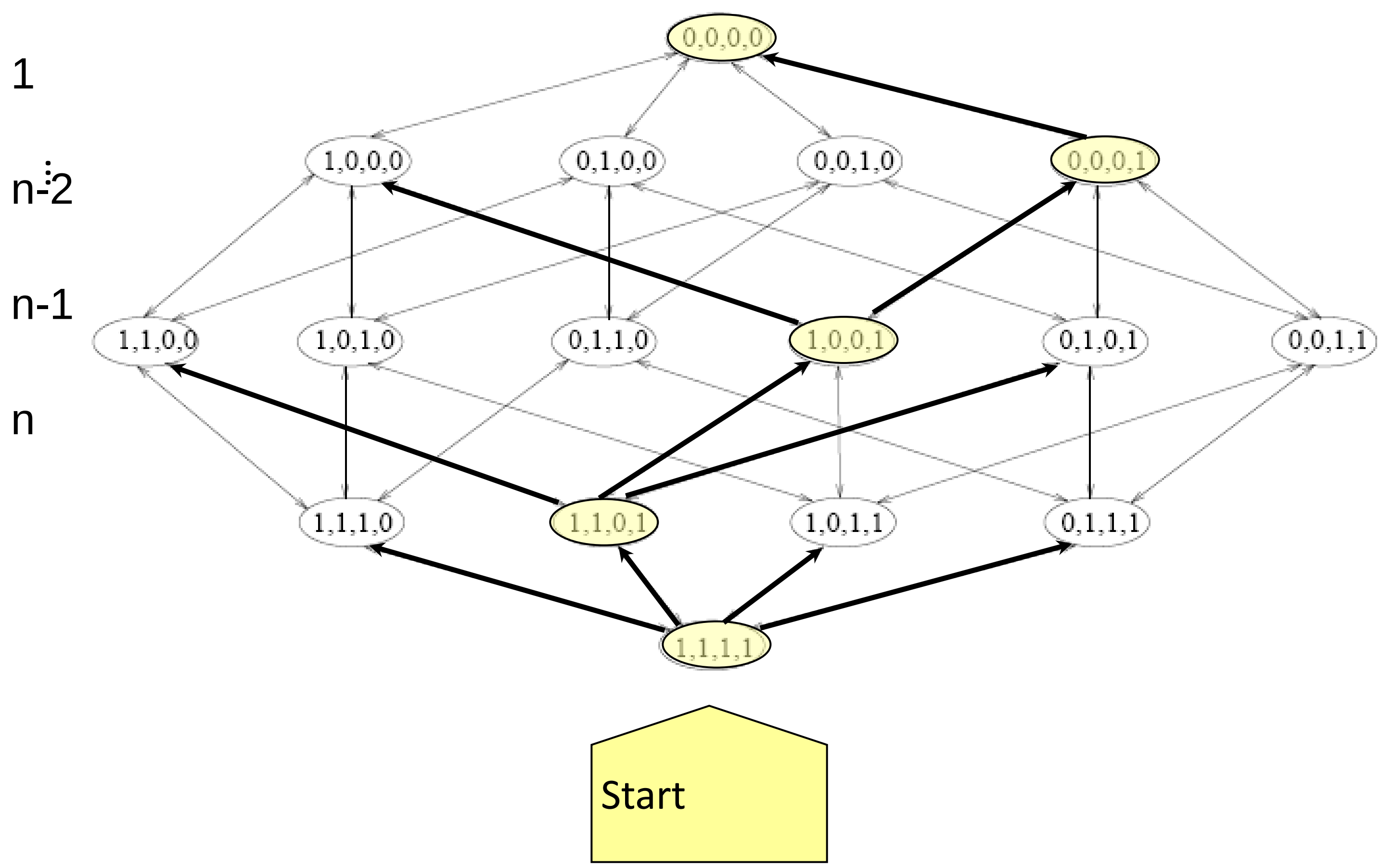
Forward Selection (embedded)



Guided search: we do not consider alternative paths.
Typical ex.: Gram-Schmidt orthog. and tree classifiers.

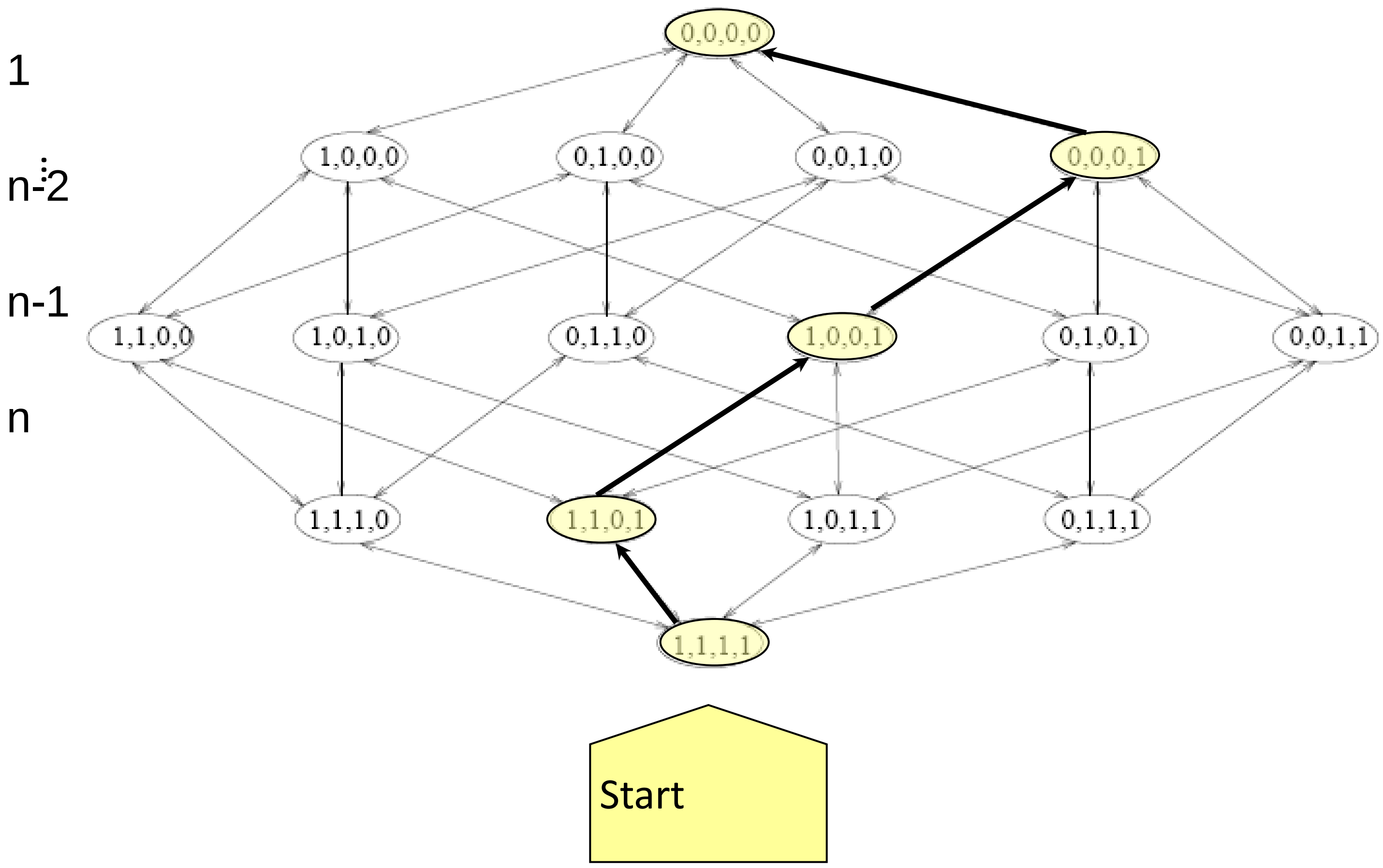
Backward Elimination (wrapper)

Also referred to as SBS: Sequential Backward Selection



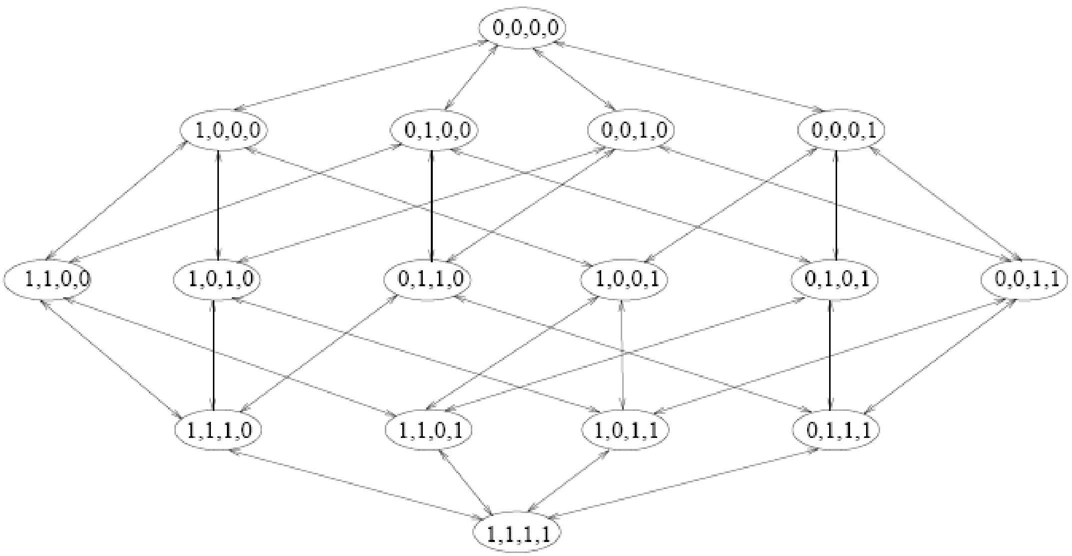
Backward Elimination (embedded)

Guided search: we do not consider alternative paths.
Typical ex.: “recursive feature elimination” RFE-SVM.



Scaling Factors

Idea: Transform a discrete space into a continuous space.



$\boldsymbol{\sigma}=[\sigma_1, \sigma_2, \sigma_3, \sigma_4]$

- Discrete indicators of feature presence: $\sigma_i \in \{0, 1\}$
- Continuous scaling factors: $\sigma_i \in \mathbb{R}$

Now we can do gradient descent!

Formalism

- Many learning algorithms are cast into a minimization of some regularized functional:

$\|\alpha\|_1$

$$\underbrace{\min_{\alpha} \hat{R}(\alpha, \sigma)}_{G(\sigma)} = \min_{\alpha} \sum_{k=1}^{m} L(f(\alpha, \sigma \circ x_k), y_k) + \Omega(\alpha)$$

Empirical error

Regularization capacity control

Justification of RFE and many other embedded methods.

Next few slides: André Elisseeff

Embedded method

- Embedded methods are a good inspiration to design new feature selection techniques for your own algorithms:
  - Find a functional that represents your prior knowledge about what a good model is.
  - Add the σ weights into the functional and make sure it's either differentiable or you can perform a sensitivity analysis efficiently
  - Optimize alternatively according to α and σ
  - Use early stopping (validation set) or your own stopping criterion to stop and select the subset of features

- Embedded methods are therefore not too far from wrapper techniques and can be extended to multiclass, regression, etc…

The l_1 SVM

$\|w\|_1$

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i$$

$$\text{s.t. } \forall i,\ y^{(i)}(w^T x^{(i)} + b) \geq 1-\xi_i$$

$$\forall i,\ \xi_i \geq 0$$

- A version of SVM where $\Omega(w)=||w||^2$ is replaced by the l_1 norm $\Omega(w)=\sum_i |w_i|$
- Can be considered an embedded feature selection method:
  - Some weights will be drawn to zero (tend to remove redundant features)
  - Difference from the regular SVM where redundant features are included

Bi et al 2003, Zhu et al, 2003

Other examples: L1 regularized algorithms

- Generally, just add L1 regularization to objective function

$\|W\|_2^2$ L2 regularized

- L1 Logistic regression

$$L = \sum_i \log P(y^{(i)}|\boldsymbol{x}^{(i)}, \boldsymbol{w}) + \|\boldsymbol{w}\|_1$$

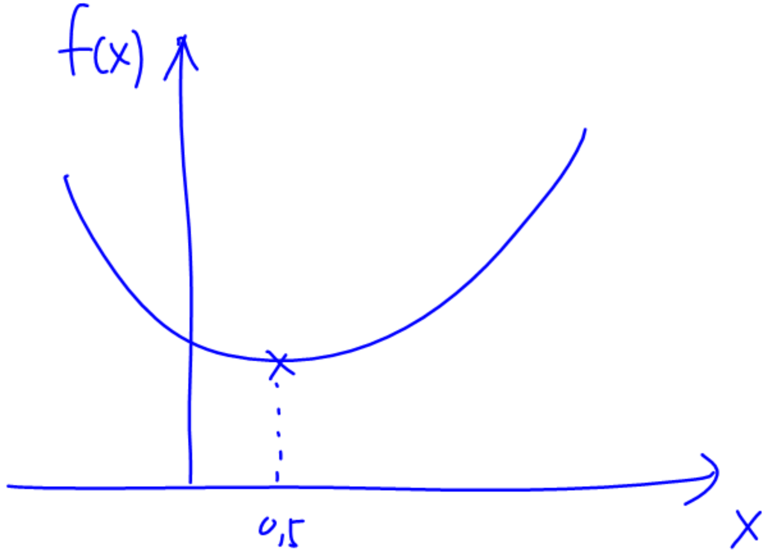
$\|W\|_2^2$ L2 reg.

- L1 Lest squares

$$L = \sum_i \left\|y^{(i)} - \boldsymbol{w}^T \boldsymbol{x}^{(i)}\right\|^2 + \|\boldsymbol{w}\|_1$$

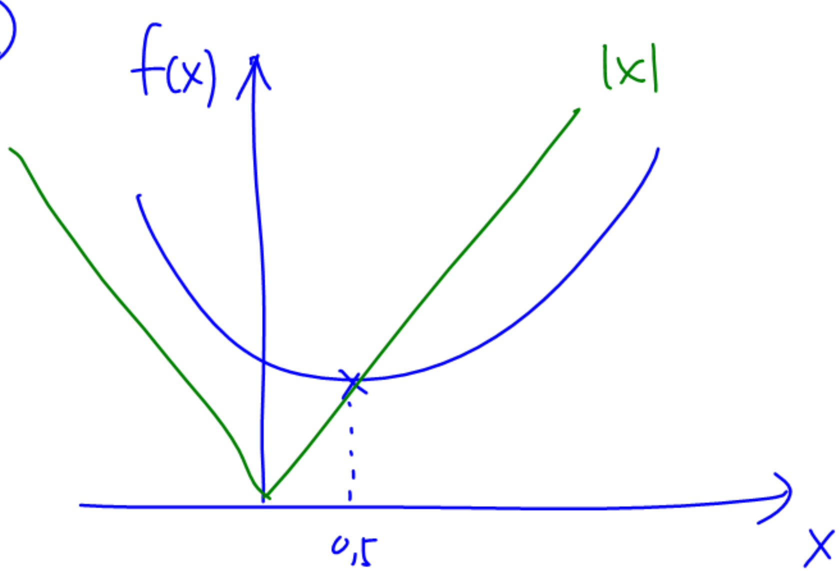
* Both problems are convex, but need a specialized solver to deal with L1 norm.

1

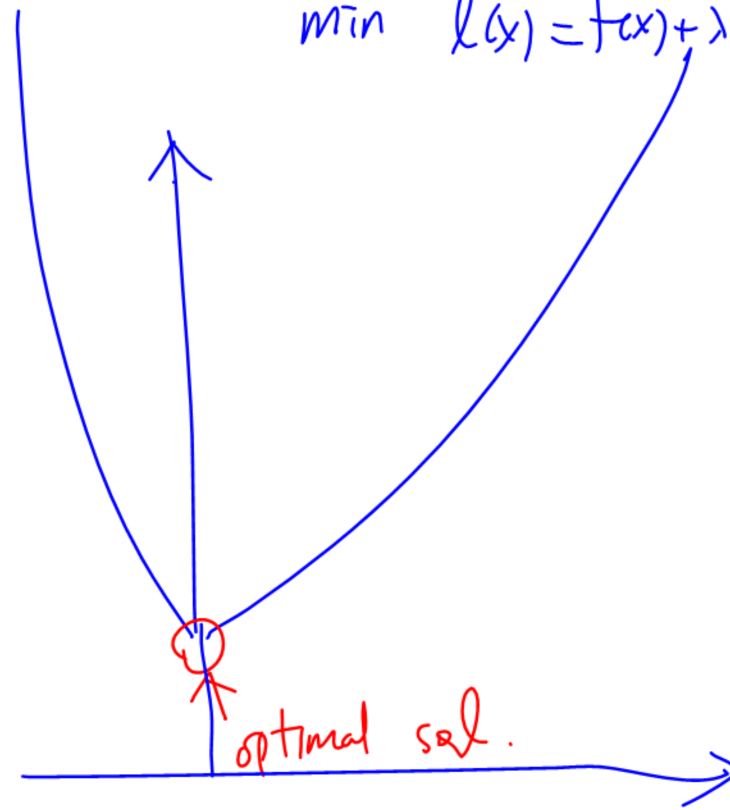


min f(x)

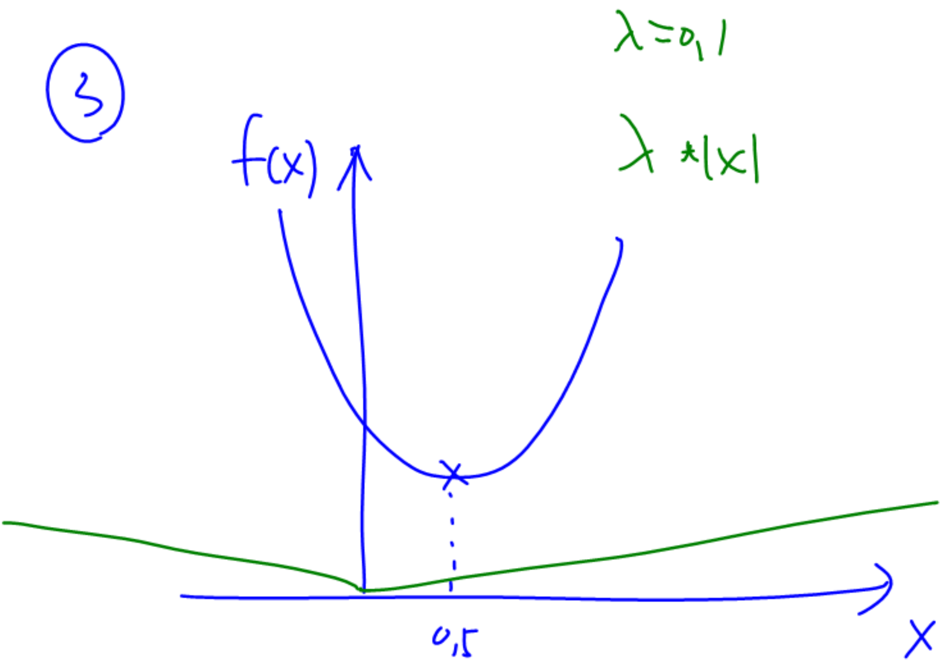
2

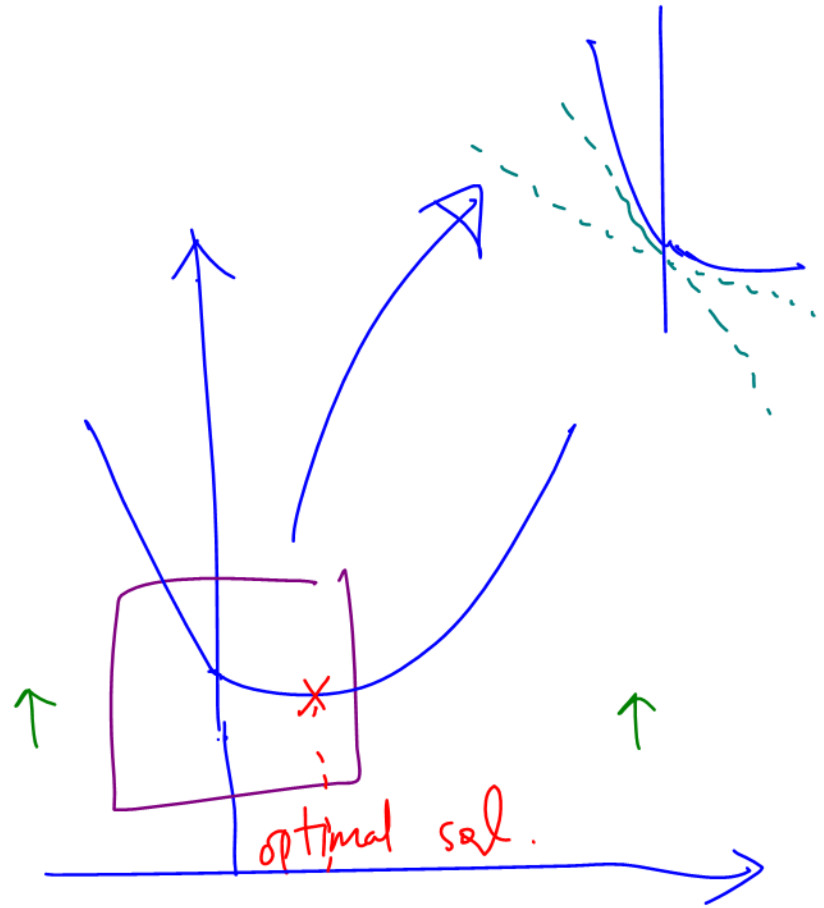


min $l(x) = f(x) + \lambda|x|$



3

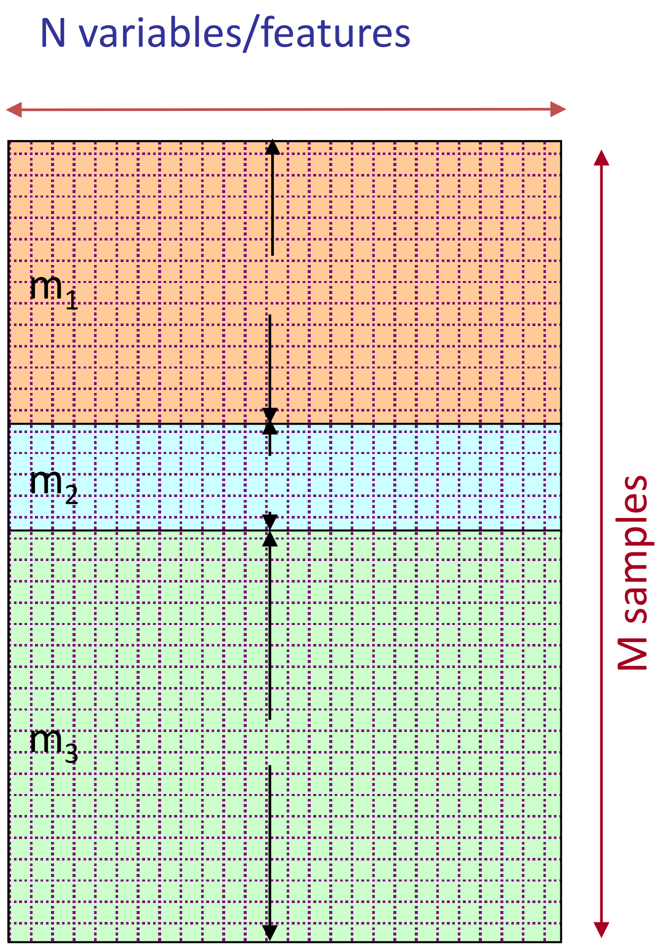




$$\text{if } \begin{cases} \left| \dfrac{\partial f}{\partial x}\Big|_{x=0} \right| > \lambda \\ \phantom{\left| \dfrac{\partial f}{\partial x}\Big|_{x=0} \right|} < \lambda \end{cases} \quad \text{then} \quad x^* \neq 0$$

Wrapping up

Bilevel optimization



Split data into 3 sets:

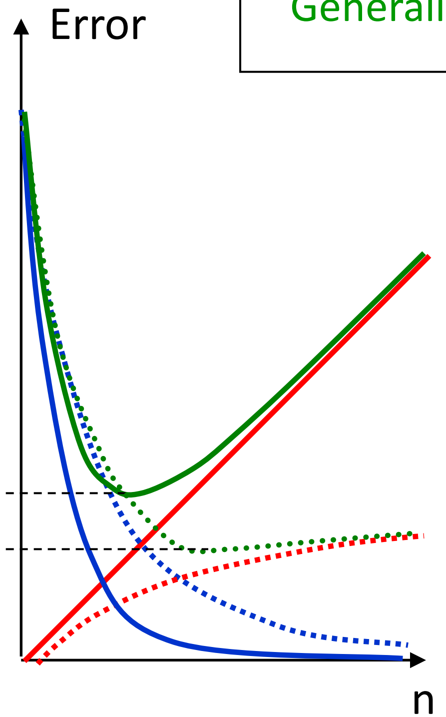
training, validation, and test set.

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.

Complexity of Feature Selection

With high probability:

$$\text{Generalization\_error} \le \text{Validation\_error} + \varepsilon(C/m_2)$$



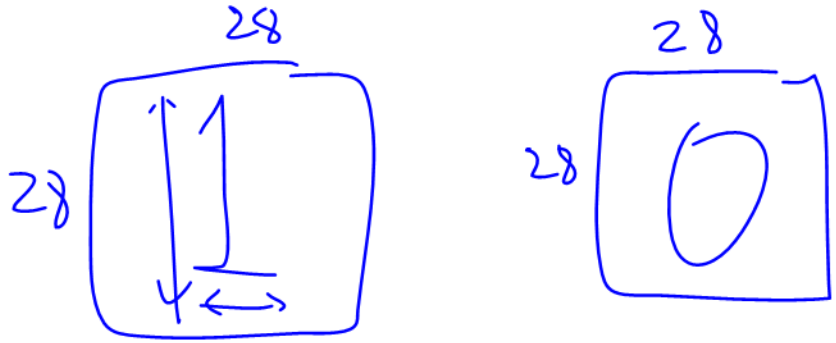
Method	**Number of subsets tried**	**Complexity C**
Exhaustive search wrapper	2^N	N
Nested subsets Feature ranking	N(N+1)/2 or N	log N

m_2: number of *validation* examples,
N: total number of features,
n: feature subset size.

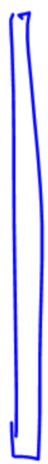
Try to keep C of the order of m_2.

Conclusion

- Feature selection focuses on uncovering subsets of variables X1, X2, … predictive of the target Y.
- Multivariate feature selection is in principle more powerful than univariate feature selection, but not always in practice.
- No method is universally better
  - wide variety of types of variables, data distributions, learning machines, and objectives.
- Match the method complexity to #examples/#features ratio:
  - non-linear classifiers are not always better.
- Feature selection is not always necessary to achieve good performance.



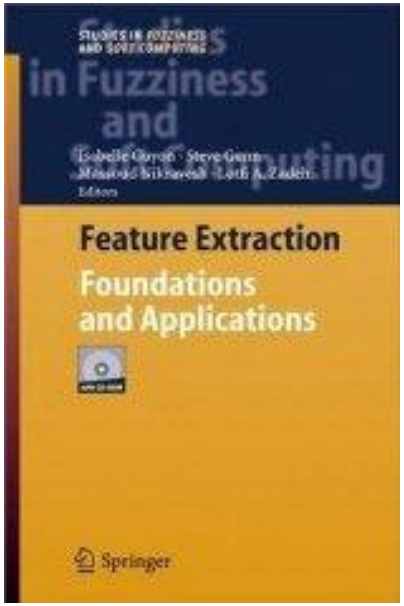
784 pixels

$\phi^1 =$ 

$\phi^2 =$ aspect ratio.

Acknowledgements
and references

1) **Feature Extraction,**
Foundations and Applications
I. Guyon et al, Eds.
Springer, 2006.
http://clopinet.com/fextract-book



2) **Causal feature selection**
I. Guyon, C. Aliferis, A. Elisseeff
To appear in "Computational Methods of Feature Selection",
Huan Liu and Hiroshi Motoda Eds.,
Chapman and Hall/CRC Press, 2007.
http://clopinet.com/causality