

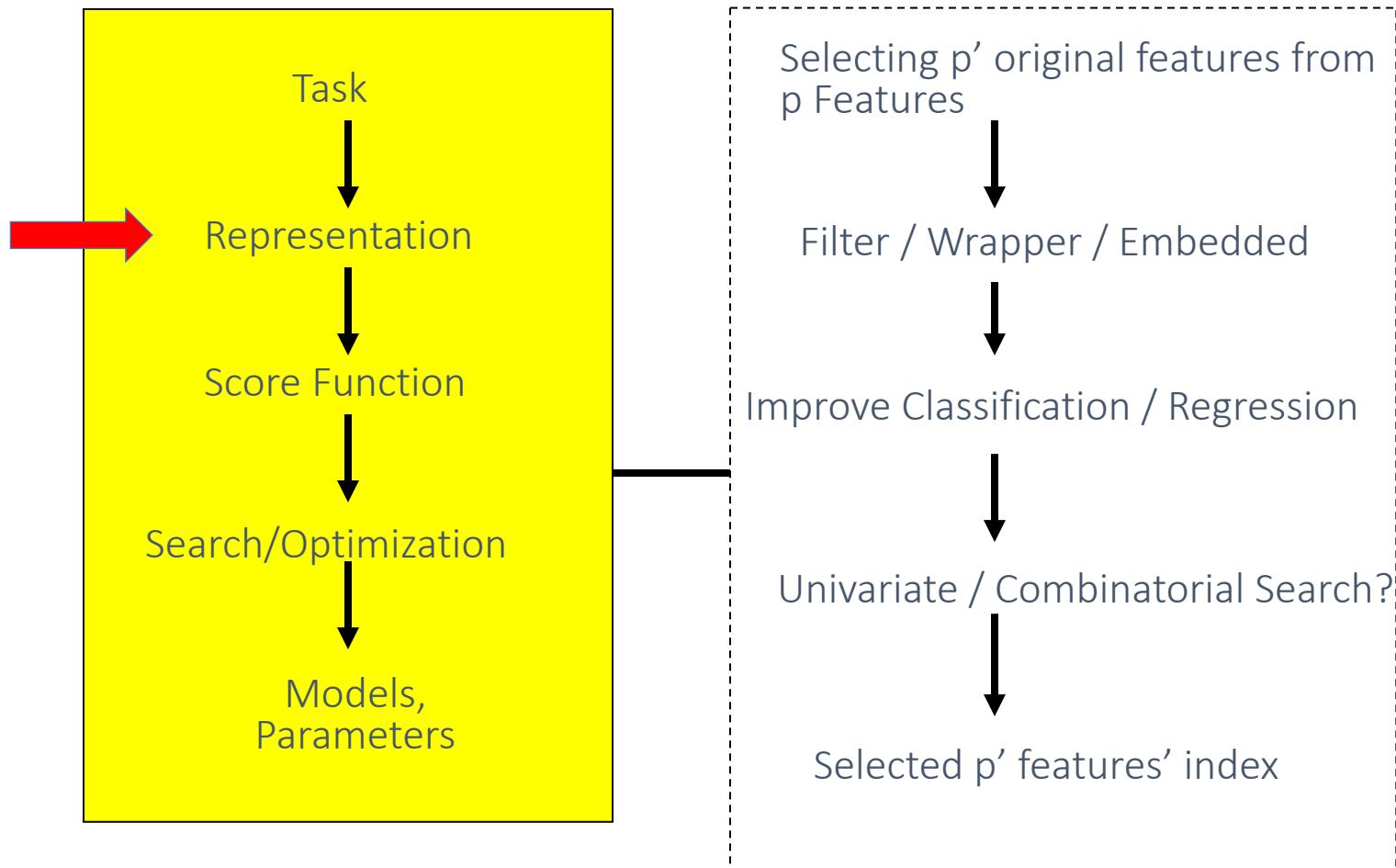
UVA CS 4774: Machine Learning

Lecture S3 Extra: Feature Selection

Dr. Yanjun Qi

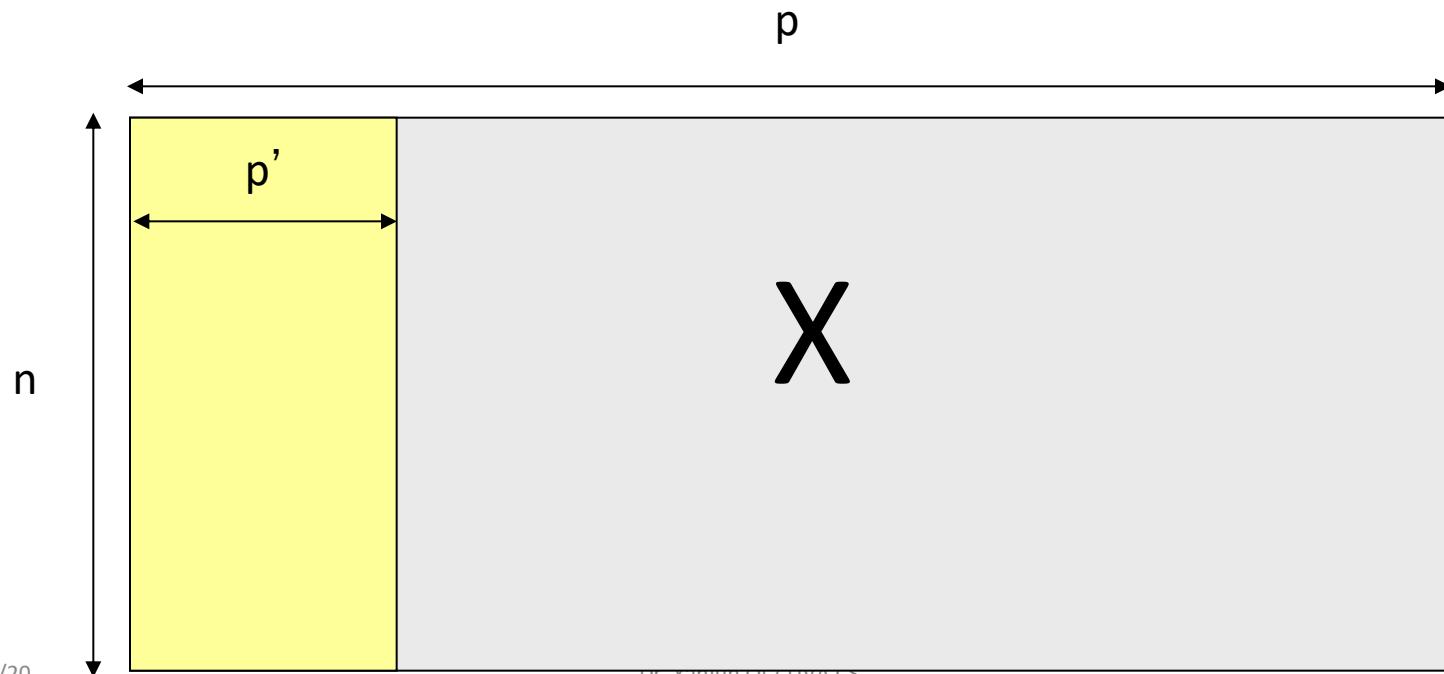
University of Virginia
Department of Computer Science

Today: Feature Selection



Feature Selection

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



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

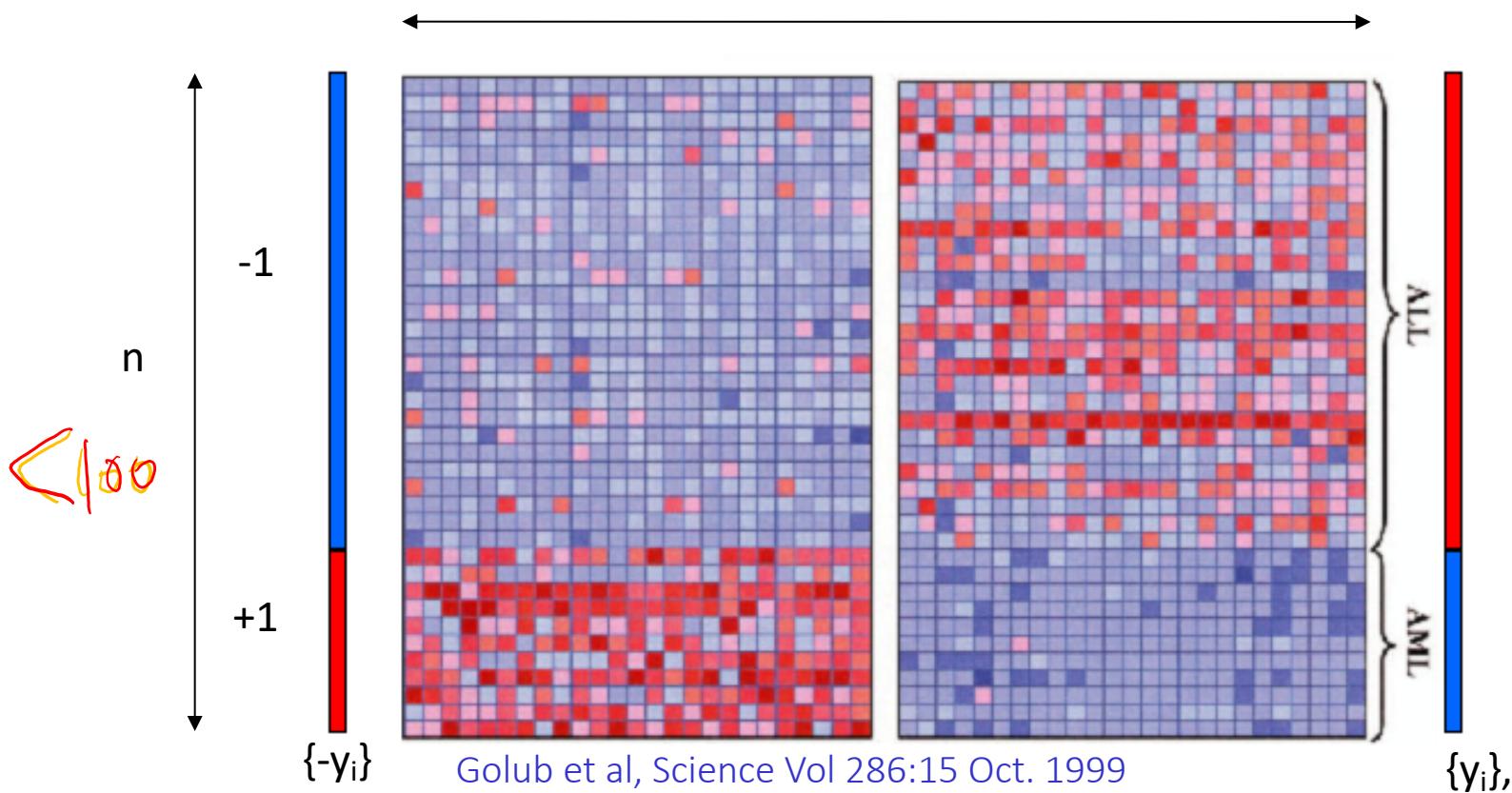
IV. Features

| | |
|------|--|
| 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) |

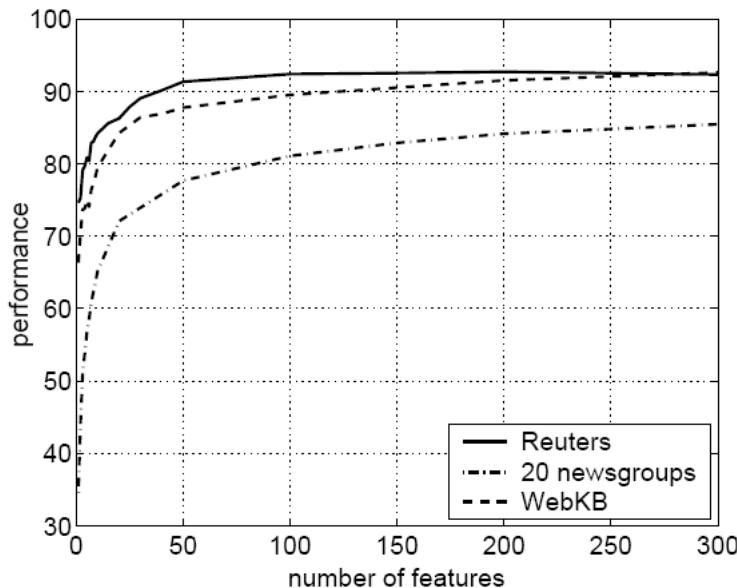
e.g. counts
of a ngram in
the text

e.g., Leukemia Diagnosis

$$p' > 30, 000$$



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

We aim to make the learned model:
Feature Selection → Simpler models

- 1. Generalize Well
 - Less sensitive to noise
 - Lower variance - Occam's razor– (More later!)
- 2. Computationally Scalable and Efficient
 - Easier to train (to need less labeled examples)
 - Simpler to use (computationally)
- 3. Robust / Trustworthy / **Interpretable**
 - Especially for some domains, this is about trust!
 - Easier to explain (more interpretable!)

Occam's razor: law of parsimony

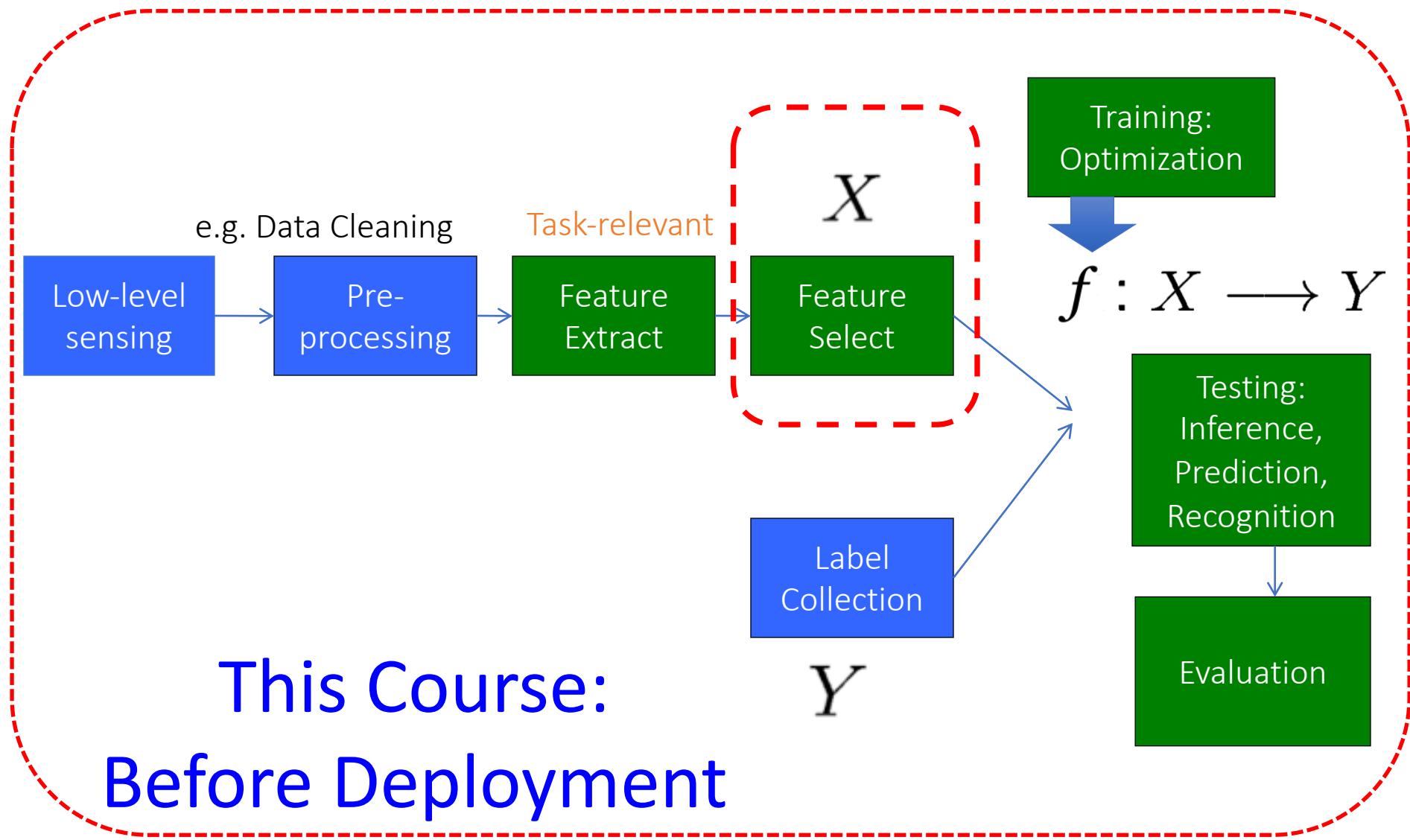
image at:
www.butterflyeffect.ca/.../OccamsRazor.html
remove frame

The principle of Occam's razor

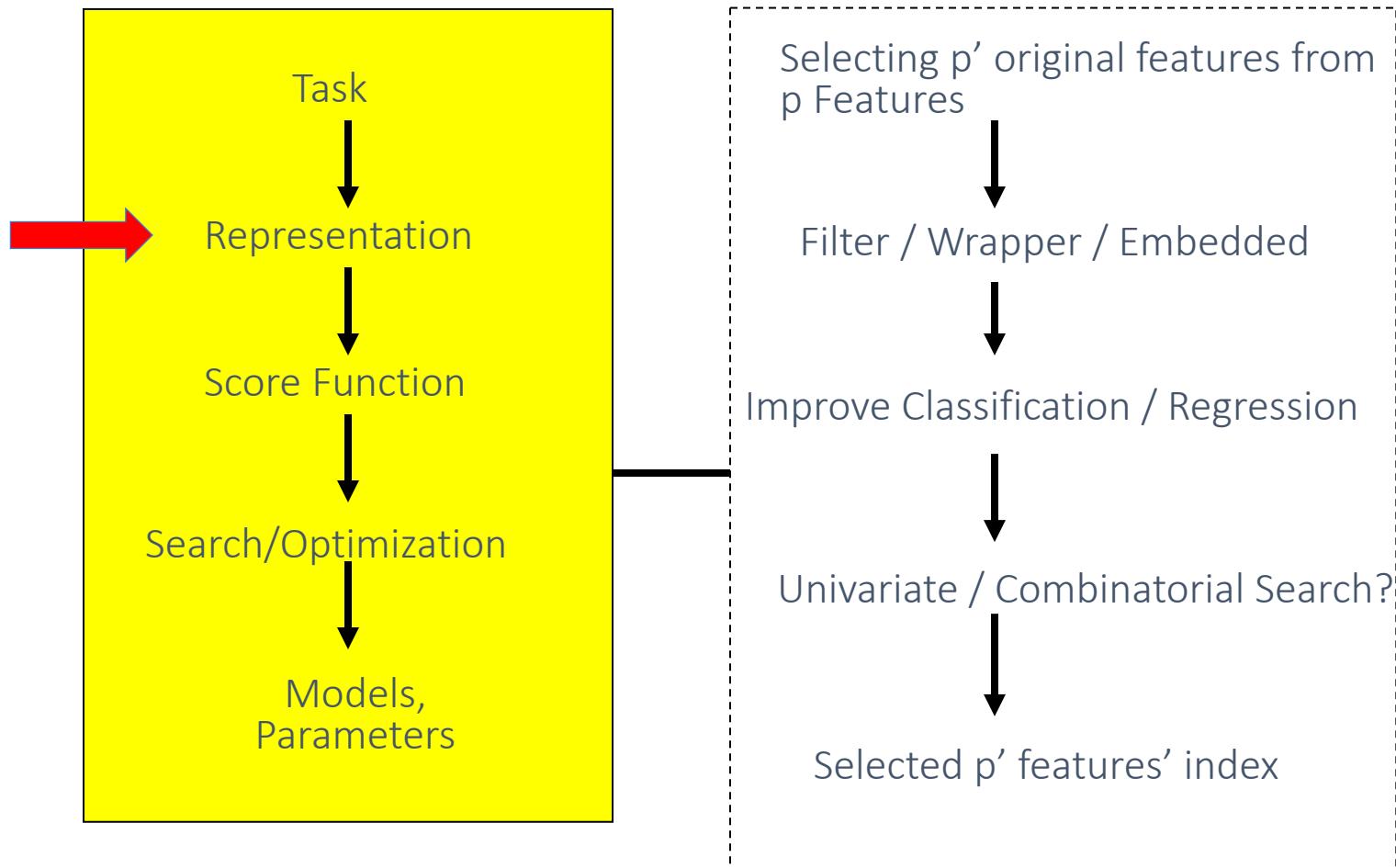
states that the explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference to any observable predictions of the theory



parsimony: extreme unwillingness to spend money or use resources.



Today: Feature Selection



Summary of Feature Selection Methods:

- 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** features or feature subsets.
- Embedding approach:
uses a **predictor to build** a (single) model with a subset of features that are internally selected.

Summary of Feature Selection Methods:

- 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** features or feature subsets.

- Embedding approach:

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

(I) Filtering: Univariate: $\Rightarrow (\Sigma, \Sigma) \rightarrow (\Gamma)$

e.g., Pearson Correlation

- Pearson correlation coefficient

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

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

2 features

$(\Sigma, \Sigma) \rightarrow (\Gamma)$

$|r(\Sigma, \Gamma)| > |r(\Sigma, \Sigma)|$

$\Rightarrow \Sigma_1, \Sigma_2, \dots, \Sigma_p$
 $r(\Sigma_1, \Gamma), \dots, r(\Sigma_p, \Gamma)$

- 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**.

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

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

- Pearson correlation coefficient

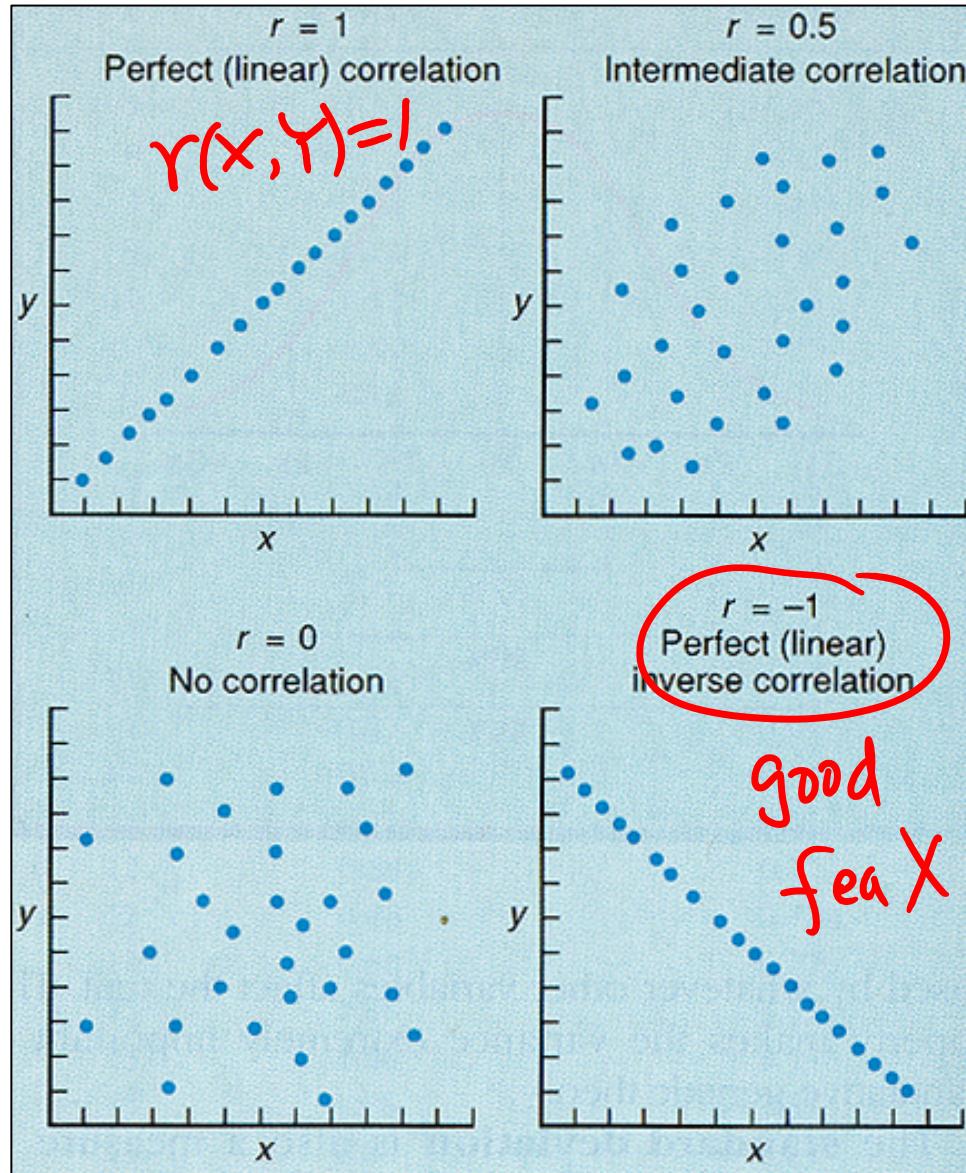
$$r(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \times \sum_{i=1}^n (y_i - \bar{y})^2}}$$
$$|r(x, y)| \leq 1$$

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

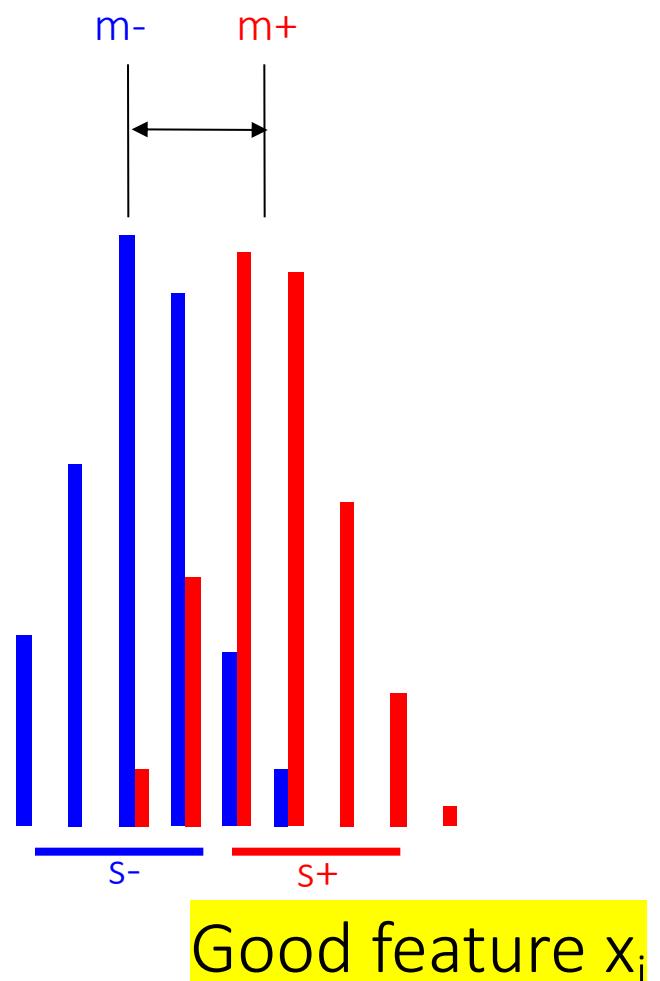
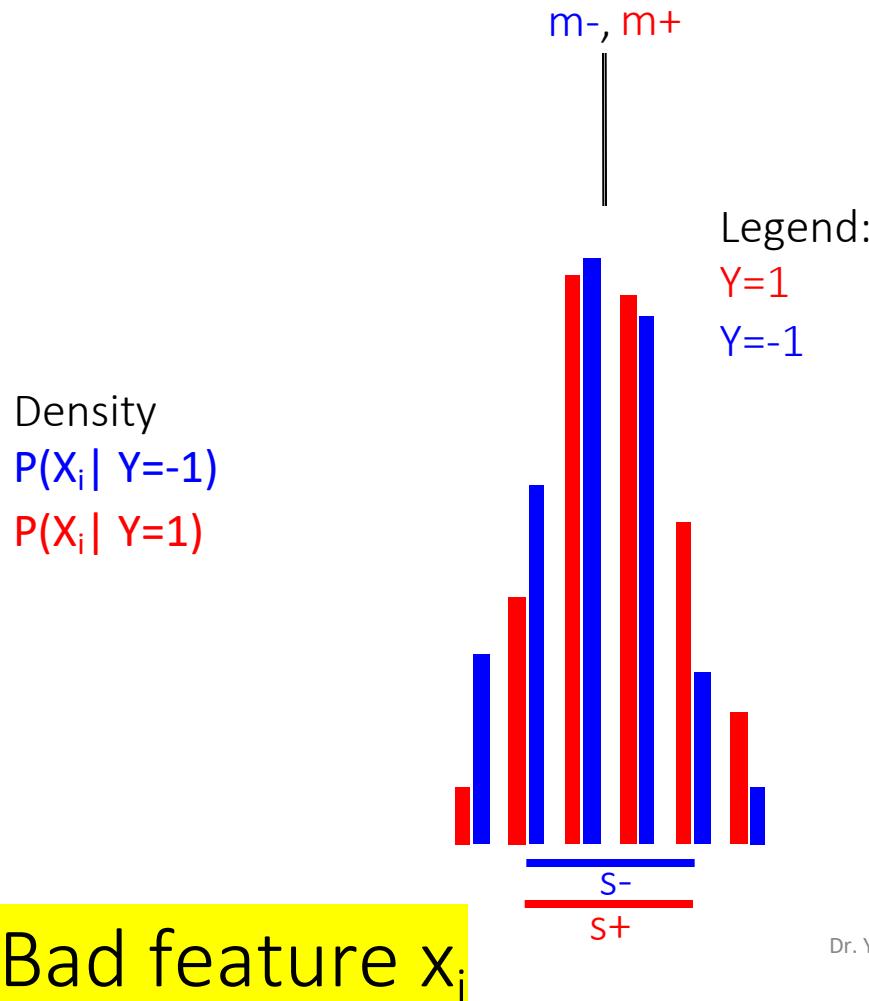


- can only detect **linear dependencies** between two variables
- (e.g. between one feature vs. target)

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

(X_i, Y)
 $\{ \cdot \}_i^{-1}$

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



(I) Filtering: univariate filtering

e.g. T-test

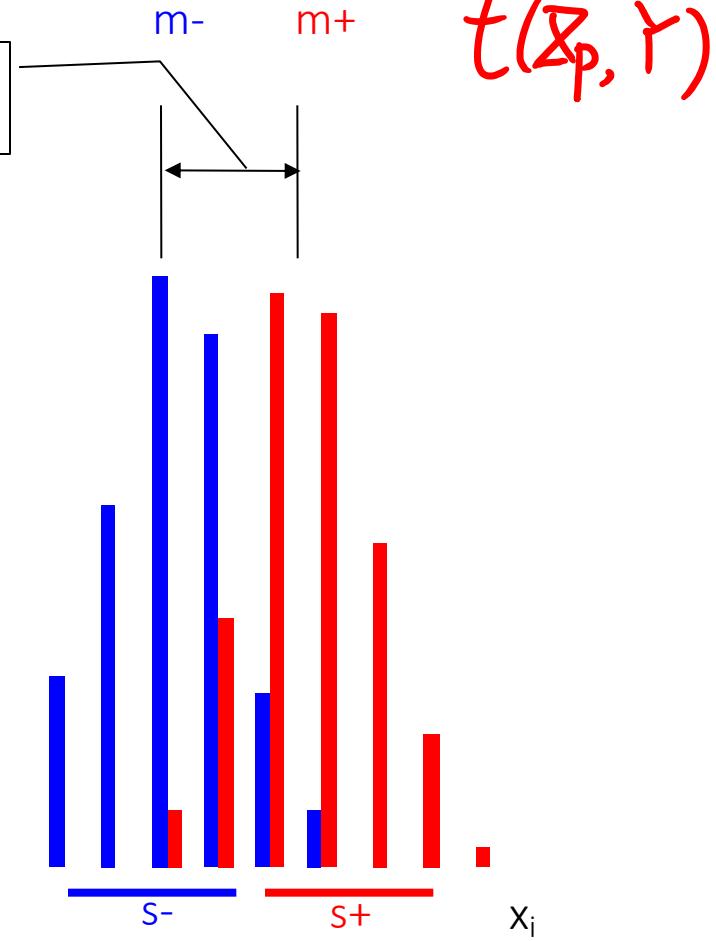
$$\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_p$$

$$t(\mathcal{X}_1, Y), t(\mathcal{X}_2, Y), \dots;$$

$$t(\mathcal{X}_p, Y)$$

T-test

Is this distance significant?



- Assumption: Two Normally distributed classes with equal variance s^2 unknown; estimated from data as s^2_{within} .

- Null hypothesis $H_0: m+ = m-$

- T statistic:

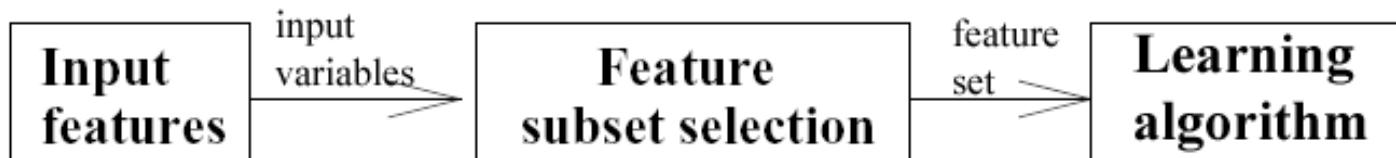
If H_0 is true, then

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

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

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

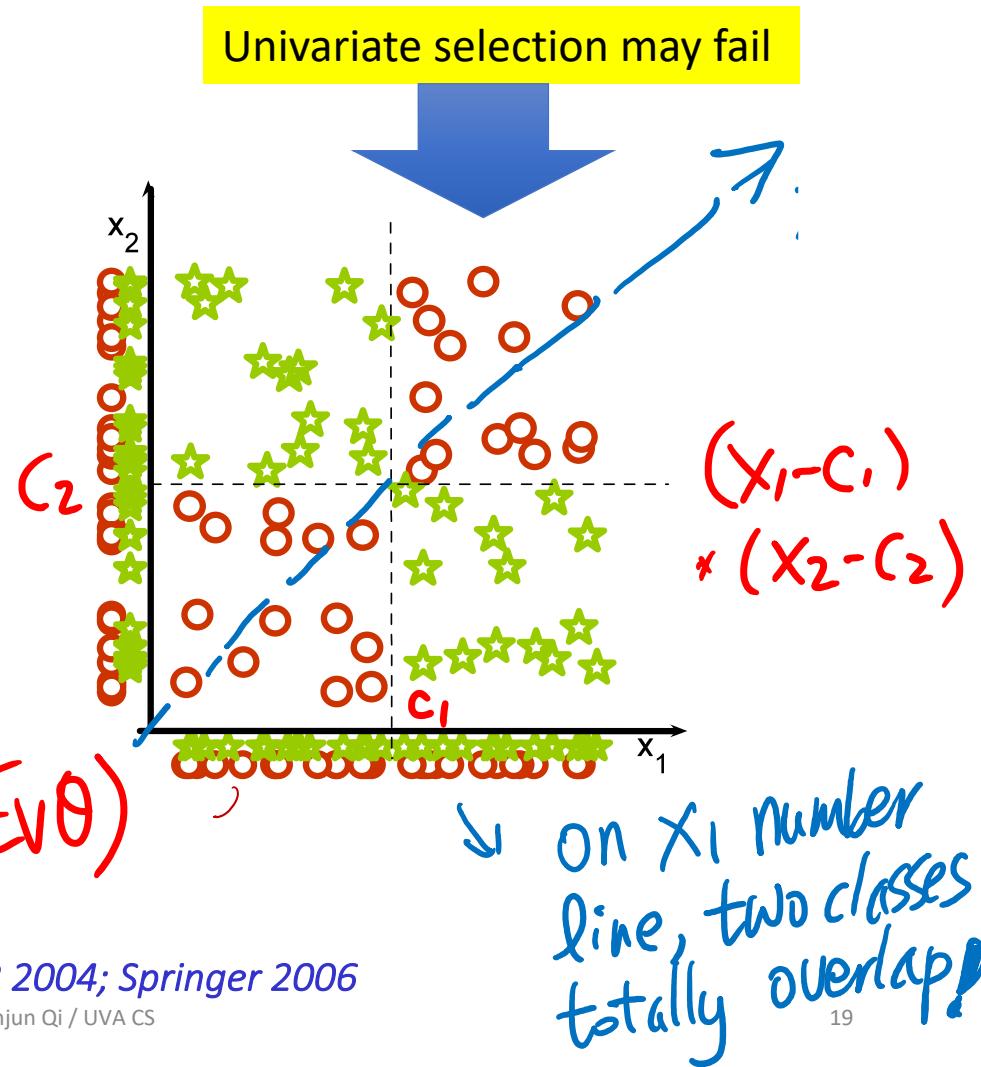
- Filter Methods
 - Select subsets of variables as a pre-processing step, independently of the used classifier!!



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

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

$$\begin{aligned}
 & t(\Sigma_1, Y) \\
 & t(\Sigma_2, Y) \\
 & (\Sigma_1, \Sigma_2, \dots, \Sigma_p) \\
 & \vec{\theta}_{\text{px}} = (\%, \%, \dots, \%) \\
 & O(2^P \times \text{CostEv} \theta)
 \end{aligned}$$



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

Sentiment Classification

e.g. Amazon review

text

X

$\rightarrow Y$

review

score

$1 \sim 5$

many possible

features

words

2gram

3grams

.

k grams

good, not, boring, ...
not good, not boring, ...

Very good,

Very very good,

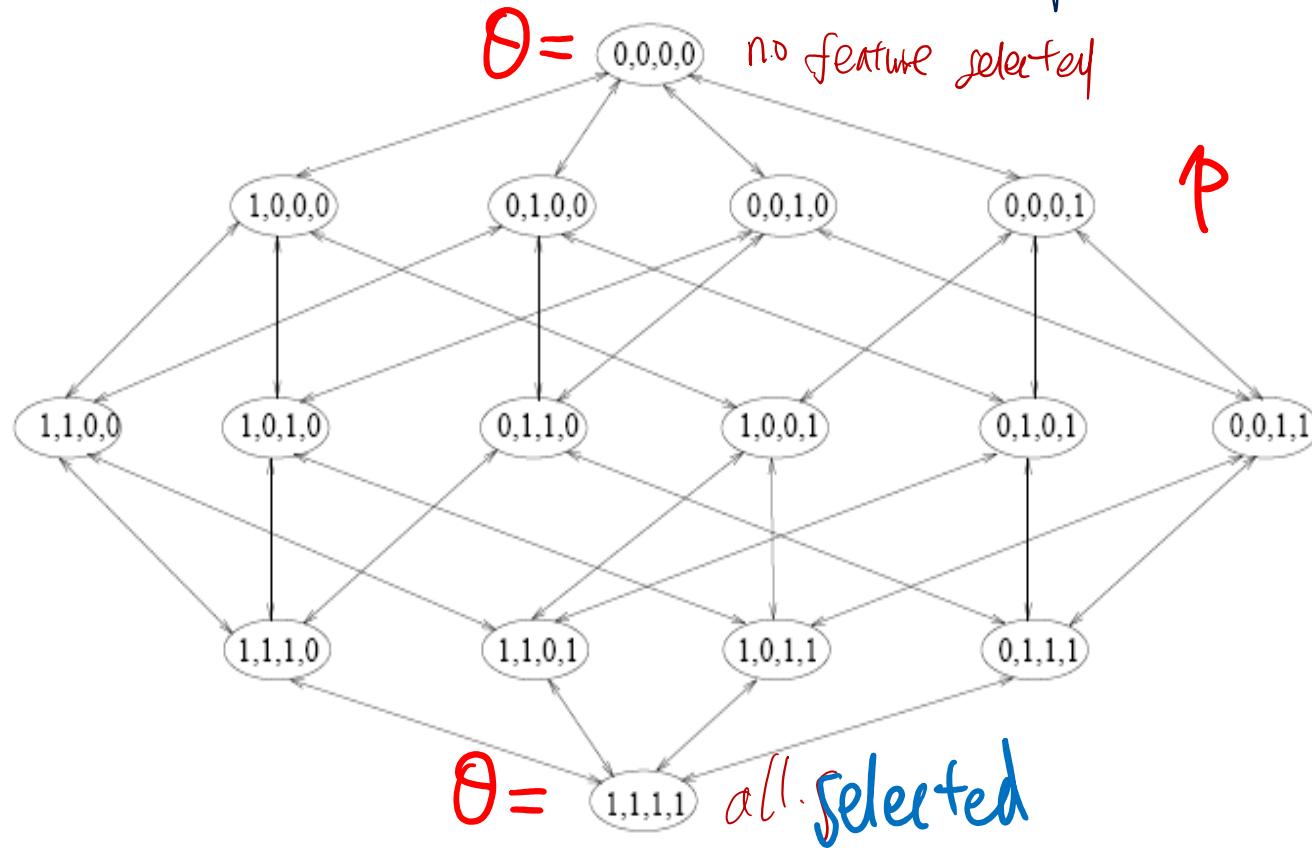
not very boring,

.....

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

- You need:
 - a **measure** for assessing the goodness of a feature subset (scoring function)
evaluating Θ
 - a **strategy** to search the space of possible feature subsets
search Θ
- Finding a minimal optimal feature set for an arbitrary target is **NP-hard**
=> Good heuristics are needed!

each feature subset can be described by $\Theta = [0/1, 0/1, 0/1, \dots, 0/1]^T$
 $p \times 1$ Vector



p features, 2^p possible feature subsets!

(I) Filtering : Summary

- usually fast
- provide generic selection of features, not tuned by given learner (universal, learner-agnostic)
- this is also often criticised (feature set not optimized for used learner)
- Often used as a pre-processing 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 | Harmonic of recall and precision, popular in information retrieval. | |
| Odds ratio | ✓ | Eq. 3.6 | + 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 | 10/12/20 | Eq. 3.37 | + s | + + s | So far rarely used. |
| MDL | | Eq. 3.38 | + s | + s | Dr. Yanjun Qi / UVA CS Low bias for multivalued features. |

Summary of Feature Selection Methods:

- Filtering approach:
 - ...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** features or feature subsets.
- Embedding approach:
 - uses a **predictor to build** a (single) model with a subset of features that are internally selected.

(2) Wrapper : Feature Subset Selection

- 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.
 $\theta \rightarrow (\mathbb{R}^{n \times p'}, Y)$
- Results vary for different learners

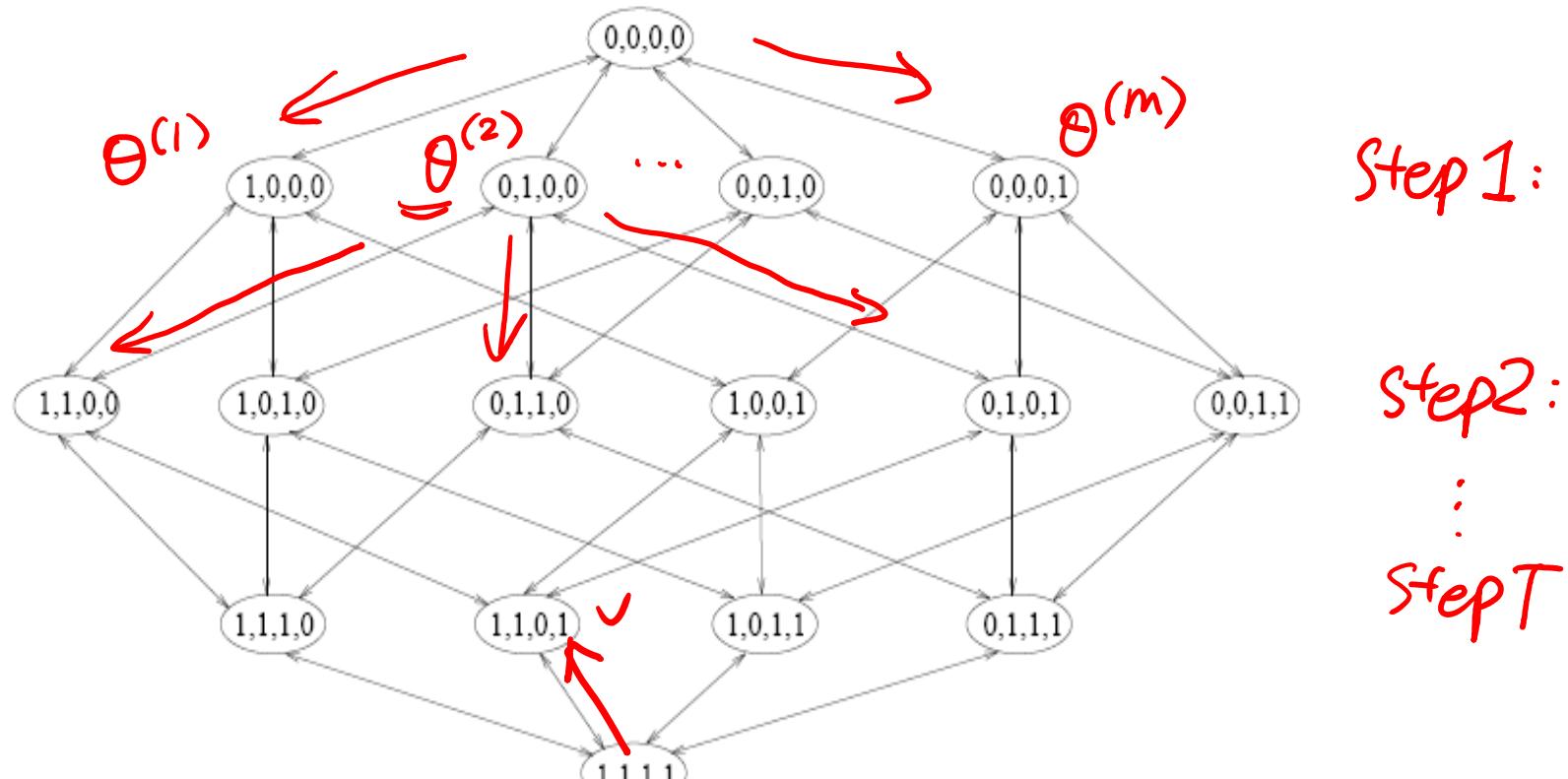
(2) Wrapper : Feature Subset Selection

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

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

- 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.

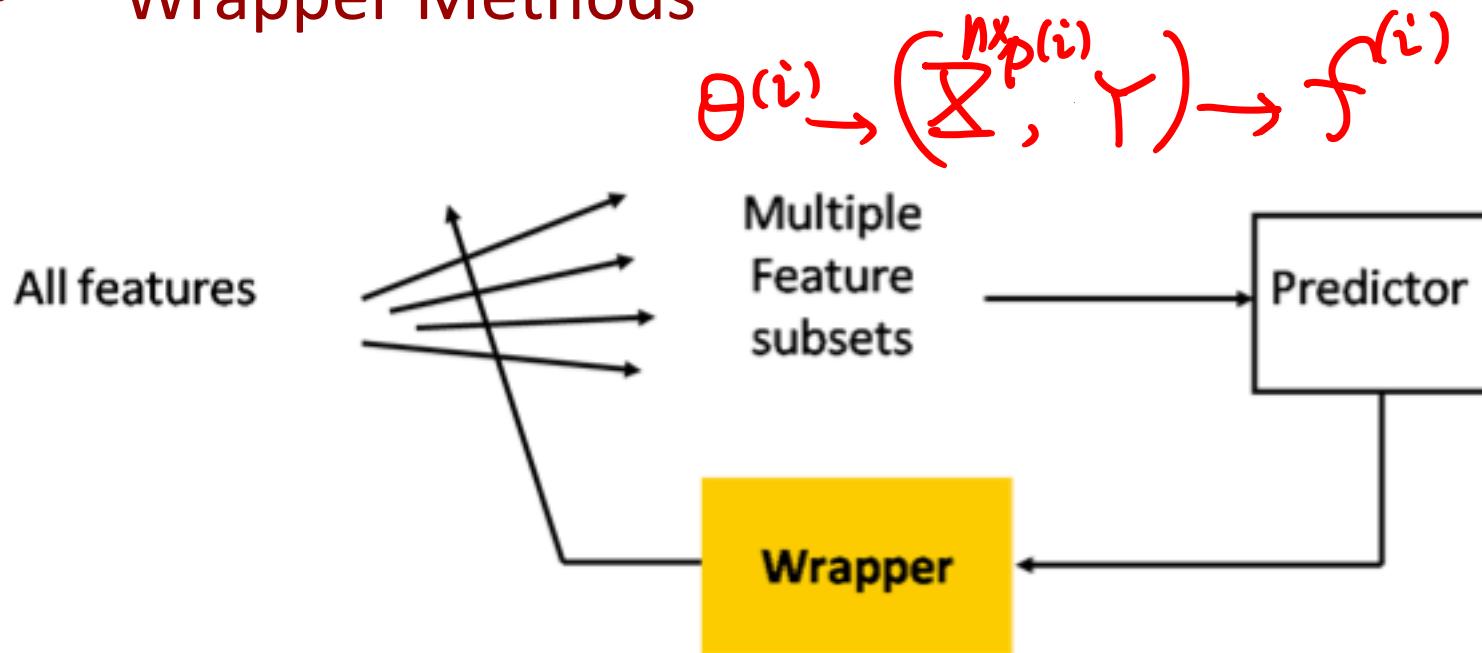
(b). Search: How to search the space of all feature subsets ?



$$2^P \Rightarrow P(P-1)(P-2)\dots(P-T)$$

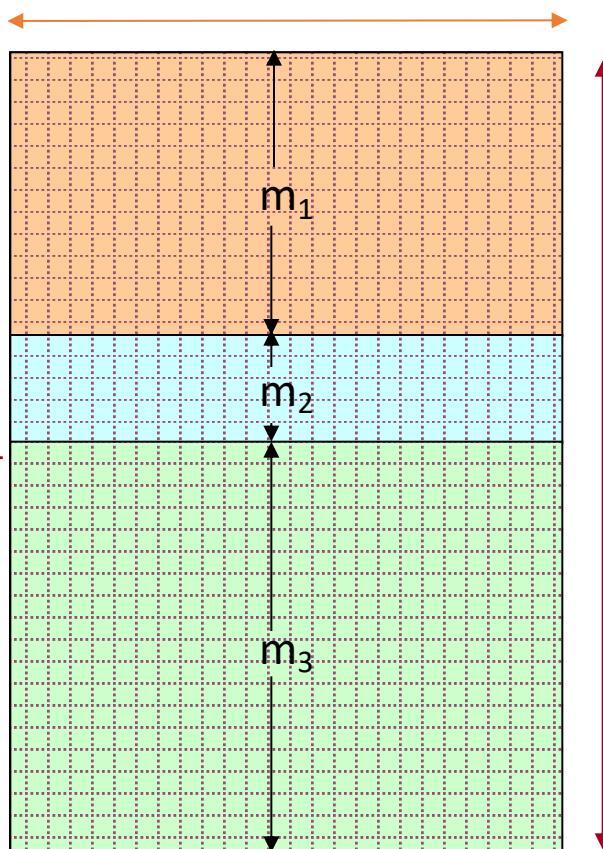
(a). Assessment: How to access multiple candidates of feature subsets

- Wrapper Methods



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

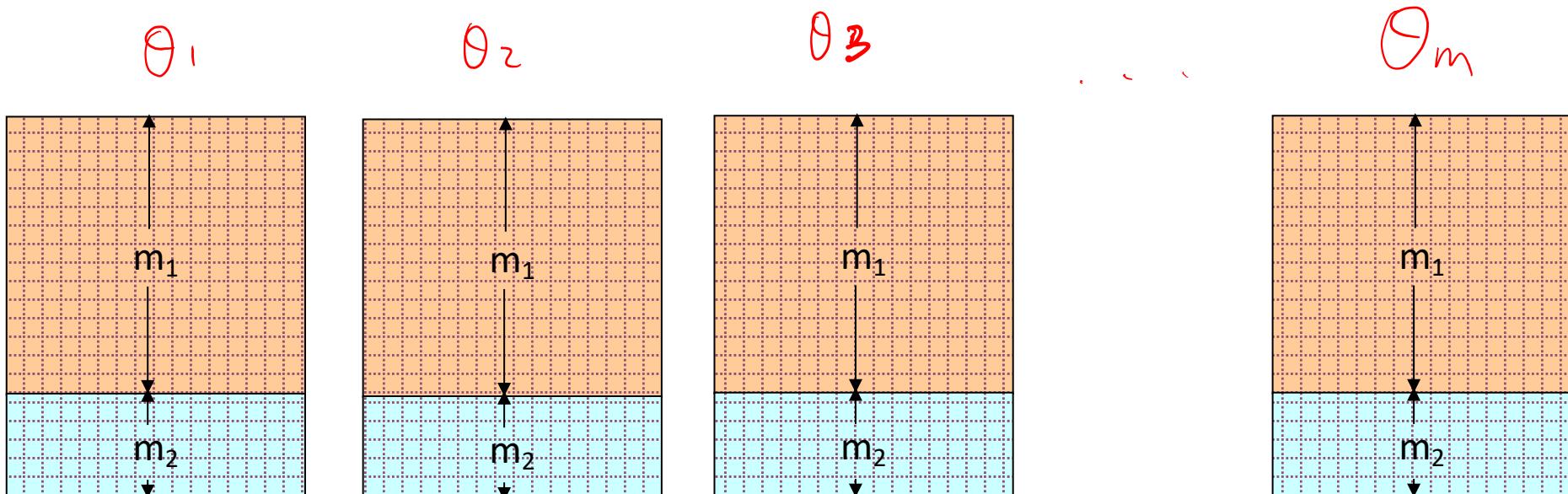
p variables/features



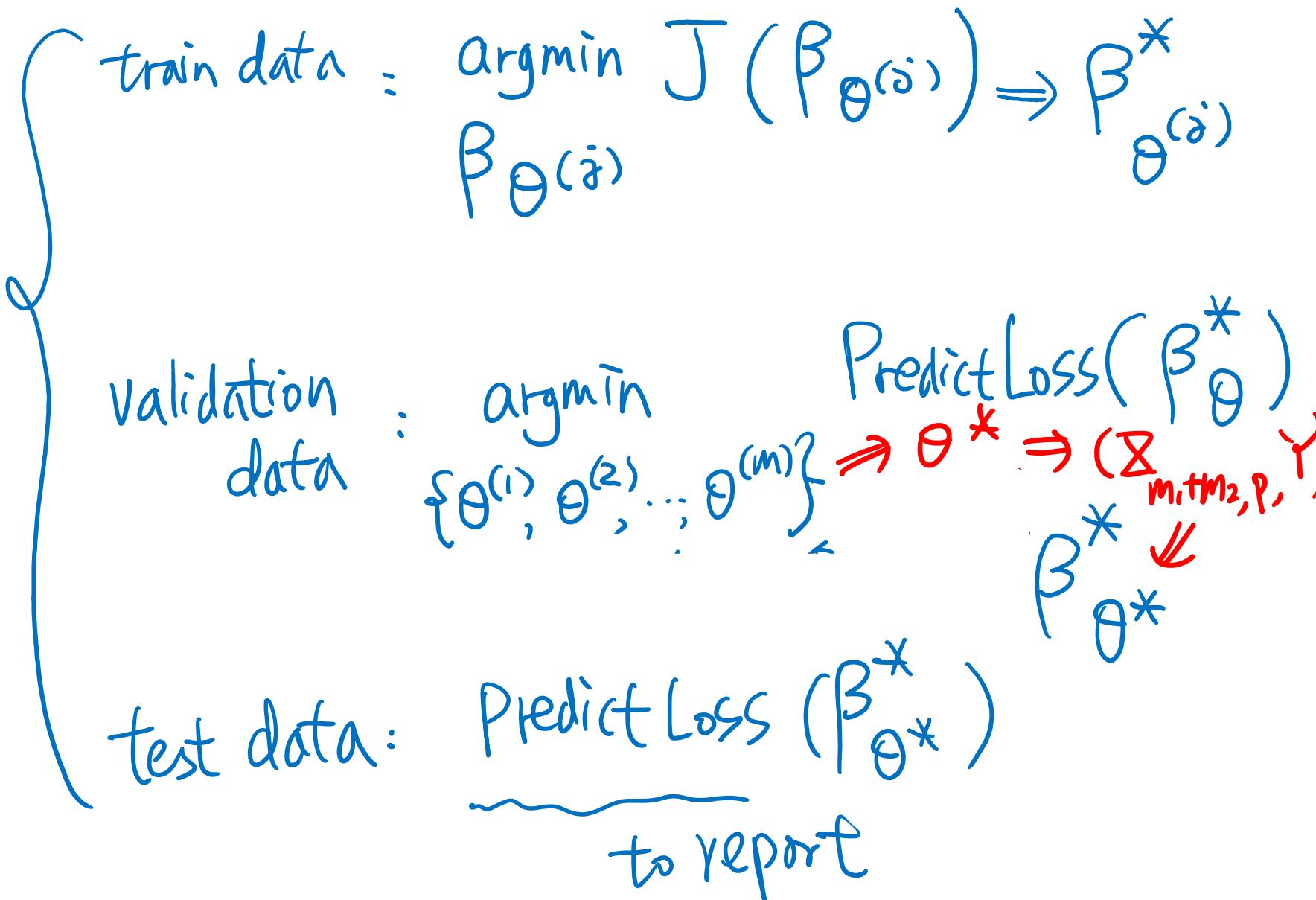
Split data into 3 sets:
training, validation, and test set.

- 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**.

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



{ train for m times on train fold
 ~~test~~ for m times on validation fold
validate \downarrow
argmin ValError $\Rightarrow \theta^*$

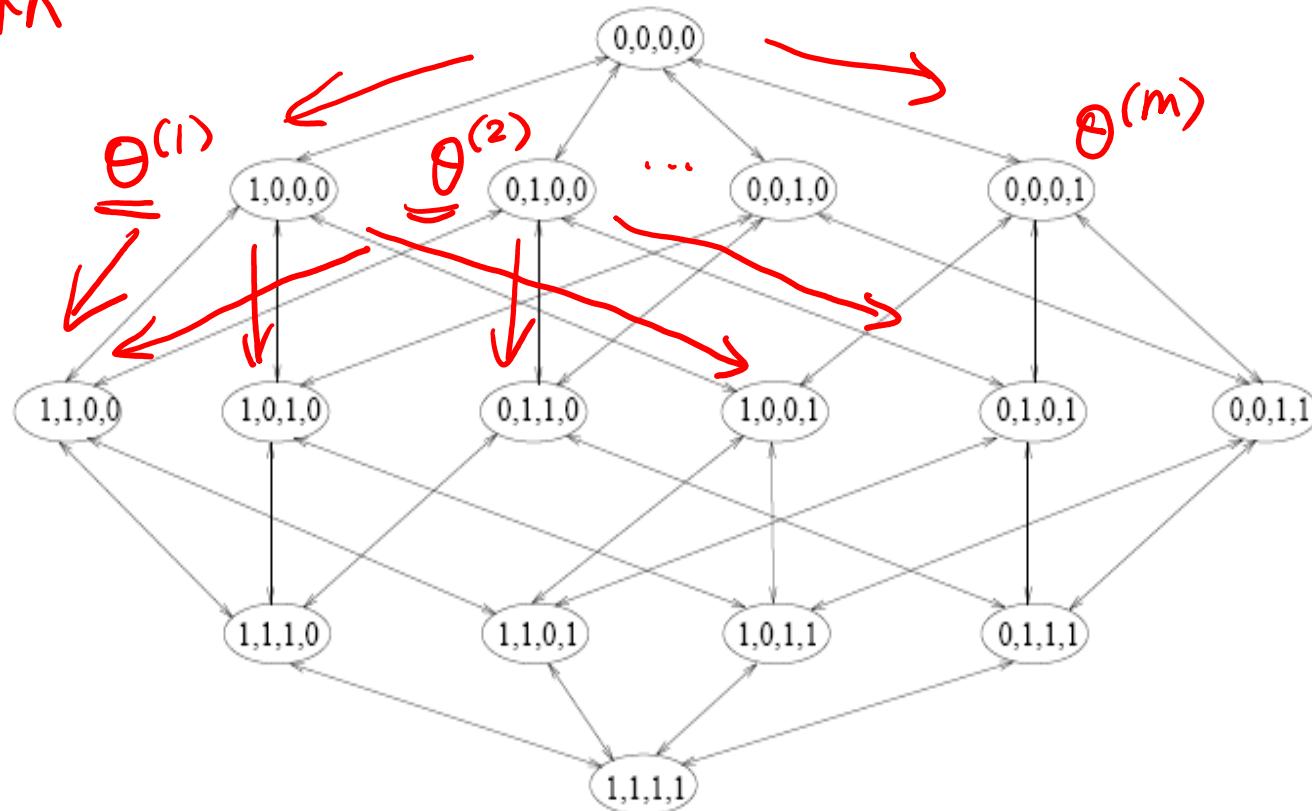


(b). Search: even more search strategies for selecting feature subset

- **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.

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

e.g. BEAM
search
[keep
top $k=2$
path]



Step 1:

Step 2:

:

Step T

$$2^P \Rightarrow P(P-1)(P-2)\dots(P-T)$$

Summary of Feature Selection Methods:

- 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** features or feature subsets.
- Embedding approach:
uses a **predictor to build** a (single) model with a subset of features that are internally selected.

(3) Embedded: Feature Subset Selection

L_1 , $L_1 + L_2$, Structured L_1

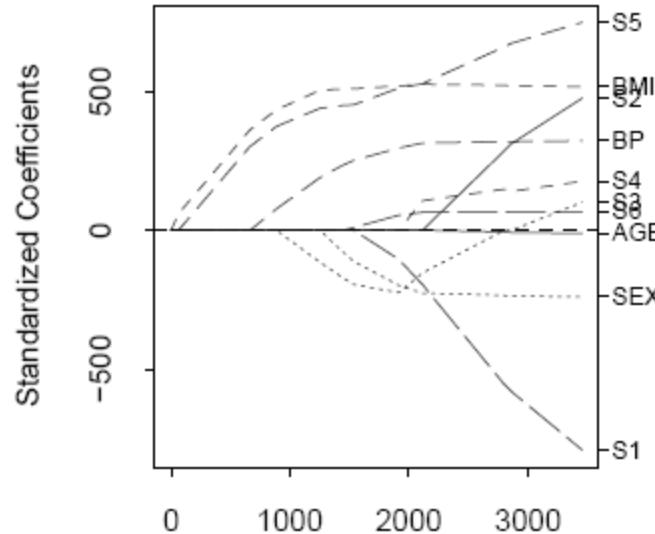
- Specific to a given learning machine!
- Performs variable selection (implicitly) in the process of training
- Just train a (single) model

(3) Embedded: e.g. Feature Selection via Embedded Methods: e.g., L₁-regularization

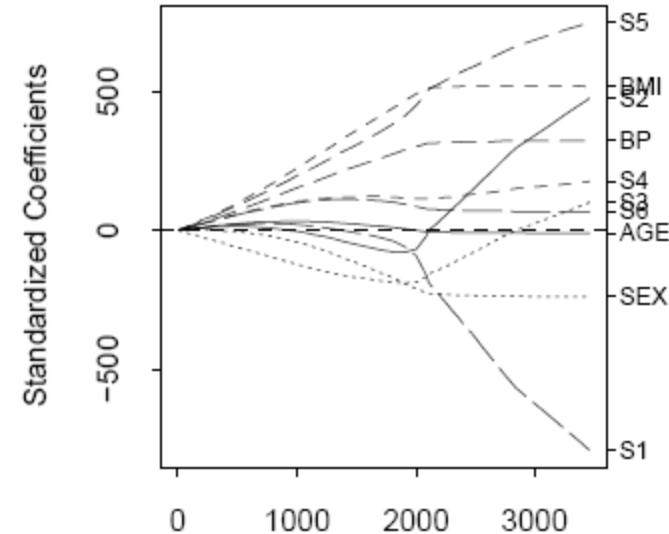
L_1 penalty: $y \sim Model(X\beta) + \lambda \sum |\beta_i|$ (lasso)

L_2 penalty: $y \sim Model(X\beta) + \lambda \sum \beta_i^2$ (ridge regression)

LASSO

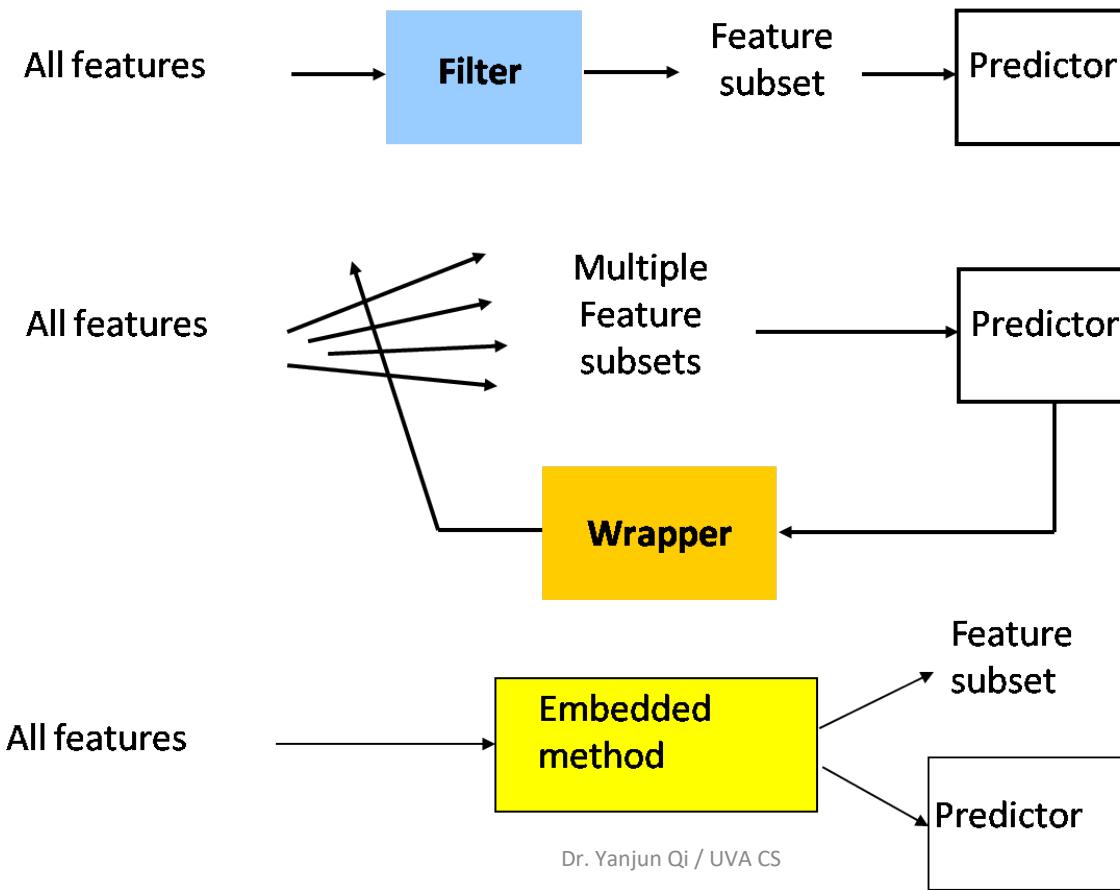


Ridge Regression



Summary: filters vs. wrappers vs. embedding

- Main goal: rank subsets of useful features



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>

Recap: Feature Extraction

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

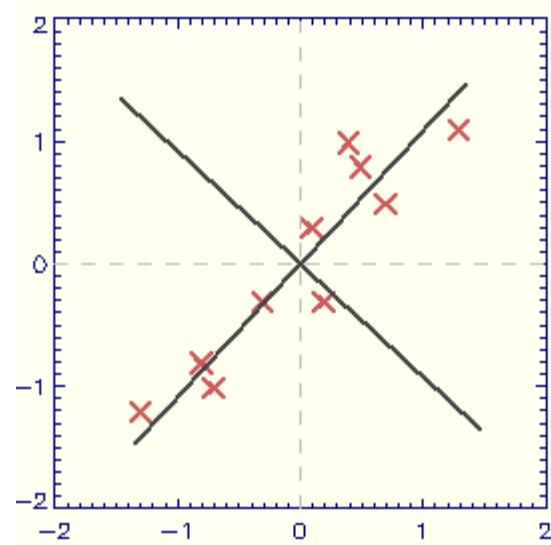
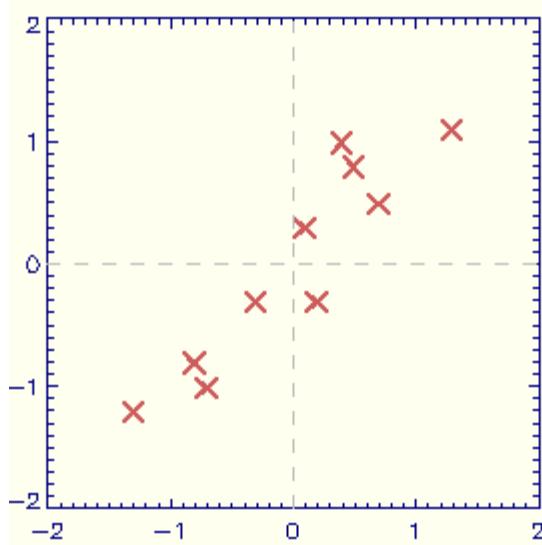
Feature Selection: $X_1, \dots, X_p \rightarrow X_{k1}, \dots, X_{kp'}$

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

Recap: Feature Extraction

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.

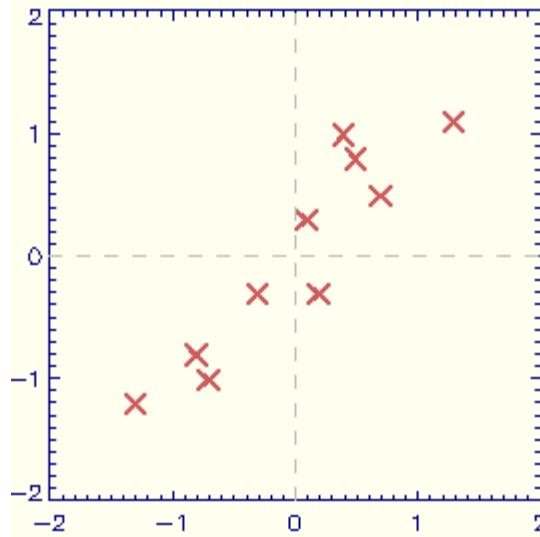


Dataset X is mapped to dataset X' , here of the same dimensionality. The first dimension in X' (= the first principal component) is the direction of maximal variance. The second principal component is orthogonal to the first.

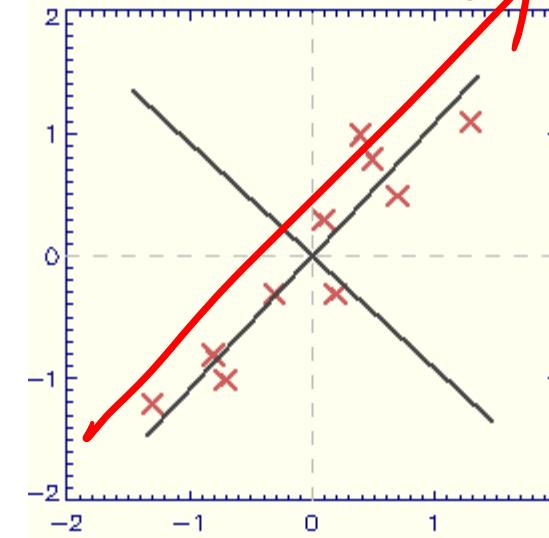
Later: Dimensionality Reduction, e.g., (Linear) Principal Components Analysis

- 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.

x_2



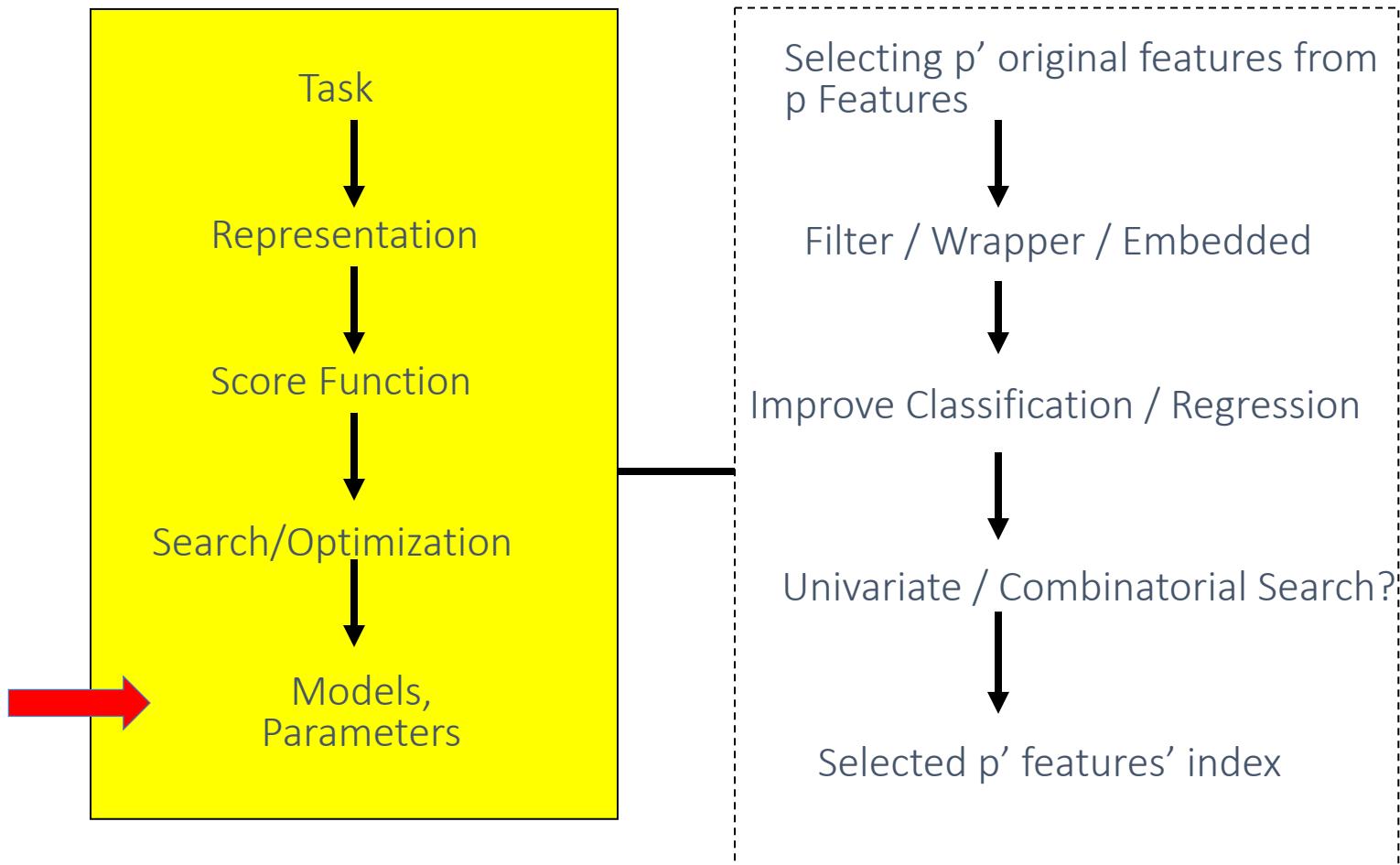
x_1



g_1

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.

Today: Feature Selection

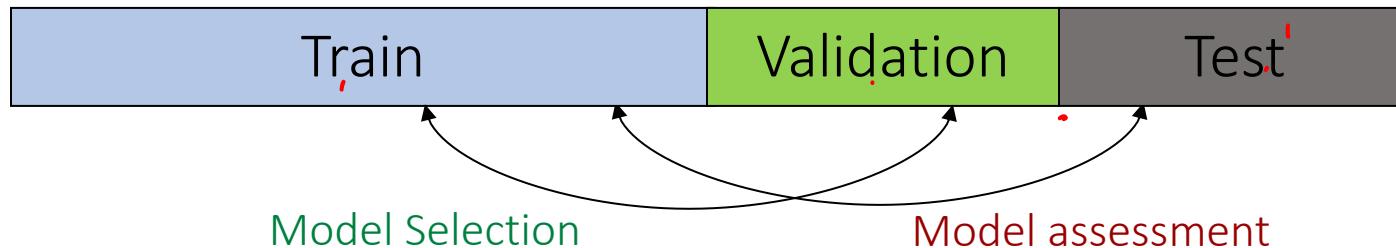


Model Selection and Assessment

- Model Selection
 - Estimating performances of different models to choose the best one
- Model Assessment
 - Having chosen a model, estimating the prediction error on new data

Model Selection and Assessment

- When Data Rich Scenario: Split the dataset



- When Insufficient data to split into 3 parts
 - Approximate validation step analytically
 - AIC, BIC, MDL, SRM
 - Efficient reuse of samples
 - Cross validation, bootstrap

Model Selection (Hyperparameter Tuning) Model Assessment Pipelines in HW2

- (1) train / Validation / test
- (2) k-CV on train to choose hyperparameter / then test

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