UVA CS 6316: Machine Learning

Lecture 7: Feature Selection

Dr. Yanjun Qi

University of Virginia
Department of Computer Science

Course Content Plan ➔ Six major sections of this course

- ❑ Regression (supervised) — Y is a continuous
- ❑ Classification (supervised) — Y is a discrete
- ❑ Unsupervised models — NO Y
- ❑ Learning theory — About f()
- ❑ Graphical models — About interactions among X1,… Xp
- ❑ Reinforcement Learning — Learn program to Interact with its environment

Last: Regularized multivariate linear regression

Task → Representation → Score Function → Search/Optimization → Models, Parameters

Regression → Y = Weighted linear sum of X' s → Least-squares + Regularization → Linear algebra for Ridge / sub-GD for Lasso & Elastic → Regression coefficients (regularized weights)

$$\min J(\beta) = \sum_{i=1}^{n} \left(Y - \hat{Y} \right)^2 + \lambda \left(\sum_{j=1}^{p} \beta_j^q \right)^{1/q}$$

Today: Feature Selection

Task

Representation

Score Function

Search/Optimization

Models, Parameters

Selecting p' original features from p Features

Filter / Wrapper / Embedded

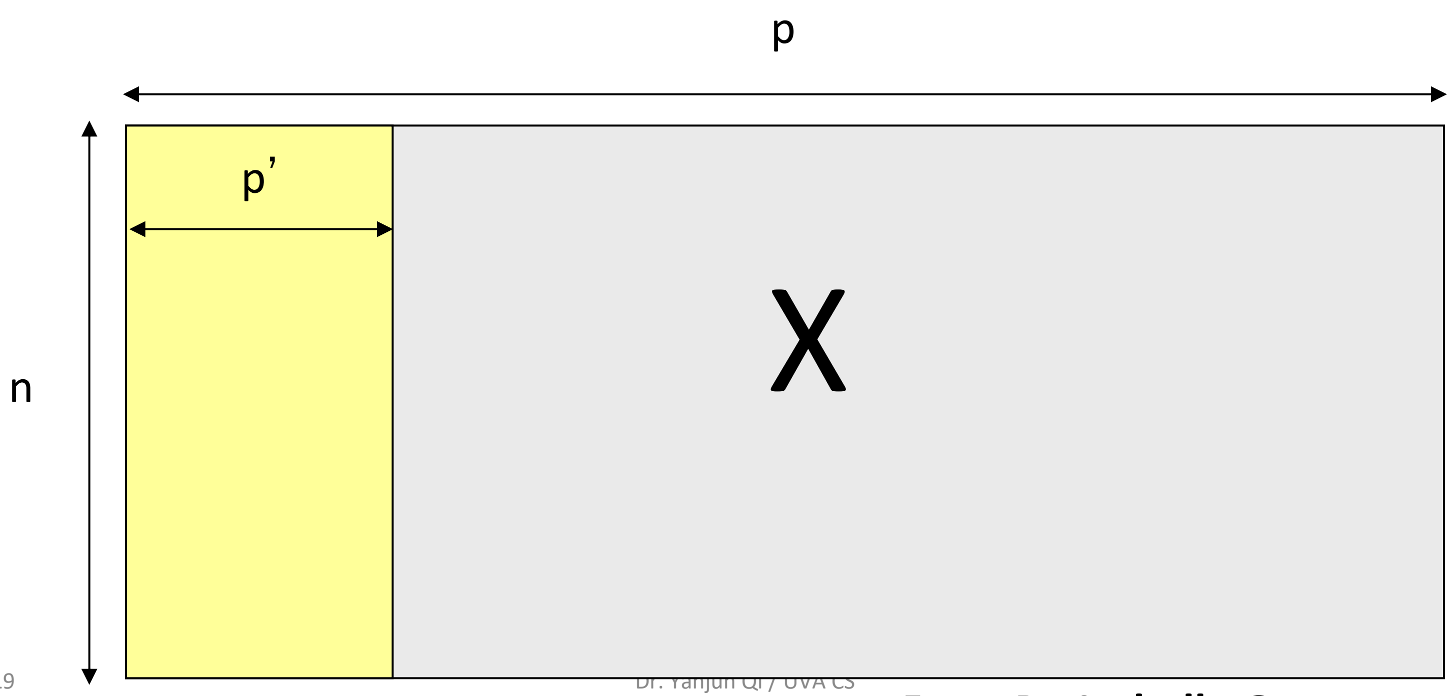
Improve Classification / Regression

Univariate / Combinatorial Search?

Selected p' features' index

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)

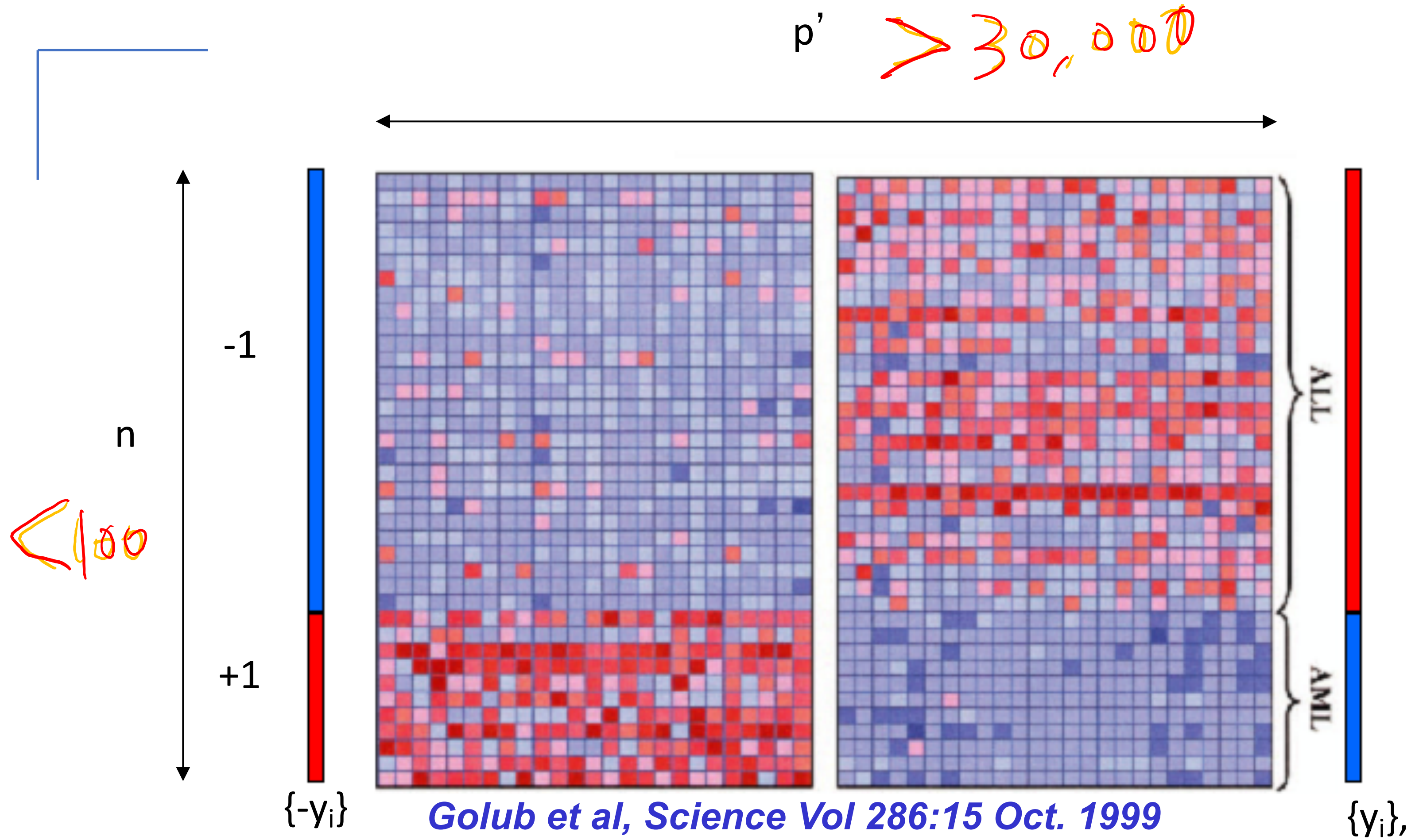
IV. Features

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

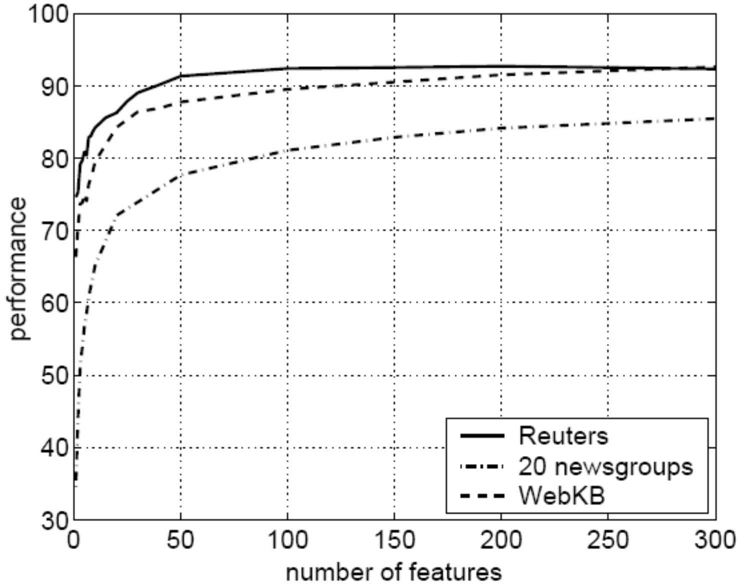
$n \approx 1700$ / $p > 30{,}000$

e.g., Leukemia Diagnosis



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

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!)

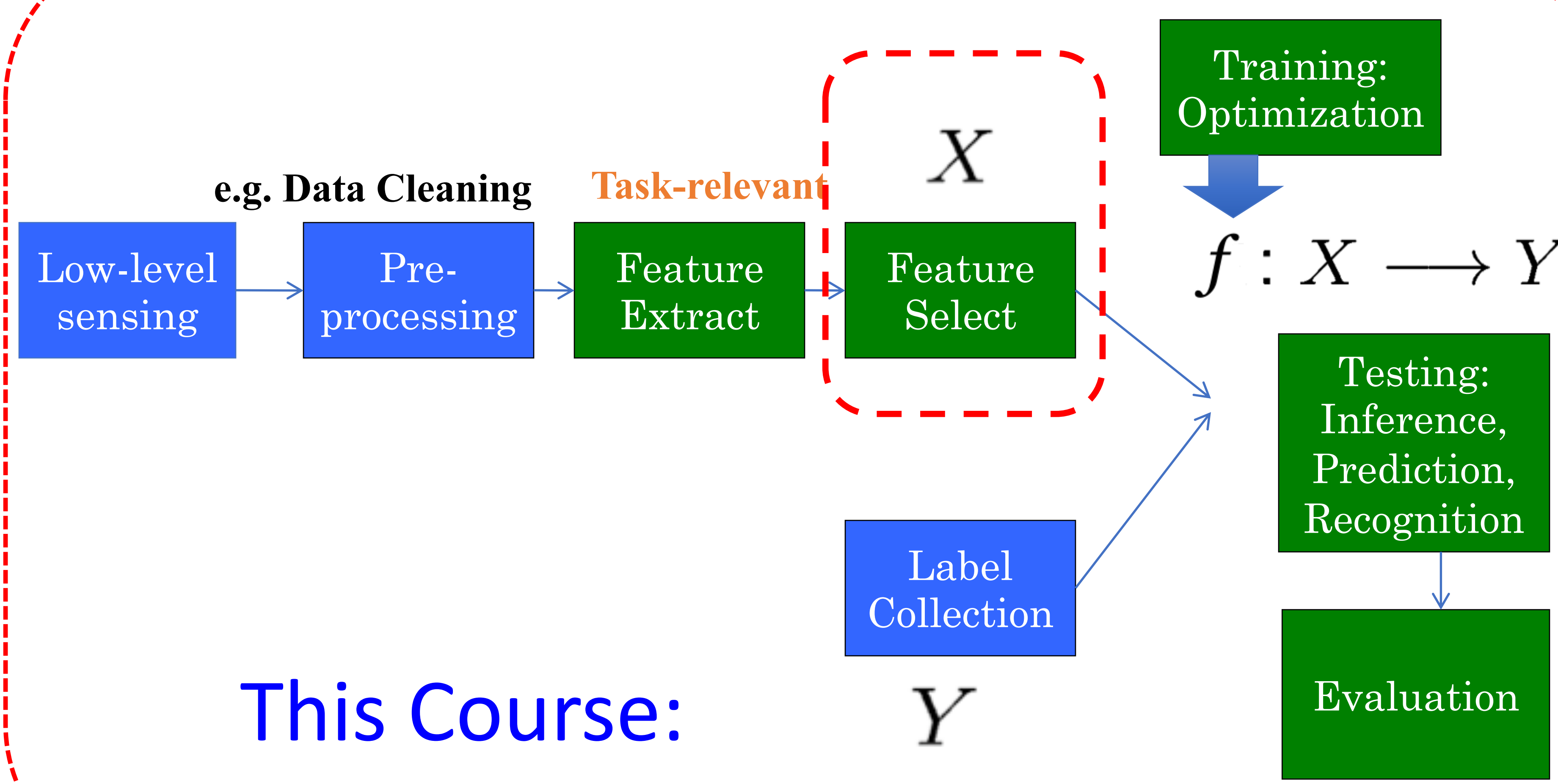
Occam's razor: law of parsimony

image at:
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.



This Course: Before Deployment

Today: Feature Selection

Task

↓

Representation

↓

Score Function

↓

Search/Optimization

↓

Models, Parameters

Selecting p' original features from p Features

↓

Filter / Wrapper / Embedded

↓

Improve Classification / Regression

↓

Univariate / Combinatorial Search?

↓

Selected p' features' index

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:

e.g., Pearson Correlation

2 features

$\Rightarrow (X, Z) \longrightarrow (Y)$

$|r(X,Y)| > |r(Z,Y)|$

$\Rightarrow X_1, X_2, \dots, X_p$

$r(X_1,Y), \dots \quad r(X_p,Y)$

- Pearson correlation coefficient

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

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

- 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**.

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

Correlation is unit independent

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

- Pearson correlation coefficient

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

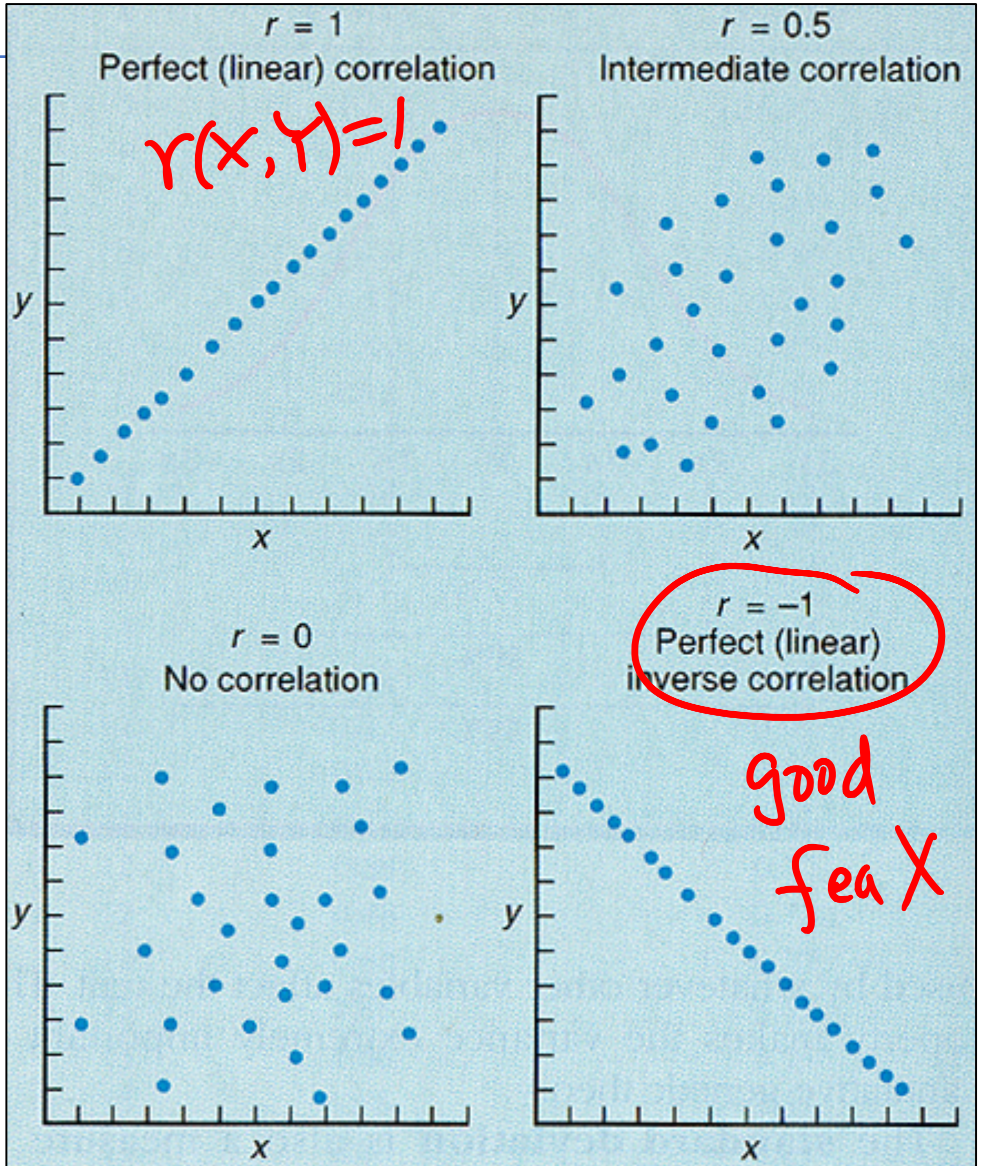
$$|r(x,y)|\leq 1$$

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

- 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)

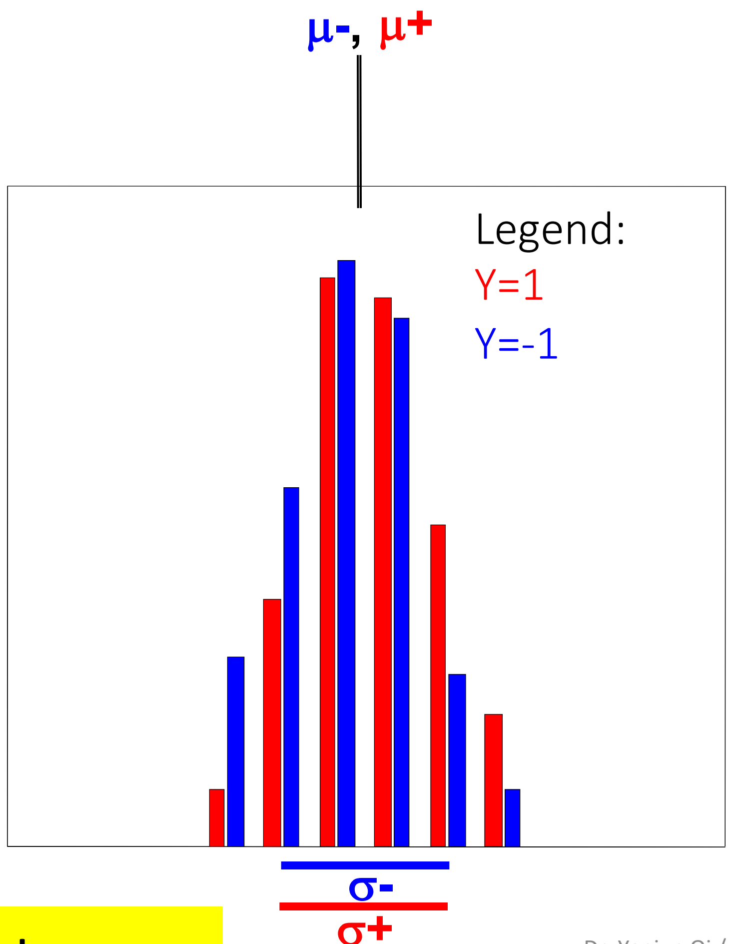
(I) Filtering: univariate filtering

e.g. T-test

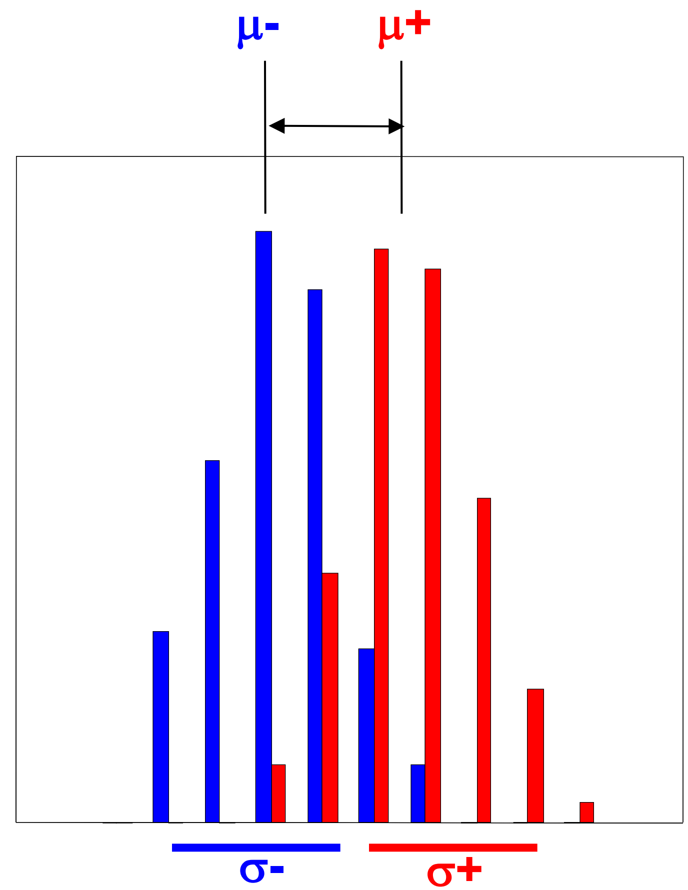
(X_i, Y) $\{-1, 1\}$

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

Density
$P(X_i | Y=-1)$
$P(X_i | Y=1)$



Bad feature x_i



Good feature x_j

(I) Filtering: univariate filtering

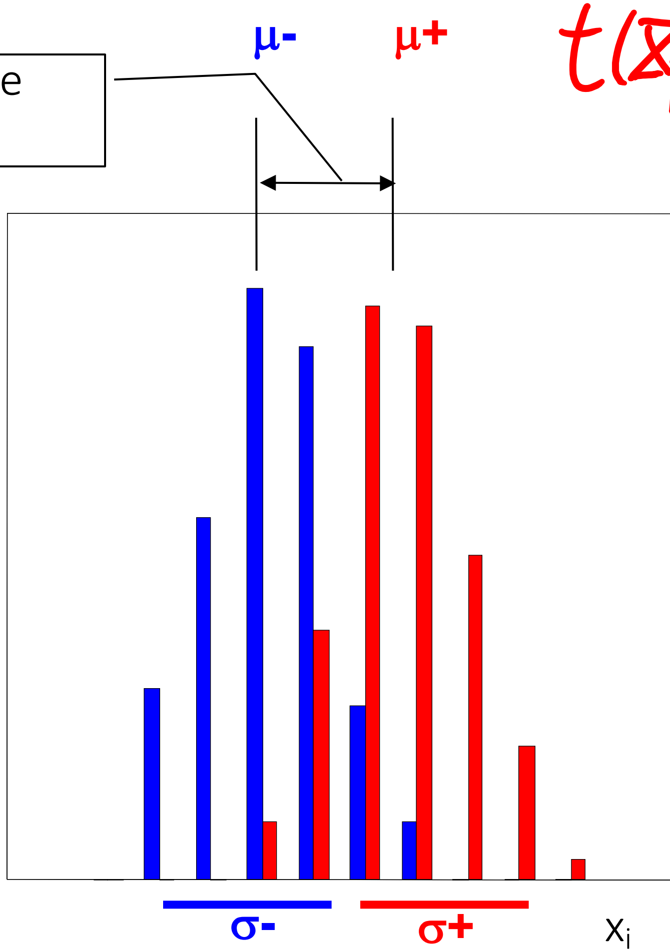
$X_1, X_2, \ldots, X_p$

$t(X_1, Y), t(X_2, Y), \ldots, t(X_p, Y)$

e.g. T-test

T-test

Is this distance significant?



- Assumption: Two Normally distributed classes with equal variance σ^2 unknown; estimated from data as σ^2_{within}.
- Null hypothesis H_0: $\mu+ = \mu-$
- T statistic:

If H_0 is true, then

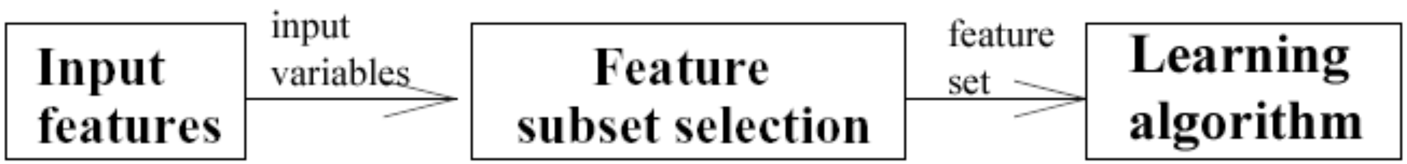
$$t = (\mu+ - \mu-)/(\sigma_{within}(1/|m^+| + 1/|m^-|)\perp(1/2))$$

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

From Dr. **Isabelle Guyon**

(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

Univariate selection may fail

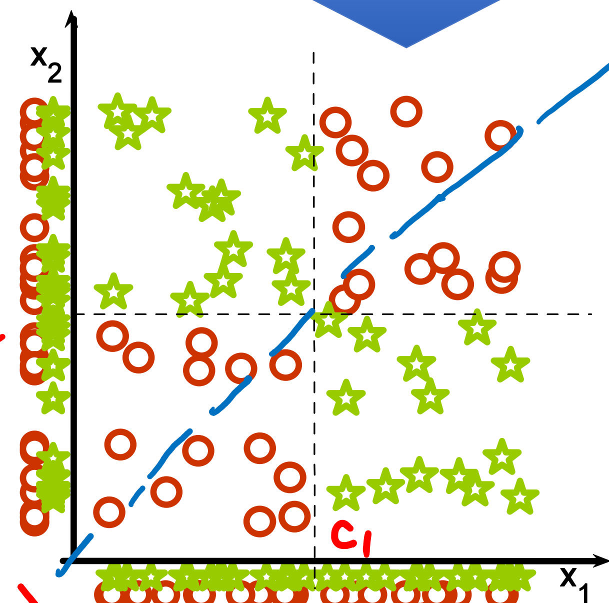
$t(X_1, Y)$

$t(X_2, Y)$

$(X_1, X_2, \ldots, X_p)$

$\vec{\theta}_{p\times 1} = (0/1, 0/1, \ldots, 0/1)$

$O(2^p \times \text{costEv}\theta)$



$(x_1 - c_1) \times (x_2 - c_2)$

on x_1 number line, two classes totally overlap!

Guyon-Elisseeff, JMLR 2004; Springer 2006

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

Sentiment Classification

e.g. amazon review text $x \rightarrow y$ 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)

    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

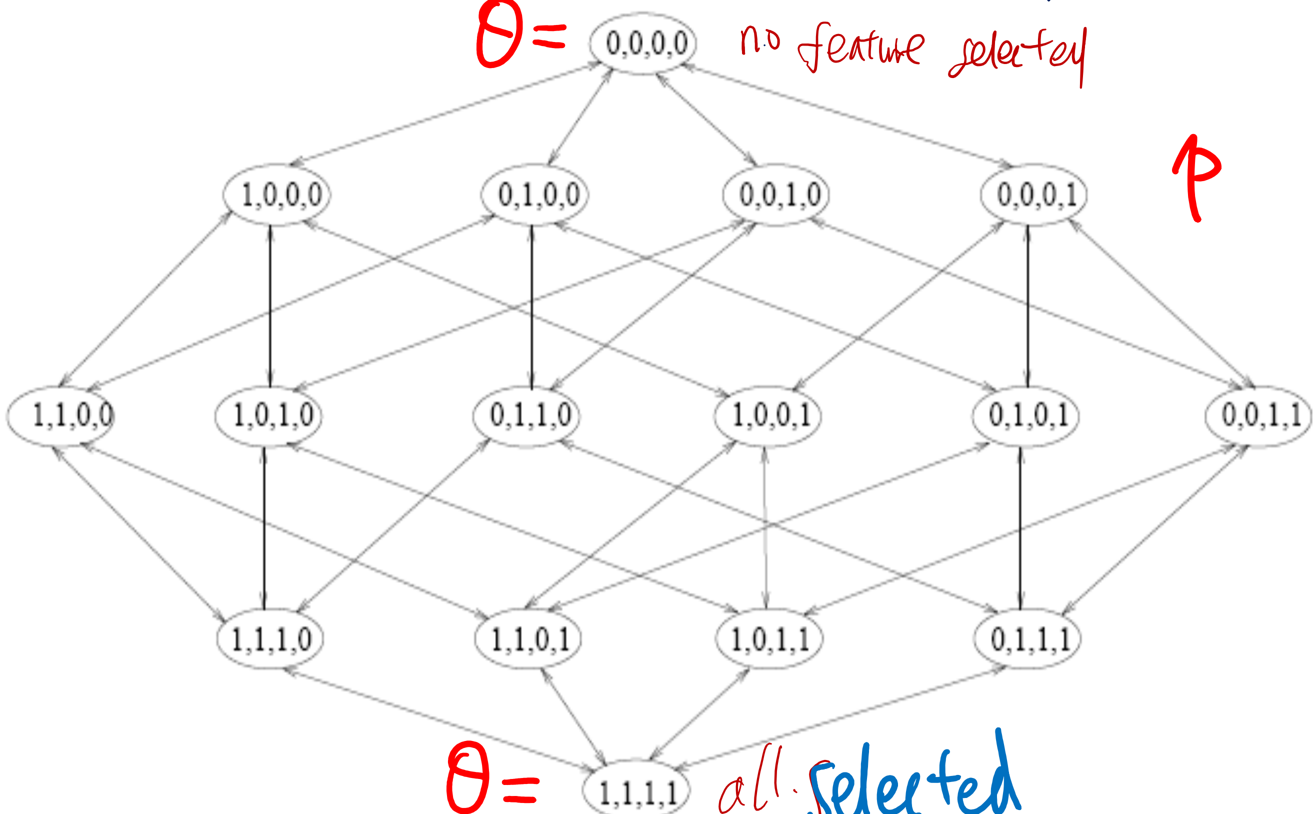
  => Good heuristics are needed!

[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 $\Rightarrow \theta = [0/1, 0/1, 0/1, \dots, 0/1]^T$

p×1 Vector

$\theta =$ no feature selected



p

$\theta =$ all selected

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		+	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.

Guyon-Elisseeff, JMLR 2004; Springer 2006

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.

(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.

$\theta \rightarrow (\cancel{X}_{n \times p'}, Y)$

- 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 ?

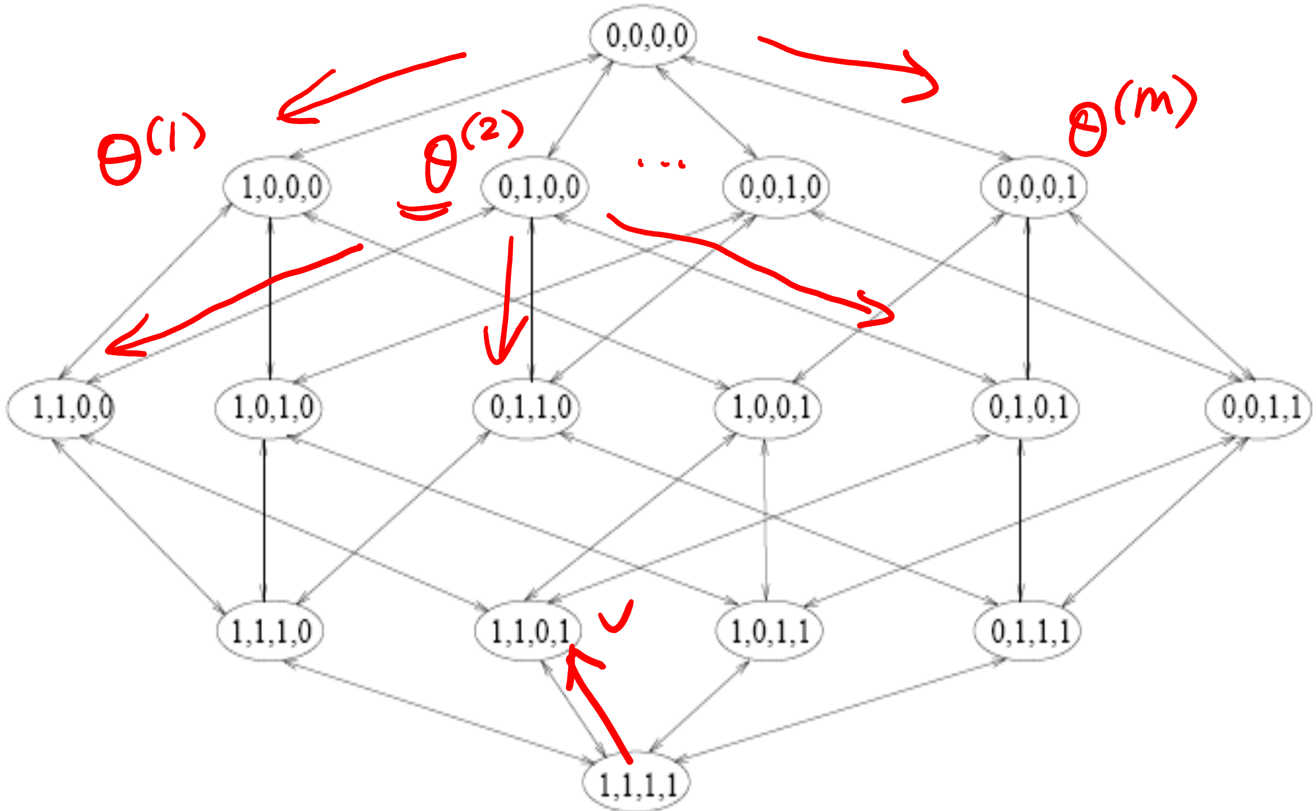
(b). Search: How to search the space of all feature subsets ?

- 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 ?



Step 1:

Step2:

⋮

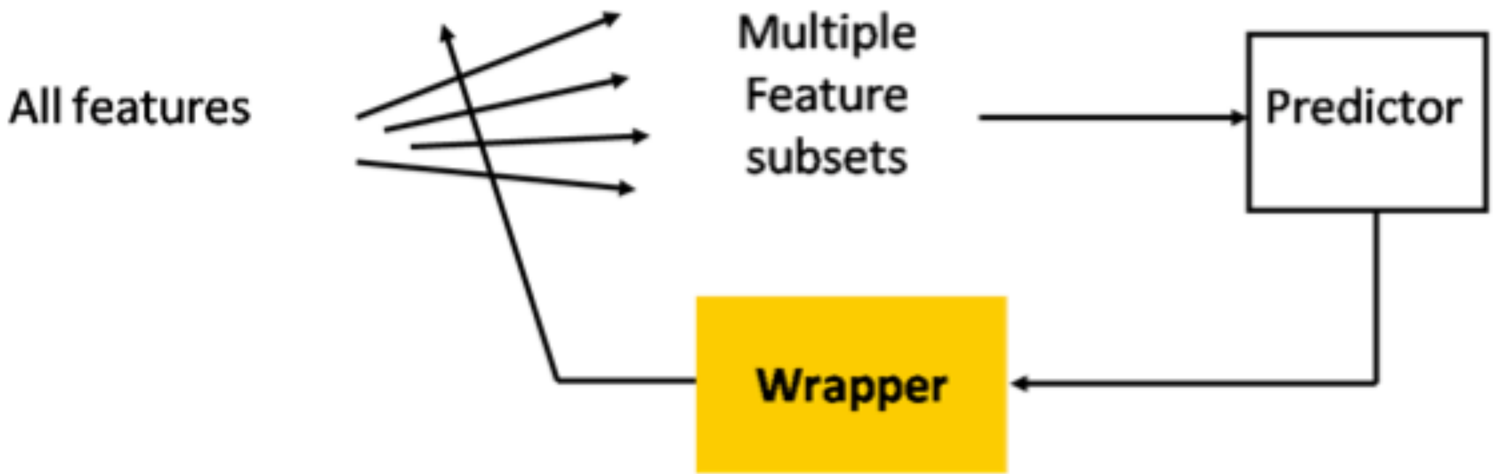
Step T

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

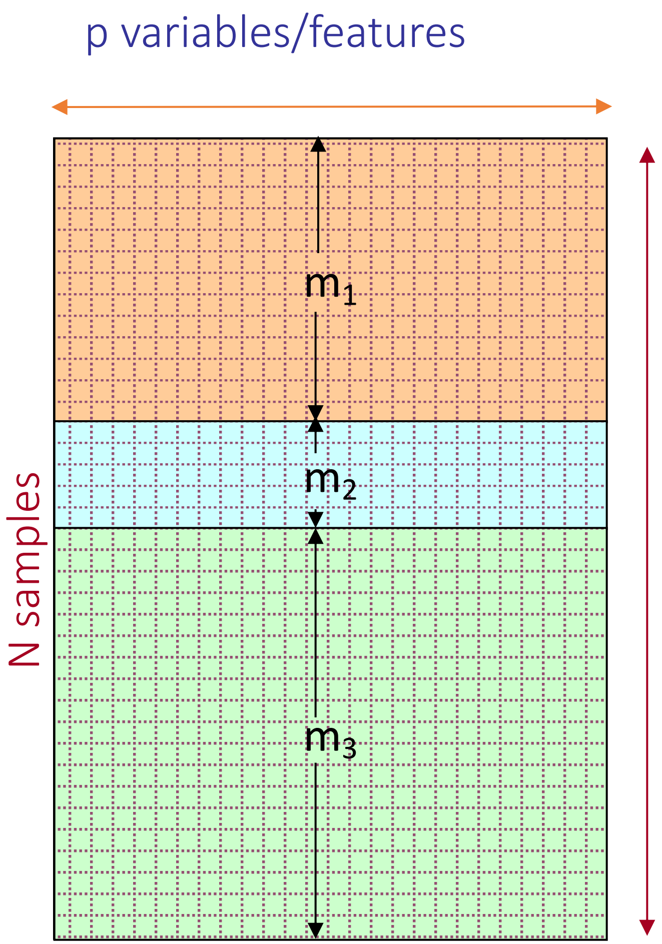
(a). Assessment: How to access multiple candidates of feature subsets

- Wrapper Methods

$\theta^{(i)} \to (\mathbf{X}^{n \times p^{(i)}}, Y) \to f^{(i)}$



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



Split data into 3 sets:

training, validation, and test set.

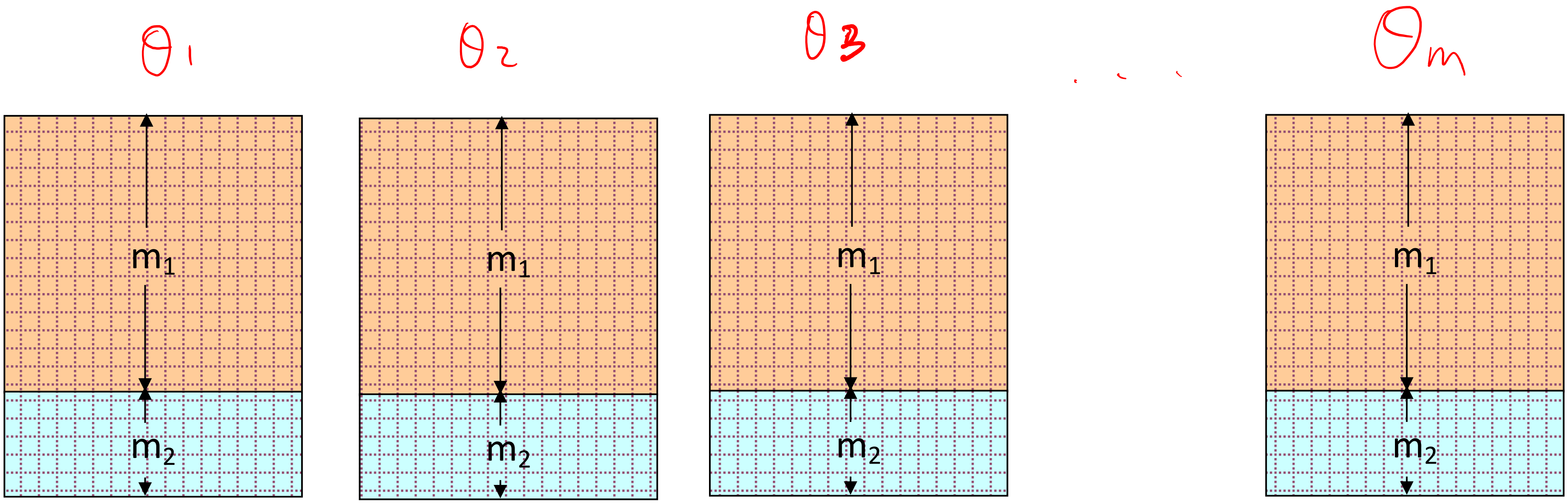
1) For each feature subset, train predictor on training data.

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.

From Dr. **Isabelle Guyon**

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



$\begin{cases} \text{train for } m \text{ times on train fold} \\ \text{~~test~~ validate for } m \text{ times on validation fold} \end{cases}$

$\Downarrow$

argmin valiError $\Rightarrow \theta^*$

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

validation data: $\underset{\{\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)}\}}{\operatorname{argmin}} \text{PredictLoss}(\beta^*_{\theta}) \Rightarrow \theta^* \Rightarrow (X_{m_1+m_2,p}, Y) \Rightarrow \beta^*_{\theta^*}$

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

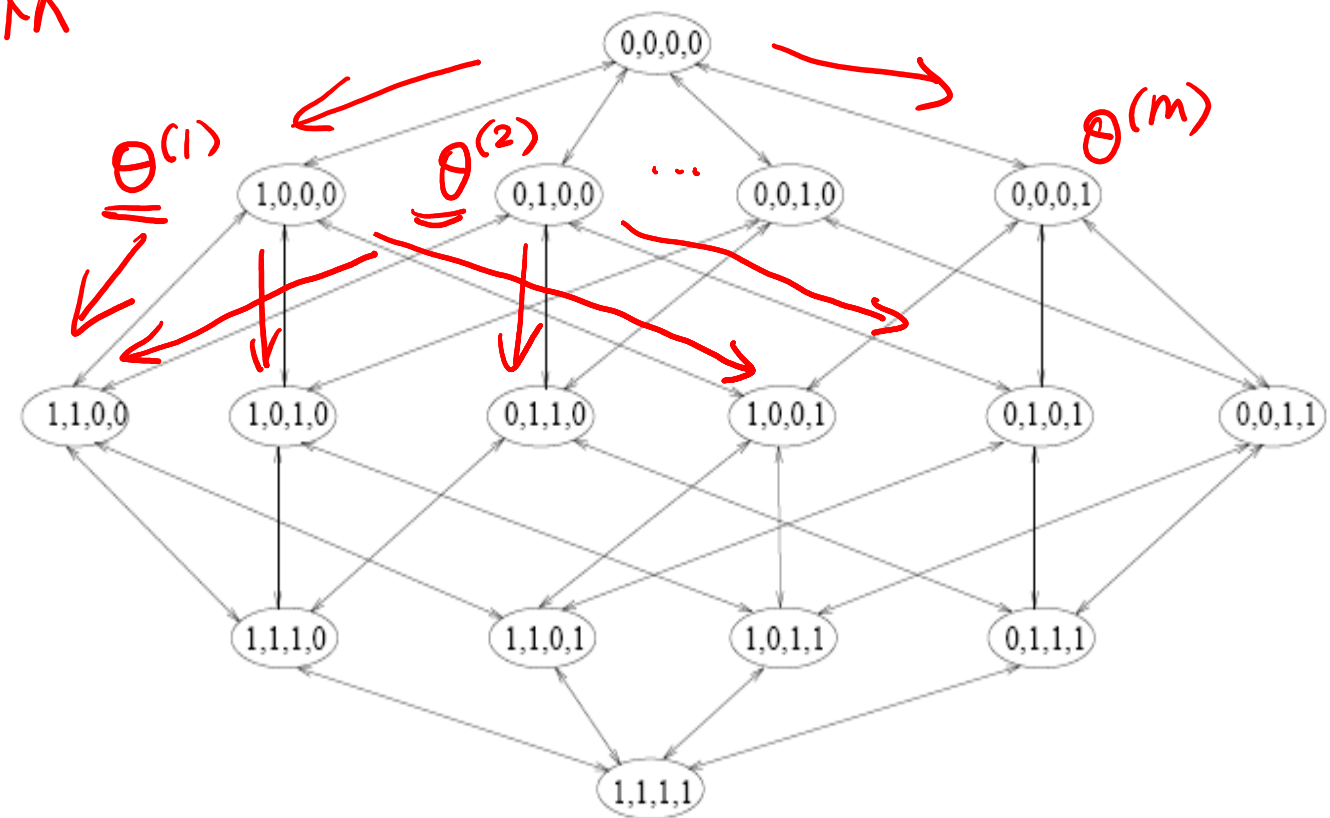
(b). Search: even more search strategies for selecting feature subset

- **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.

From Dr. **Isabelle Guyon**

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

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



Step 1:

Step2:

⋮

Step T

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

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

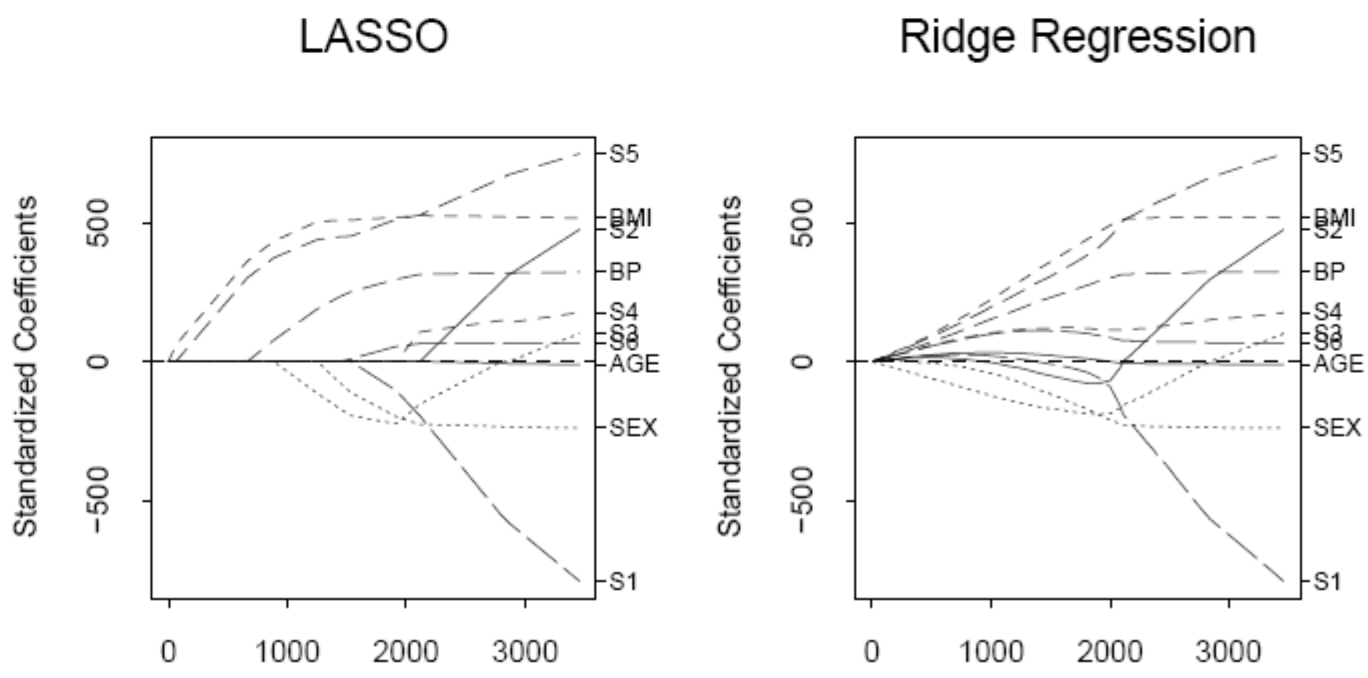
L1, L1+L2, Structured L1

- 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)

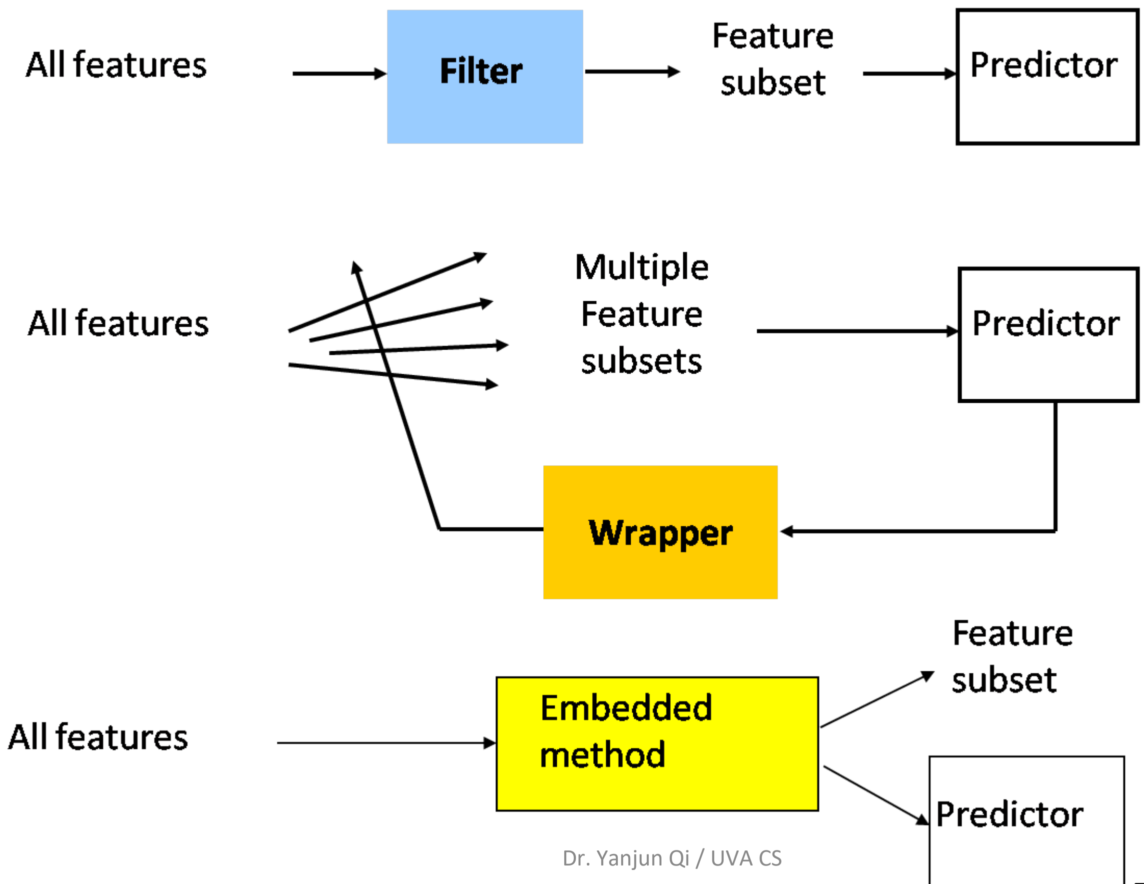
l_2 penalty: $y \sim Model(X\beta) + \lambda \sum \beta_i^2$ (ridge regression)



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**

Later: Dimensionality Reduction,

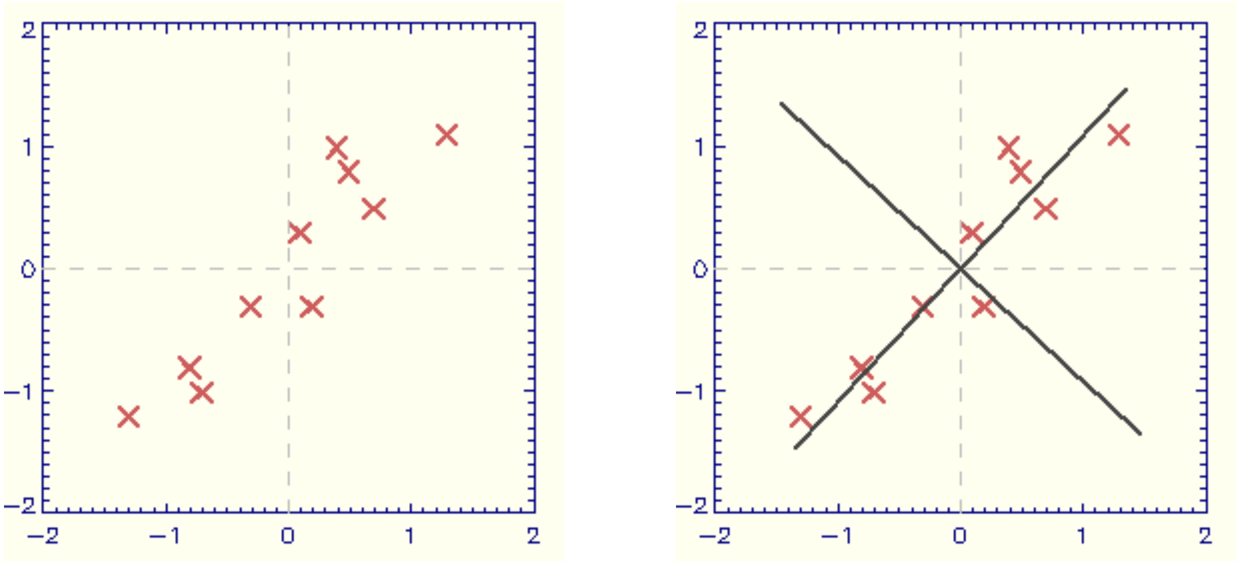
In the presence of many of features, select the most relevant subset of (weighted) combinations of features.

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

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

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

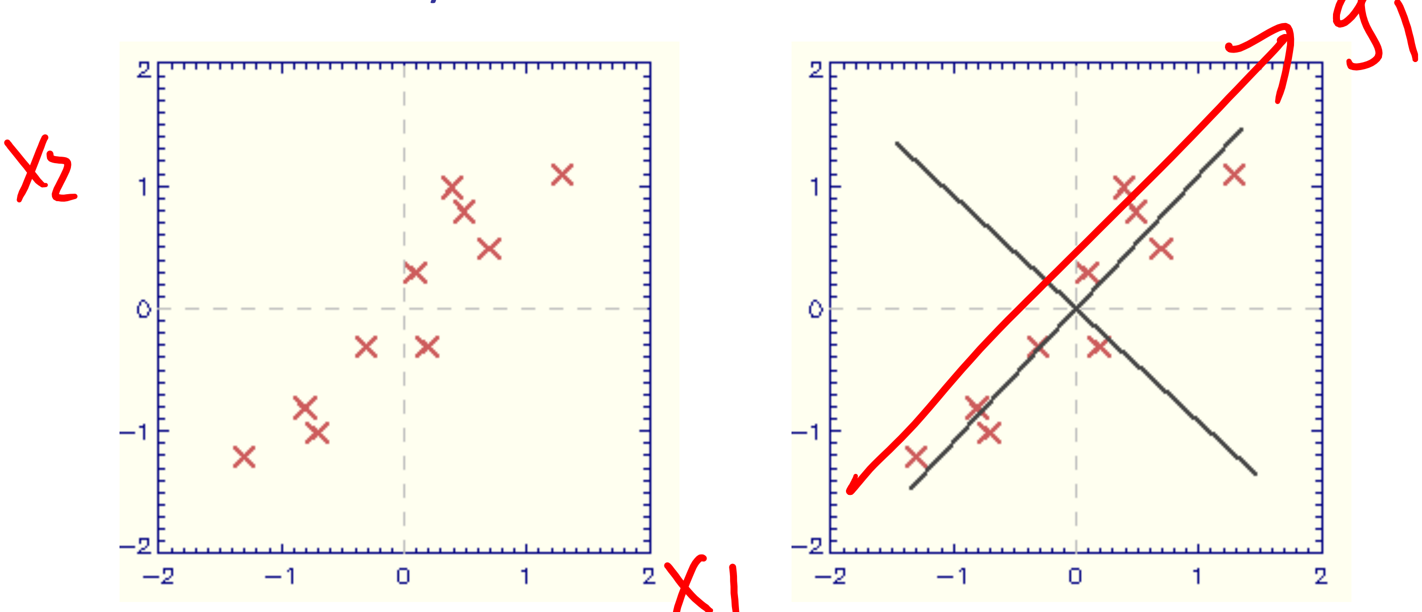
- **PCA** finds a *linear* mapping of dataset X to a dataset X' of lower dimensionality. The variance of X that is remained in X' is maximal.



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.



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

Task

↓

Representation

↓

Score Function

↓

Search/Optimization

↓

Models, Parameters

Selecting p’ original features from p Features

↓

Filter / Wrapper / Embedded

↓

Improve Classification / Regression

↓

Univariate / Combinatorial Search?

↓

Selected p’ features’ index

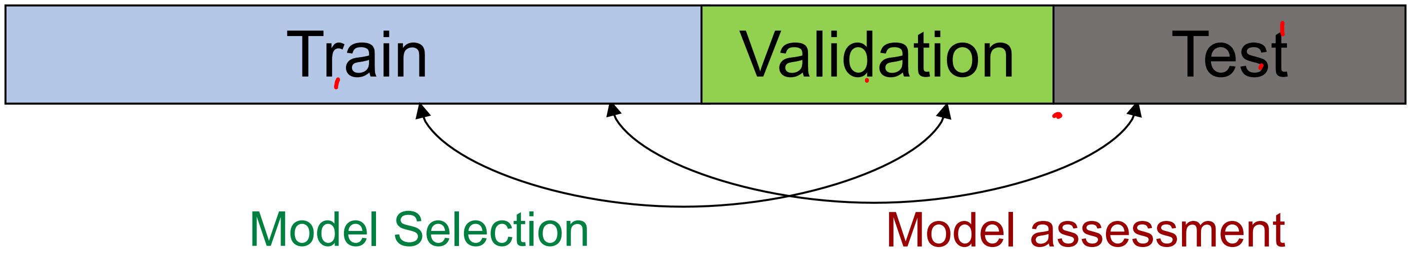
Model Selection and Assessment

- Model Selection
  - Estimating performances of different models to choose the best one

- Model Assessment
  - Having chosen a model, estimating the <u>prediction error</u> 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

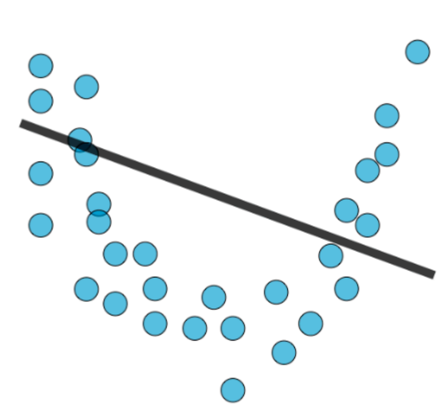
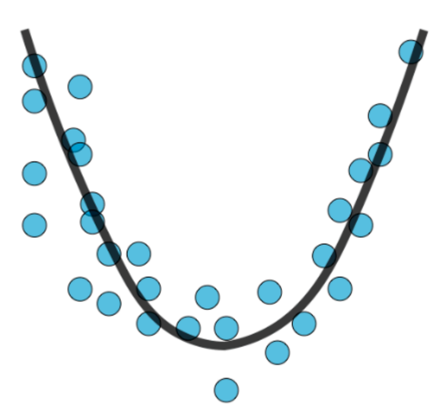
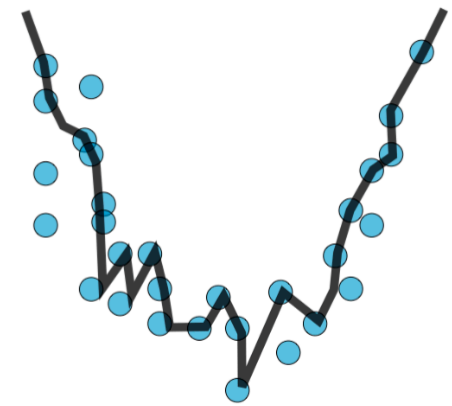
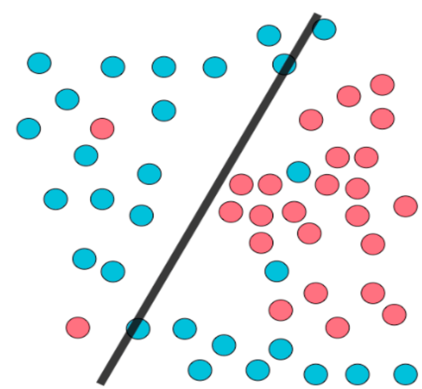
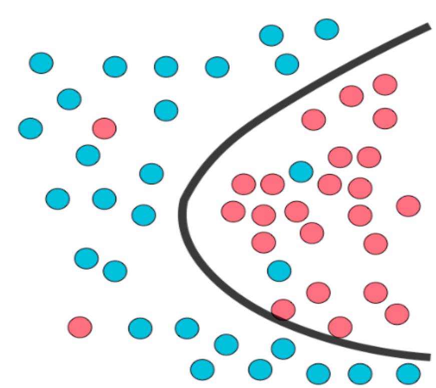
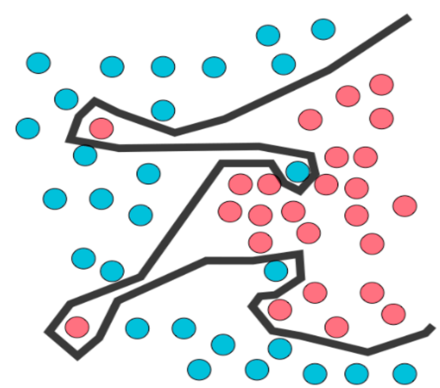
need to make assumptions that are able to generalize

- **Underfitting:** model is too "simple" to represent all the relevant characteristics
  - High bias and low variance
  - High training error and high test error
- **Overfitting:** model is too "complex" and fits irrelevant characteristics (noise) in the data
  - Low bias and high variance
  - Low training error and high test error

A Gentle Touch of Bias - Variance Tradeoff

(More details … Later)

Slide credit: L. Lazebnik

	Underfitting	Just right	Overfitting
Symptoms	- High training error - Training error close to test error - High bias	- Training error slightly lower than test error	- Low training error - Training error much lower than test error - High variance
Regression			
Classification			
Remedies	- Complexify model - Add more features - Train longer		- Regularize - Get more data - Feature selection

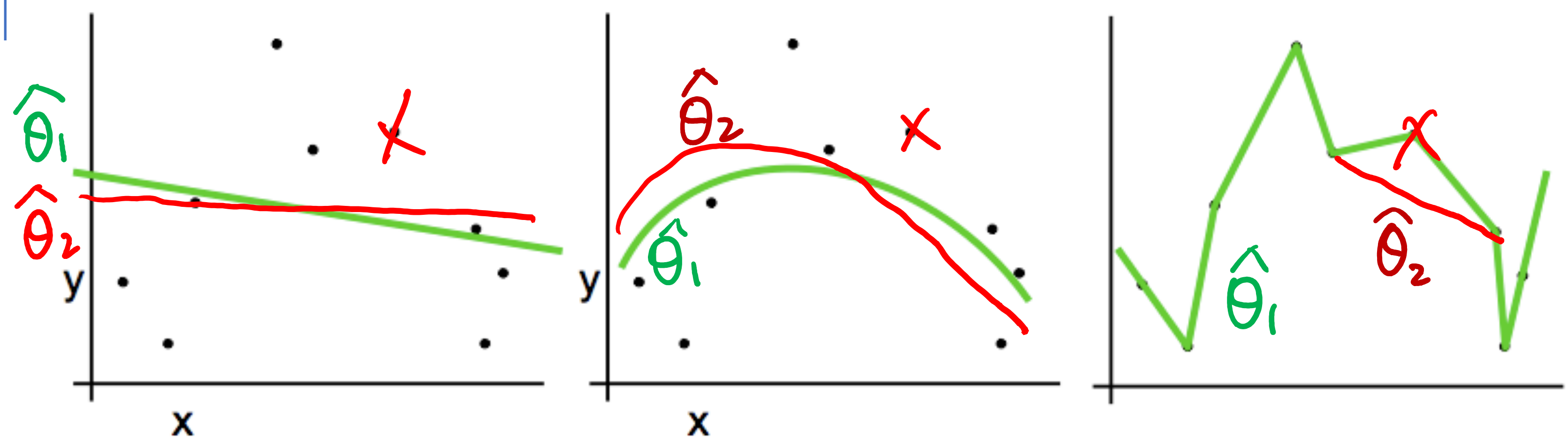
Credit: Stanford Machine Learning

need to make assumptions that are able to generalize

- Components
  - **Bias:** how much the average model over all training sets differ from the true model?
    - Error due to inaccurate assumptions/simplifications made by the model
  - **Variance:** how much models estimated from different training sets differ from each other

Slide credit: L. Lazebnik

Randomness of Train Set
=> Variance of Models, e.g.,

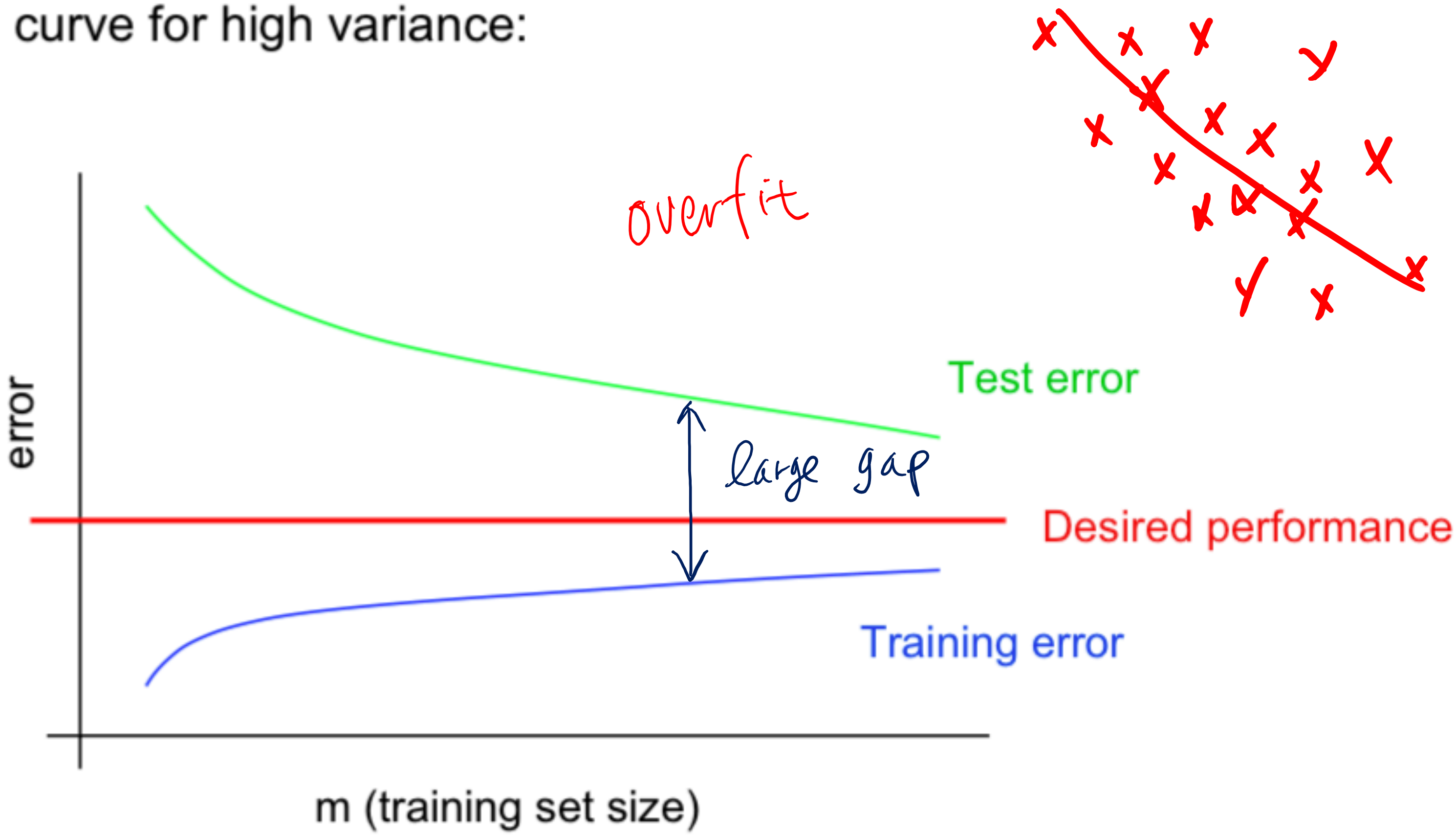


e.g. removing one training sample

model complexity ↑ ⇒ model variance ↑

(1) Overfitting / High variance / Model too Complex

Typical learning curve for high variance:



- Test error still decreasing as m increases. Suggests larger training set will help.
- Large gap between training and test error.
- Low training error and high test error

Slide credit: A. Ng

How to reduce Model High Variance?

- Choose a simpler classifier
  - More Bias

- Regularize the parameters
  - More Bias

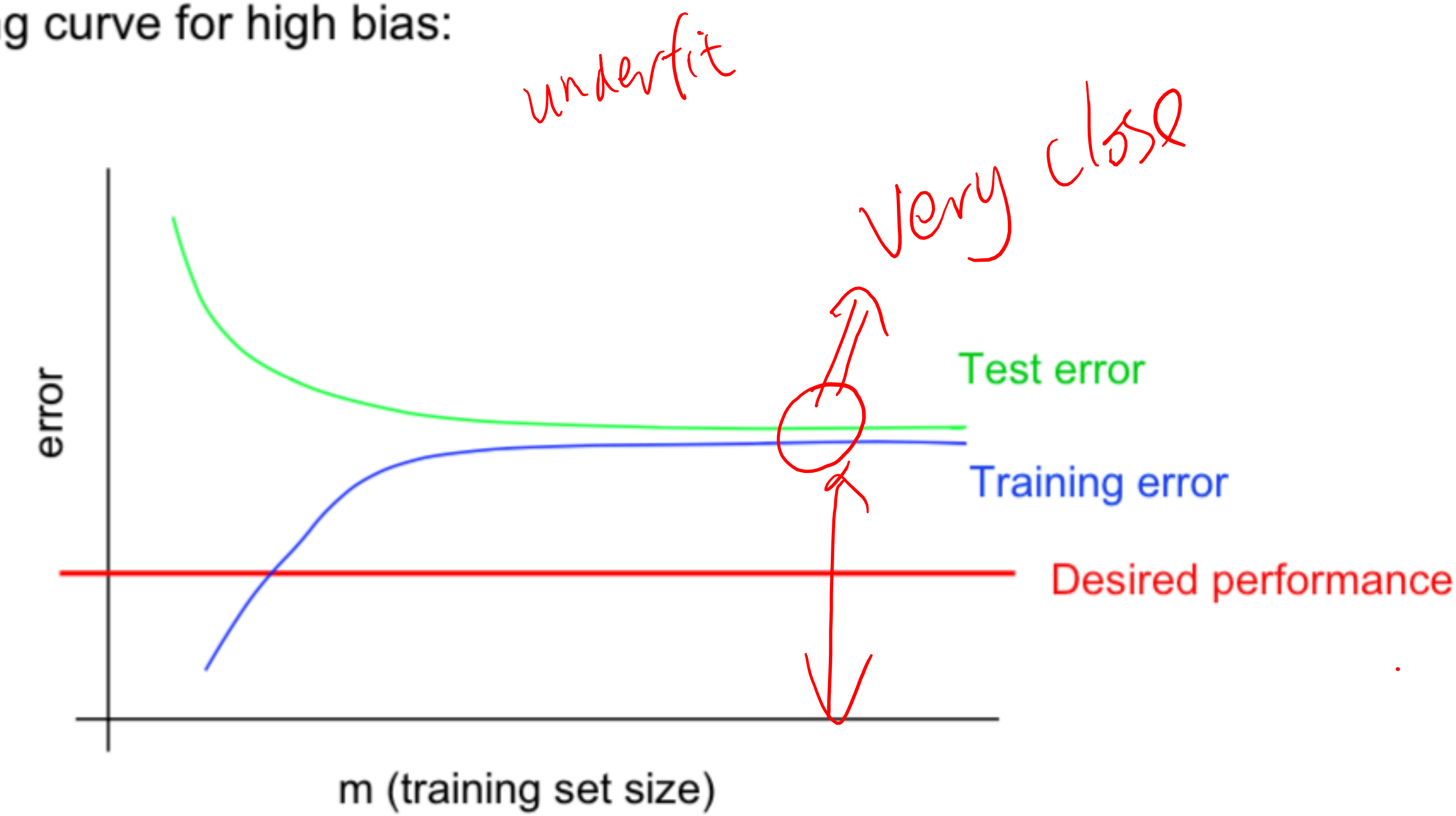
- Get more training data

- Try smaller set of features
  - More Bias

Slide credit: D. Hoiem

(2) Underfitting / High bias / Model too Simple

Typical learning curve for high bias:



- Even training error is unacceptably high.
- Small gap between training and test error.

High training error and high test error

Slide credit: A. Ng

How to reduce Model High Bias ?

- E.g.

  - Get additional features

  - Try more complex learner

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**