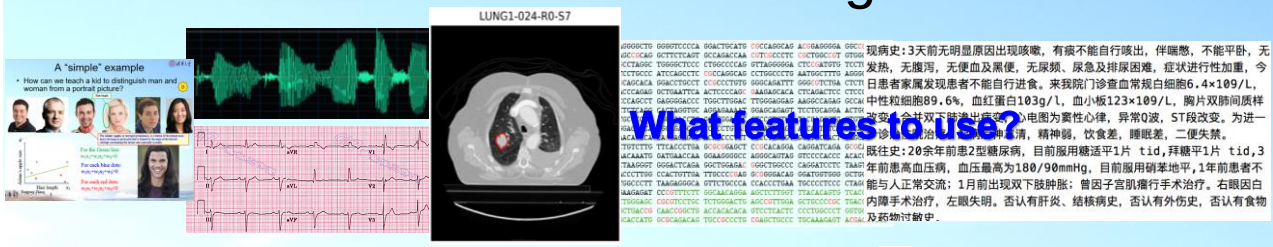
Course ID 80250993 Dates: 9/16-12/23/2021 @ I-205
Tencent/VooV Meeting 521 4678 6257 (passwd 1205)

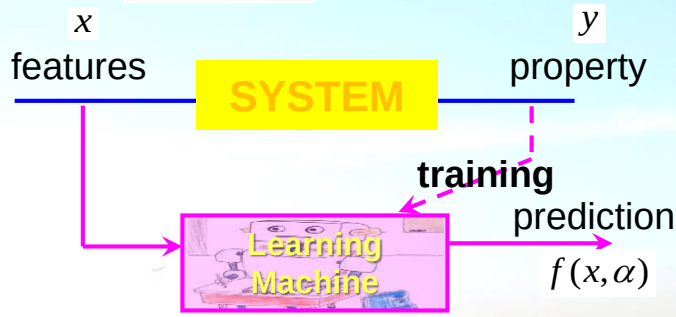
Chapter 8
Feature Selection and Extraction for Classification

Xuegong Zhang
Oct. 28, 2018

8.1
Introduction

Machine Learning





Quick review

- The basic elements of a (supervised) learning machine
  - A specific task of predicting something (labels, values)
    - A well-defined task:

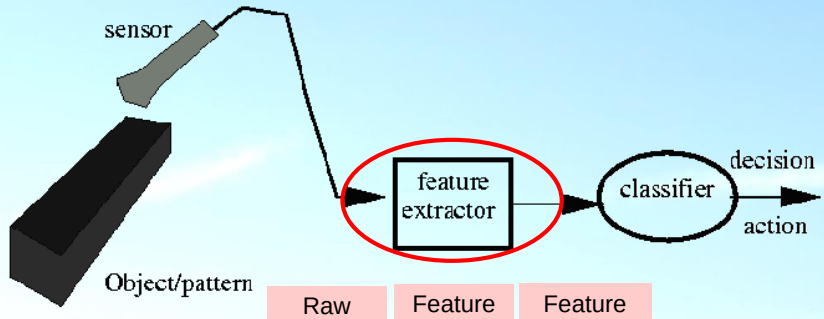
      Given features to be used for the prediction
    - A more general task:

      Not sure about features to be used for the prediction
  - Training data
  - A proper model of the machine
    - Perceptron, Linear regression, Logistic regression, MLP, SVM, …
  - The objective of learning (for the model)
  - The learning algorithm (to optimize the objective function)

Machine Learning Methods

Machine Learning Methods

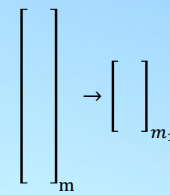
An ML System

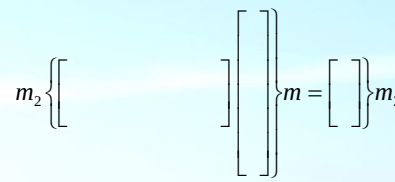


Raw features | Feature selection | Feature transform

"Feature Engineering"

- Terminology
  - Feature acquisition → raw/initial features
    - Real-value features
    - Nominal features
  - Feature selection
    - Select m_1 features from m original features ($m_1 < m$)
  - Feature extraction / feature transformation
    - Obtain m_2 features from m original features (usually $m_2 < m$)
- Why?
  - Noisy original features
  - Computational considerations
  - Finding key features





8.2
Metrics for class-separability of feature

What criteria shall we use for the selection?

- Error rate?
  - That's what we want
  - But how can we know before we do it?
- Metrics J_{ij} on separability of classes i and j with the features
  - They should be
    - Be correlated with error rate, e.g., $J_{ij} \uparrow \rightarrow P_e \downarrow$
    - $J_{ij} \geq 0,\ J_{ii} = 0,\ J_{ij} = J_{ji}$
    - Addable for independent features
    - Not decreasing when adding new features

8.1.1 Metrics based on distance

- Average distance between classes

$$J_D = \frac{1}{2}\sum_{i=1}^{c} P_i \sum_{j=1}^{c} P_j \frac{1}{n_i n_j} \sum_{k=1}^{n_i}\sum_{l=1}^{n_j} \delta\left(\boldsymbol{x}_k^{(i)}, \boldsymbol{x}_l^{(j)}\right)$$

where $\boldsymbol{x}_k^{(i)} \in \omega_i, k = 1,\cdots,n_i$, $\boldsymbol{x}_l^{(j)} \in \omega_j, l = 1,\cdots,n_j$, P_i, P_j are the prior probability of the classes, and $\delta(\boldsymbol{x}_k^{(i)}, \boldsymbol{x}_l^{(j)})$ are the distance between two samples, e.g.,

$$\delta\left(\boldsymbol{x}_k^{(i)}, \boldsymbol{x}_l^{(j)}\right) = \left(\boldsymbol{x}_k^{(i)} - \boldsymbol{x}_l^{(j)}\right)^T \left(\boldsymbol{x}_k^{(i)} - \boldsymbol{x}_l^{(j)}\right).$$

- Calculation using scatter matrixes

$$J_D = tr(\tilde{S}_w + \tilde{S}_b)$$

Between-class scatter matrix

$$\tilde{S}_b = \sum_{i=1}^{c} P_i(\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^T$$

Within-class scatter matrix

$$\tilde{S}_w = \sum_{i=1}^{c} P_i \frac{1}{n_i} \sum_{k=1}^{n_i} \left(\boldsymbol{x}_k^{(i)} - \boldsymbol{m}\right)\left(\boldsymbol{x}_i^{(i)} - \boldsymbol{m}\right)^T$$

Class mean $\boldsymbol{m}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} \boldsymbol{x}_k^{(i)}$

Total mean $\boldsymbol{m} = \sum_{i=1}^{c} P_i \boldsymbol{m}_i$

Ref. Fisher's criterion

$$J_F = \frac{(\tilde{m}_1 - \tilde{m}_2)^2}{\tilde{S}_1^2 + \tilde{S}_2^2}$$

- Other types of distance-based criteria

$$J_1 = tr(S_w + S_b)$$

$$J_2 = tr(S_w^{-1} S_b)$$

$$J_3 = \ln \frac{|S_b|}{|S_w|}$$

$$J_4 = \frac{trS_b}{trS_w}$$

$$J_5 = \frac{|S_b - S_w|}{|S_w|}$$

Note:

- Intuitive and easy to compute
- No explicit relation with error rate
- Better when covariances are not very different between classes

Some popular distance measurements

- Minkovski Metric (of order s): $\delta(\boldsymbol{x}_k, \boldsymbol{x}_l) = \left[\sum_{i=1}^{d} |x_{ki} - x_{li}|^s\right]^{\frac{1}{s}}$
- Euclidean Distance: $\delta_E(\boldsymbol{x}_k, \boldsymbol{x}_l) = [(\boldsymbol{x}_k - \boldsymbol{x}_l)^{\mathrm{T}}(\boldsymbol{x}_k - \boldsymbol{x}_l)]^{\frac{1}{2}}$
- City-Block Distance: $\delta(\boldsymbol{x}_k, \boldsymbol{x}_l) = \sum_{i=1}^{d} |x_{ki} - x_{li}|$
- Chobychev Distance: $\delta(\boldsymbol{x}_k, \boldsymbol{x}_l) = \max_i |x_{ki} - x_{li}|$
- Squared Distance: $\delta(\boldsymbol{x}_k, \boldsymbol{x}_l) = (\boldsymbol{x}_k - \boldsymbol{x}_l)^{\mathrm{T}} Q (\boldsymbol{x}_k - \boldsymbol{x}_l)$
- Nonlinear distances, e.g., $\delta(\boldsymbol{x}_k, \boldsymbol{x}_l) = \begin{cases} H & \text{if } \delta_E(\boldsymbol{x}_k, \boldsymbol{x}_l) \geq T \\ 0 & \text{if } \delta_E(\boldsymbol{x}_k, \boldsymbol{x}_l) < T \end{cases}$

8.1.2 Metrics based on distributions

- To measure the overlapping of two distributions

$$J_p(\cdot) = \int g[p(\boldsymbol{x}|\omega_1), p(\boldsymbol{x}|\omega_2), P_1, P_2]\, d\boldsymbol{x}$$

- Example:
  - Bhattacharyya distance

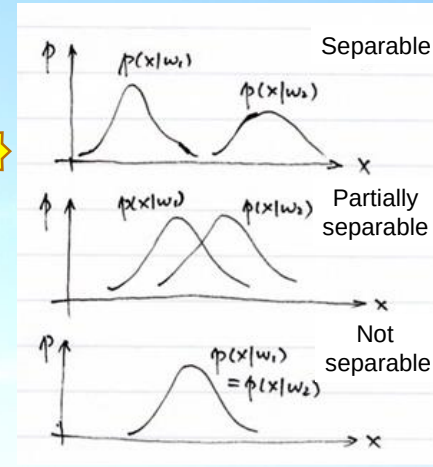
$$J_B = -\ln \int [p(x|\omega_1)p(x|\omega_2)]^{\frac{1}{2}} dx$$

  - Chernoff distance

$$J_c = -\ln \int p^s(x|\omega_1)p^{1-s}(x|\omega_2)\, dx$$

  - Divergence

$$J_D = \int_x [p(x|\omega_1) - p(x|\omega_2)] \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx$$



- Similarly, we can consider difference between the density function of one class with that of the whole sample.
  - Since $p(x, \omega_i) = p(x|\omega_i)P(\omega_i)$, if $p(x) = p(x|\omega_i)$, feature x is independent with class ω_i and cannot provide information for the classification.
- Probabilistic correlation criterion:

$$J_i = \int g(p(x|\omega_i), p(x), P(\omega_i))\, dx$$

Some probability distance measures

- Chernoff

$$J_c = -\ln \int p^s(x|\omega_1)p^{1-s}(x|\omega_2)\,dx \;,\; I_c = -\ln \int p^s(x|\omega_i)p^{1-s}(x)\,dx$$

- Bhattacharyya

$$J_B = -\ln \int \sqrt{p(x|\omega_1)p(x|\omega_2)}\,dx \;,\; I_B = -\ln \int \sqrt{p(x|\omega_i)p(x)}\,dx$$

- Matusita

$$J_M = \left[\int \left(\sqrt{p(x|\omega_1)} - \sqrt{p(x|\omega_2)}\right)^2 dx\right]^{\frac{1}{2}} \;,\; I_M = \left[\int \left(\sqrt{p(x|\omega_i)} - \sqrt{p(x)}\right)^2 dx\right]^{\frac{1}{2}}$$

- Divergence

$$J_D = \int [p(x|\omega_1) - p(x|\omega_2)] \ln \frac{p(x|\omega_1)}{p(x|\omega_2)} dx \;,\; I_D = \int [p(x|\omega_i) - p(x)] \ln \frac{p(x|\omega_i)}{p(x)} dx$$

- Patrick-Fisher

$$J_P = \left[\int (p(x|\omega_1)P_1 - p(x|\omega_2)P_2)^2\,dx\right]^{\frac{1}{2}} \;,\; I_P = \left[\int (p(x|\omega_i)P_i - p(x))^2\,dx\right]^{\frac{1}{2}}$$

- Lissack-Fu

$$J_L = \int |p(x|\omega_1)P_1 - p(x|\omega_2)P_2|^{\alpha} p(x)^{1-\alpha}\,dx$$

- Kolmogorov

$$J_K = \int |p(x|\omega_1)P_1 - p(x|\omega_2)P_2|\,dx$$

8.1.3 Metrics based on information theory

- Entropy
  - Information Entropy (Shannon, 1949)

$$I = -(P_1 \log_2 P_1 + P_2 \log_2 P_2 + \cdots + P_k \log_2 P_k) = -\sum_{i=1}^{k} P_i \log_2 P_i$$

  - Shannon Entropy for feature selection

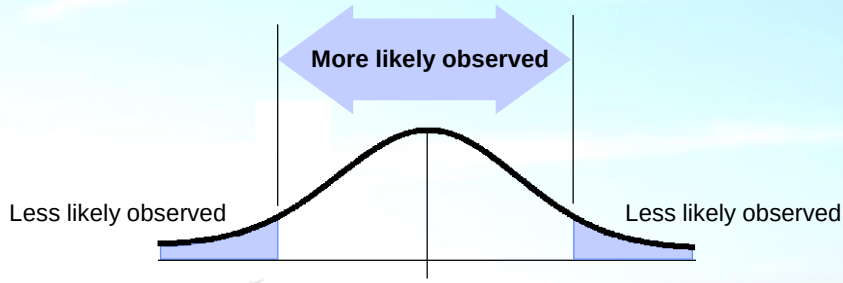
$$H = -\sum_{i=1}^{c} P(\omega_i|x) \log_2 P\,(\omega_i|x)$$

- Mutual Information (MI)

$$I(X;Y) = \int_y \int_x p_{(X,Y)}(x,y) \log \left(\frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)}\right) dxdy$$

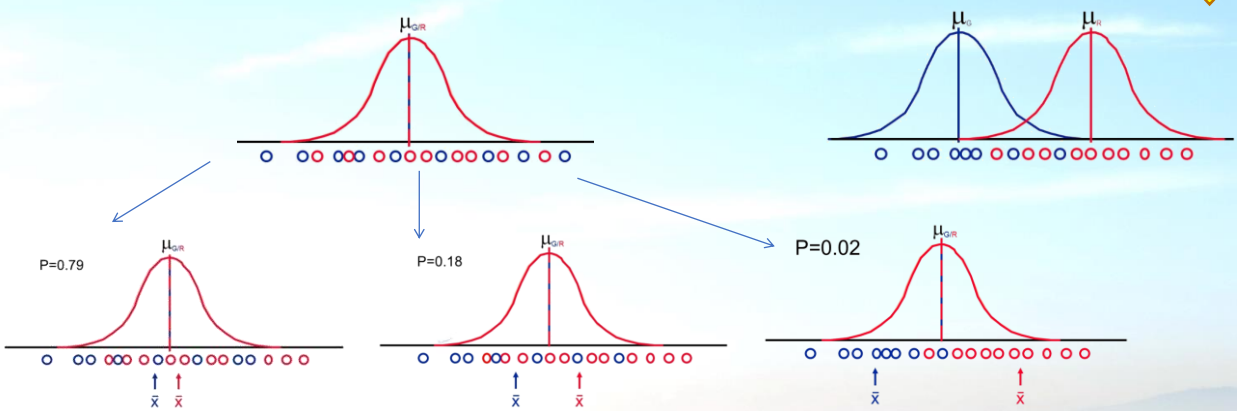
8.1.4 Using statistical tests to measure separability

- Hypothesis testing (Statistical Inference)
  - Null hypothesis vs. Alternative hypothesis
  - Basic idea
    - Using a test statistic to measure likelihood of observing the data under the null hypothesis



- Evaluating a feature for classification
  - Comparing two samples (two sample sets)
  - Null hypothesis: two samples from the same distribution
  - Alternative: they are from two distributions

→ Test for difference in means



- The t-test

sample 1: $X_1, \cdots, X_n \quad X \sim N(\mu_X, \sigma^2)$

sample 2: $Y_1, \cdots, Y_m \quad Y \sim N(\mu_Y, \sigma^2)$

- pooled sample variance

$$s_P^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{m+n-2}$$

- the null hypothesis

$$H_0: \ \mu_X = \mu_Y$$

- alternative hypotheses

two-sided $H_1: \ \mu_X \neq \mu_Y$

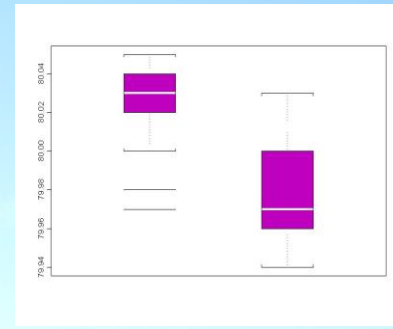
one-sided $H_2: \ \mu_X > \mu_Y$

$H_3: \ \mu_X < \mu_Y$

- t-statistic

$$t = \frac{\bar{X} - \bar{Y}}{s_{\bar{X}-\bar{Y}}} = \frac{\bar{X} - \bar{Y}}{s_P\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

A	B
79.98	80.02
80.04	79.94
80.02	79.98
80.04	79.97
80.03	79.97
80.03	80.03
80.04	79.95
79.97	79.97
80.05	
80.03	
80.02	
80.00	
80.02	



Box-plot

- Permutation test for means
  - No assumption for the underlying distribution
  - Necessary when sample size small
- Idea
  - Q: Whether measurements in A and B are exchangeable?
    - 13+8=21 measurements, $\sim 2 \times 10^5$ ways of assigning the two classes
    - Is the observed assignment unusual, in the sense that the two means are significantly different?
  - A: Check distribution of difference under null hypothesis using permutation
    - Choose a random sample of all possible permutations (e.g, 1000)
    - Make a histogram of the resulting values $mean_A - mean_B$
    - Check the position of the observed $mean_A - mean_B$ in the histogram for a *p-value*

A	B
79.98	80.02
80.04	79.94
80.02	79.98
80.04	79.97
80.03	79.97
80.03	80.03
80.04	79.95
79.97	79.97
80.05	
80.03	
80.02	
80.00	
80.02	

10-second break

8.3
Searching for the best features

- Task: Finding the best d features from the total of D candidates
  - The space for searching

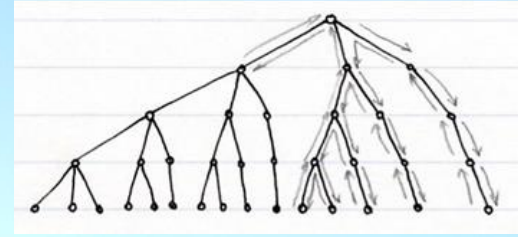
$$C_D^d = \frac{D!}{(D-d)!d!}$$

D=100, d=2, C = 4950
D=100, d=3, C = 161700
D=100, d=10, C = 1.73103e+13
D=100, d=50, C = 1.00891e+29
D=1000, d=2, C = 499500
D=10000, d=2, C = 4.9995e+07

- Challenge:
  - Optimal combination (not the best single features)
- Strategies
  - Enumerating (optimal), Heuristic (sub-optimal)
  - Bottom-up, top-down

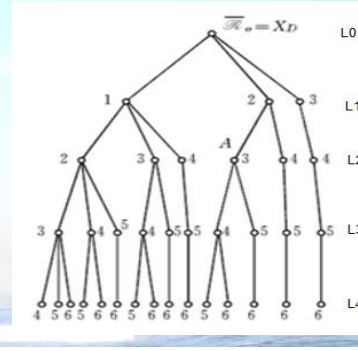
8.2.1 Optimal searching

- Enumerating
  - Enumerate all possible combinations to find the best
  - Suitable when d or $D-d$ is small



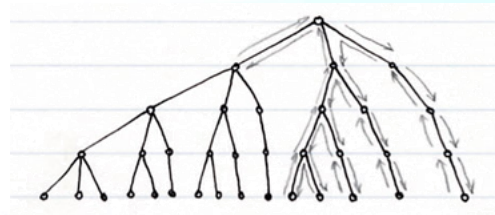
- Branch-and-Bound Algorithm
  - Organize all possible combinations as an ordered tree, in such a way that searching can be completed without browsing the whole tree.
  - Top-down searching with tracing back
  - Requirement: The criterion does not increase when reducing features.
    - $J_1,\ J_2,\ J_3,\ J_4,\ J_5,\ J_C,\ J_B,\ J_D$

- Key rules of BB searching
  - Root: level 0, all features included
  - Drop 1 feature at each node. Leaf nodes are selected combinations
  - Avoid identical combinations throughout the tree
  - Keep record of the current max as the bound (B), $B = 0$ at beginning
  - Put features least likely to be dropped at the left most of each layer (e.g., dropping them resulting in smallest metric)
  - Searching from the right-hand branch
  - Features in the left nodes of the same layer will not be dropped at the current node or its downstream nodes
  - Go back when meeting a node with $J < B$
  - Update B when reaching a leaf node, then go back
  - Put back the dropped features when going back
  - When the searching stops at the root, the leaf node that last updated B is the optimal combination.



- BB Searching
  - Optimal when the criterion is monotonic
  - Saving the most when $d \approx \frac{1}{2}D$
  - Example

D	*d*	Enumerate	BB
12	4	$\binom{12}{4} = 495$	42
24	12	$\binom{24}{12} = 2{,}704{,}156$	13,369



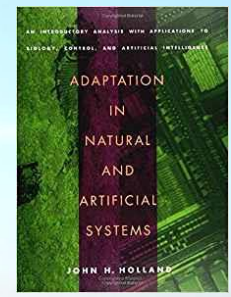
8.2.2 Sub-optimal Searching

- Combination of top single features
- SFS (Sequential Forward Selection)
  - Bottom-up, adding one feature in each step
- Generalized SFS
  - Bottom-up, adding l features in one step
- SBS (Sequential Backward Selection)
  - Top-down, eliminating one feature in each step
- Generalized SBS
  - Top-down, eliminating r features in each step
- L-R
  - Adding l features and then eliminating r features, ...
- ...

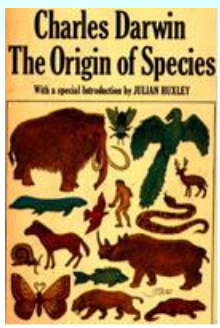
10-second break

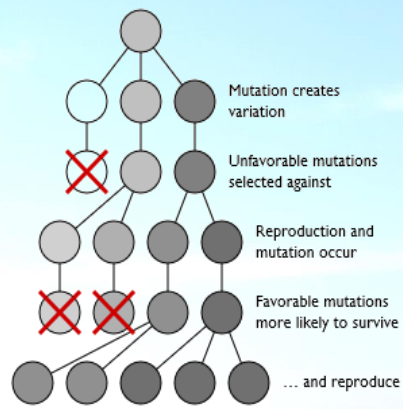
8.2.3 Genetic Algorithm for Feature Selection

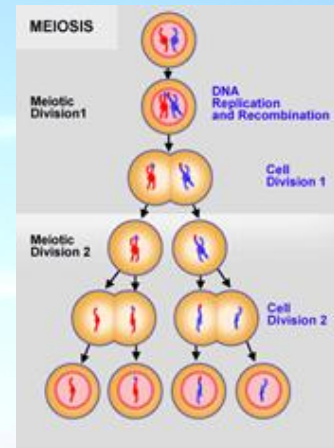
- GA (Genetic Algorithm)
  - Stochastic Searching Algorithm
  - Evolutionary Computing
  - "Genetic Algorithms are good at taking large, potentially huge search spaces and navigating them, looking for optimal combinations of things, solutions you might not otherwise find in a lifetime." ---- Salvatore Mangano
  - 1960-1970's, John H. Holland, a Ph.D. student in U. Michigan
  - "*teaching computers how to have sex*"
  - John Henry Holland, *Adaptation in Natural and Artificial Systems*, MIT Press, 1975



- Two Key Factors in Evolution
  - Variation (Mutation)
    - Point Mutations
    - Recombination
  - Selection







Genetic Algorithm:

Using the mechanism of variation-selection for optimization.

Task: Selecting d features from D using GA

- Encoding the problem as a chromosome
  - E.g., (0 1 0 0 1 … 0 1 1 0 1 0 1) of D binary elements
  - Searching space: C_D^d possible combinations
  - Task: finding **the most fit** *"individual"* from the *"population"*
- Fitness function $f(m)$
  - Certain metrics of separability
- Basic Operation: Inheritance, variation, selection
  - Point mutation
  - Recombination or crossover
  - Reversion, transposition, duplication, …
  - Selection: $p(f(m))$

- Genetic Algorithm

① Initialization at t =0

Randomly generate a population $M(0)$ of L chromosomes

② Compute the fitness f (**m**) of each chromosome in the current population $M(t)$

③ Sample the population according to selection probability $p(f$ (**m**)), and use the selected population to reproduce (with probability of variation) the next generation of chromosomes $M(t+1)$

④ Return to ② until the stopping criterion met, then output the fittest chromosome as the optimal solution

Stopping criterion usually set as a threshold of the fitness.

10-second break

8.4
Filtering, Wrapper and Embedded Methods for Feature Selection

Three strategies for feature selection

- Filtering methods (Two-Step Procedure)
  - Feature selection (with some stand-alone criteria or metrics)
  - Classification (using the selected features)

D

- Wrapper methods (Recursive Procedure)
  - Classification (with all features)
  - Feature selection (according to classification performance)
  - Classification (using the selected features)

- Embedded methods
  - Feature selection is embedded in the classification algorithm, e.g., Lasso

$$\min_{\beta} \frac{1}{l}\sum_{i=1}^{l}\left(y_j - \boldsymbol{\beta}^T \boldsymbol{x}_j\right)^2 + \lambda\|\boldsymbol{\beta}\|_1$$

H

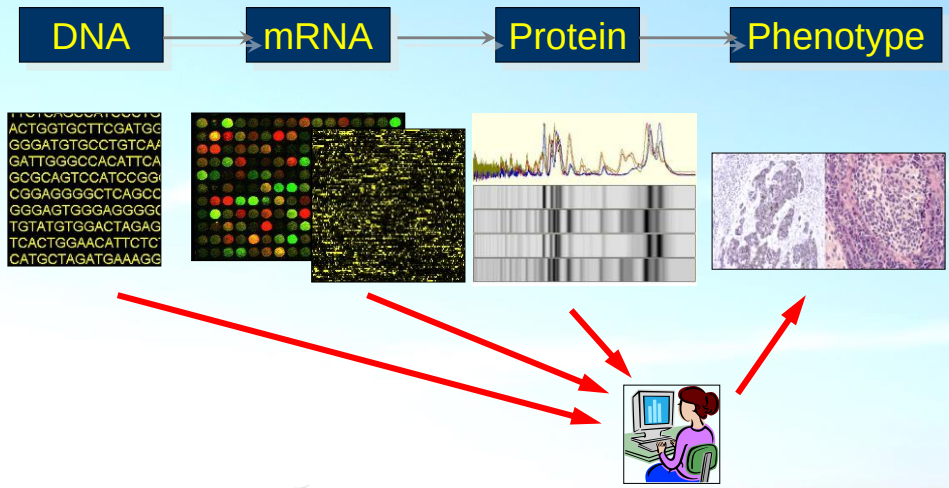
R-SVM and SVM-RFE: Examples of wrapper methods

R

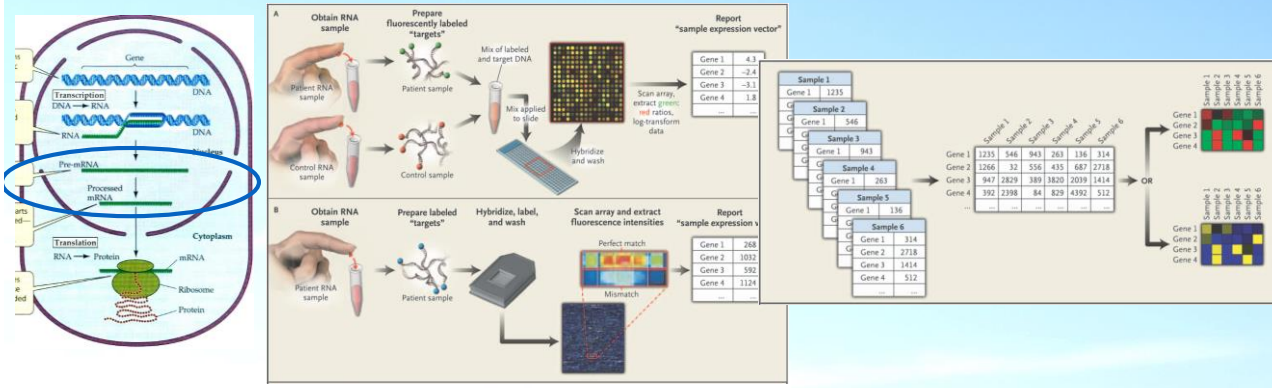
Further reading:

Xuegong Zhang, Xin Lu, Qian Shi, Xiu-qin Xu, Hon-chiu E Leung, Lyndsay N Harris, James D Iglehart, Alexander Miron, Jun S Liu and Wing H Wong, Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data, *BMC Bioinformatics*, **7**:197, 2006

Let Machines to Learn the Central Dogma



DNA microarrays for cancer study



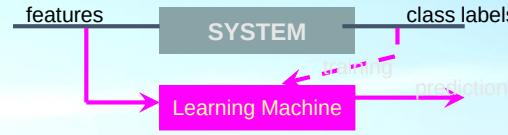
J. Quackenbush, *NEJM*, 354:2463-72, 2006

Gene Expression Analyses as Machine Learning Problems

Gene \ Case	Case 1	Case 2	...	Case 9	...
gene a	123.4	234.5		56.3	
gene b	1234.5	5678.9		78.9	
gene c	765.4	43.2		3456.1	
gene d	211.0	985.0		12.3	
gene e	432.9	543.4		102.7	
gene f	2843.0	1254.6		645.8	
gene h	23.6	65.3		564.3	
...					

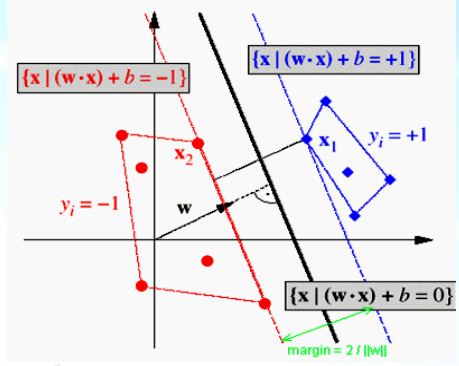
e.g. normal vs. cancer
good prognosis vs. bad prognosis
subtype A vs. subtype B

Pattern Recognition
- Classification
- Feature Selection



Choosing Linear SVM for the task

- Good generalization esp. for sparse samples in high-D spaces
- Linear SVM: Least complexity, proper for the very small sample size



R-SVM: Recursive SVM for classification and gene selection

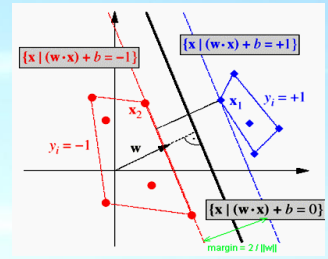
[Zhang & Wong, 2001]

- SVM training
- Select a subset of genes that gives the *best performance*:
  - Minimal error ---- always zero on training set
  - Maximal separation:

$$S = \frac{1}{n_1} \sum_{x^+ \in class1} f(x^+) - \frac{1}{n_2} \sum_{x^- \in class2} f(x^-)$$

$$S = \sum_{i=1}^{d} w_i m_i^+ - \sum_{i=1}^{d} w_i m_i^- = \sum_{i=1}^{d} w_i (m_i^+ - m_i^-)$$

$$s_i = w_i (m_i^+ - m_i^-)$$



  - Ranking genes according to their s_i
- Loop (select a subset → redo SVM training and gene ranking)

SVM-RFE: SVM - Recursive Feature Elimination

I. Guyon, J. Weston, S. Barnhill, V. Vapnik, *Machine Learning*, **46**: 389-422, 2002

- SVM training
- Select a subset of genes that gives the *best performance*:
  - Genes' contribution:

$$f(x) = \text{sgn}(w \cdot x + b)$$

$$s_i^{SVM-RFE} = w_i^2$$

$$S_{SVM-RFE} = w \cdot \left(\sum_{x_j^+ : SVs\ in\ class1} \alpha_j x_j^+ - \sum_{x_j^- : SVs\ in\ class2} \alpha_j x_j^- \right) = w \cdot w$$

  - Ranking genes according to their $s_i^{SVM-RFE}$
- Loop (select a subset → redo SVM training and gene ranking)

R-SVM vs. SVM-RFE

- Same recursive selection procedure, different ranking criteria
  - SVM-RFE:

$$S_{SVM-RFE} = \boldsymbol{w} \cdot \left(\sum_{x_j^+ : SVs\ in\ class1} \alpha_j x_j^+ - \sum_{x_j^- : SVs\ in\ class2} \alpha_j x_j^- \right) = \boldsymbol{w} \cdot \boldsymbol{w}$$

$$s_i^{SVM-RFE} = w_i (r_i^+ - r_i^-) = w_i^2$$

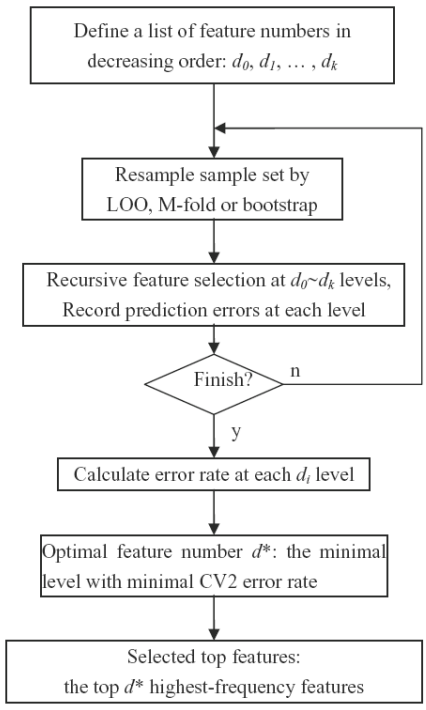
  - R-SVM:

$$S_{R-SVM} = \boldsymbol{w} \cdot \left(\frac{1}{n_i} \sum_{x^+ \in class1} x^+ - \frac{1}{n_2} \sum_{x^- \in class2} x^- \right) = \boldsymbol{w} \cdot (\boldsymbol{m}^+ - \boldsymbol{m}^-)$$

$$s_i^{R-SVM} = \boldsymbol{w}_i (\boldsymbol{m}_i^+ - \boldsymbol{m}_i^-)$$

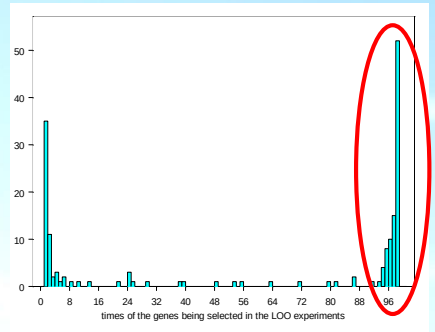
- Experiments show that R-SVM is more robust to noise

---- X. Zhang et al, *BMC Bioinformatics*, 7: 197, 2006



Feature selection with R-SVM

- Vote for the most frequently selected genes in the LOOCV experiments



- A simpler strategy:
  - LOOCV to find feature selection level of minimal CV error
  - Redo R-SVM on the whole data to select that number of features

Comparison on Simulated Data-G (with gene outliers)

Table 1. Comparison of R-SVM and SVM-RFE on Data-G (with gene outliers)

Level[a]	ReduceSV[b]	P(sv-diff)[c]	ReduceTest[d]	P(test-diff)[e]	ImproveRec[f]	P(rec-diff)[g]
800	4.01%	1.81E-42	-7.70%	4.72E-03	-3.90%	1.71E-39
600	5.77%	1.74E-49	-2.50%	4.64E-01	-1.70%	5.21E-15
500	6.83%	2.75E-51	-4.00%	1.62E-01	-0.30%	0.079189
400	8.35%	3.26E-60	2.80%	3.48E-01	1.10%	4.48E-06
300	9.33%	3.83E-58	7.40%	3.65E-02	3.70%	1.77E-31
200	8.22%	1.28E-48	19.20%	6.36E-09	6.30%	5.79E-44
150	8.55%	1.51E-53	19.50%	1.16E-08	7.10%	9.76E-46
100	4.97%	6.20E-22	11.90%	1.83E-04	6.00%	6.43E-40
90	5.84%	1.66E-27	13.70%	4.20E-06	4.60%	1.07E-30
80	5.17%	8.20E-29	12.40%	4.14E-06	4.50%	7.12E-29
70	4.14%	1.46E-27	8.50%	4.77E-04	3.80%	1.05E-24
60	3.10%	1.23E-20	10.20%	3.14E-05	3.40%	4.99E-24
50	2.27%	2.01E-15	10.20%	4.11E-06	2.90%	2.37E-21

Comparison on Simulated Data-S (with sample outliers)

Table 2. Comparison of R-SVM and SVM-RFE on Data-S (with sample outliers)

Level[a]	ReduceSV[b]	P(sv-diff)[c]	ReduceTest[d]	P(test-diff)[e]	ImproveRec[f]	P(rec-diff)[g]	ReduceOSV[h]	P(osv-diff)
800	3.25%	4.49E-41	-65.19%	5.65E-36	-10.14%	3.36E-75	50.37%	5.97E-35
600	5.80%	1.90E-57	-70.27%	3.04E-35	-7.14%	5.18E-56	72.28%	1.10E-49
500	7.02%	8.20E-63	-59.63%	1.81E-37	-5.13%	3.37E-39	80.54%	1.17E-56
400	8.26%	1.68E-67	-41.43%	8.31E-25	-2.57%	4.53E-12	89.04%	2.51E-64
300	7.72%	1.20E-58	-19.14%	2.18E-13	0.75%	4.92E-02	93.44%	7.46E-65
200	7.21%	4.54E-51	-6.53%	2.56E-04	4.00%	7.15E-16	93.91%	1.47E-61
150	9.13%	1.29E-71	2.63%	1.20E-01	6.47%	8.41E-23	93.59%	6.27E-61
100	8.30%	1.42E-64	5.56%	8.04E-04	7.69%	3.50E-22	92.44%	1.33E-61
90	8.36%	2.01E-72	4.31%	1.15E-02	6.99%	8.74E-19	91.37%	2.60E-61
80	8.01%	6.63E-71	4.45%	1.99E-02	6.99%	9.33E-18	90.26%	2.65E-60
70	7.17%	1.29E-67	6.59%	3.78E-04	7.52%	2.80E-16	88.56%	7.55E-62
60	6.67%	2.65E-65	6.16%	2.32E-03	7.27%	5.72E-13	86.38%	2.60E-62
50	5.82%	1.08E-58	7.70%	1.34E-04	7.42%	3.71E-12	83.82%	1.23E-61

Comparison on Simulated Data-R (generated from real data)

Table 3. Comparison of R-SVM and SVM-RFE on Data-R

Level[a]	ReduceSV[b]	P(sv-diff)[c]	ReduceTest[d]	P(test-diff)[e]	ImproveRec[f]	P(rec-diff)[g]
800	15.35%	1.24E-53	-3.59%	1.26E-05	-3.60%	1.50E-23
600	18.65%	3.14E-56	-7.06%	4.09E-04	2.69%	2.20E-09
500	19.58%	7.71E-58	-6.46%	1.79E-03	9.18%	1.24E-37
400	21.07%	1.80E-63	-2.74%	3.22E-05	17.32%	4.25E-59
300	22.51%	5.12E-67	-4.64%	1.26E-05	24.14%	5.43E-65
200	22.16%	9.38E-68	-0.93%	1.83E-04	30.64%	2.25E-71
150	21.78%	4.57E-64	-3.44%	8.74E-04	29.14%	5.86E-71
100	21.01%	3.21E-57	0.31%	3.22E-05	29.95%	7.74E-69
90	22.57%	1.88E-60	-2.52%	3.52E-03	27.51%	9.74E-66
80	22.88%	1.67E-65	1.84%	7.85E-05	27.92%	4.03E-62
70	21.42%	2.96E-59	0.59%	4.09E-04	27.16%	1.15E-58
60	20.20%	1.64E-55	6.16%	1.83E-04	26.83%	2.55E-60
50	18.67%	4.40E-52	4.23%	8.74E-04	25.89%	9.63E-53
40	15.37%	5.66E-46	8.99%	4.69E-06	25.39%	1.09E-55
30	11.85%	6.90E-33	9.61%	1.67E-06	24.19%	2.07E-45
20	7.87%	2.19E-18	11.43%	3.22E-05	20.86%	1.09E-34

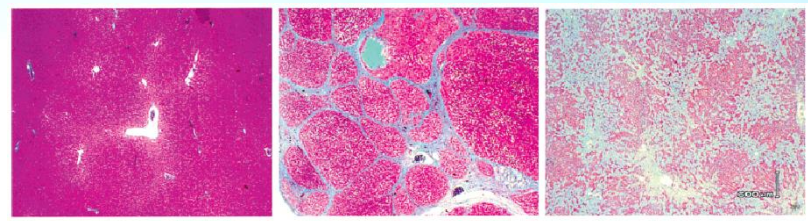
Example: R-SVM to find proteomics markers for liver cirrhosis

Proteomics 2004, 4, 3235–3245 DOI 10.1002/pmic.200400839

Molecular classification of liver cirrhosis in a rat model by proteomics and bioinformatics

Xiu-Qin Xu[1], Chon K. Leow[1, 2], Xin Lu[3], Xuegong Zhang[4], Jun S. Liu[3], Wing-Hung Wong[3], Arndt Asperger[5], Sören Deininger[5] and Hon-chiu Eastwood Leung[1]

- Diagnosis of liver cirrhosis
  - Biopsy: invasive, potential risk of internal bleeding
  - CT scanning: not able to detect early cirrhosis accurately
  - At present, there are no sensitive and specific serum or plasma markers available
  - cDNA microarray: need liver tissue by an invasive procedure
  - 2DE: not good for hydrophobic proteins, low abundant proteins and low molecular weight proteins
  - SELDI-TOF-MS: good resolution, surfaces for different proteins
- Material:
  - Normal rat (n=8)
  - Liver cirrhosis rat (n=22)
  - Liver fibrosis rat (n=5)



(a) Normal liver (x20) (b) Thioacetamide-induced cirrhosis (x20) (c) Bile duct ligation (x40)

Figure 1. Liver sections stained with Masson Trichrome. (A) normal liver, (B) cirrhotic liver, and (C) liver after bile duct ligation. The fold magnification in 1a) and 1b) was ×20; while 1c) was × 40.

Three ways of doing R-SVM on the SELDI data

- Data range: 1-10kDa
- Biomarker Wizard: R-SVM on *biomarkers* detected with Ciphergen's software
  - 78 biomarkers → 6 important markers
    - 1743.12, 3515.68, 3537.26, 4186.07, 4902.63, 8201.04 Da
- Point-to-Point: R-SVM on all 4607 points resampled from 1-10kDa
  - 7 important regions selected, covering all the previous 6 markers, but centering at different points
    - 1744.56, 3513.31, 3515.07, 3518.60, 3520.36, 4187.13, 8209.99 Da
- Sliding Window:
  - Scanning with a sliding window to pre-select some candidate regions (21 with MeanDist (Intra/Inter)<0.75), then select 6 markers by R-SVM:
    - 1743.12, 1787.89, 3515.68, 3537.26, 6207.55, 8201.04 Da

Table 1. Overall error rates of different statistical biomarkers selection approaches

Statisical Method	CV2 (external CV) errors		CV1 (internal CV) errors	
	False positive (Type 1) count	False negative (Type 2) count	False positive (Type 1) count	False negative (Type 2) count
Point-to-point RSVM	2	2	1	1
Sliding window selection	2	0	2	0
Biomarker Wizard RSVM	2	2	1	1
Type Specific Error Rate	7.7%	0 to 2.9%	3.8 to 7.7%	0 to 1.4%
Overall Error Rate	2.1 to 4.2%		2.1%	
Overall sensitivity	97.1 to 100%			
Overall specificity	92.3%			

The 3495 Da protein

- The 1743, 3515 and 3537 Da peaks were selected as important markers by 3 tests, and they are mostly dicharged peptide or sodium adducts of the 3495 Da peak.
- Taken together, … the 3495 Da peak was a fragment of some unknown histidine-rich glycoprotein.

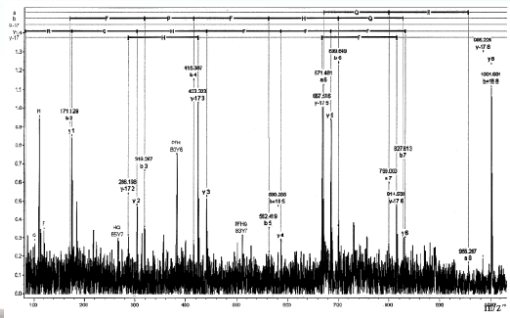


Figure 7. Peptide sequencing of the 1001 Da peak using Ultraflex MALDI TOF/TOF mass spectrometer. X-axis is the *m/z* value; the y-axis is the signal intensity.

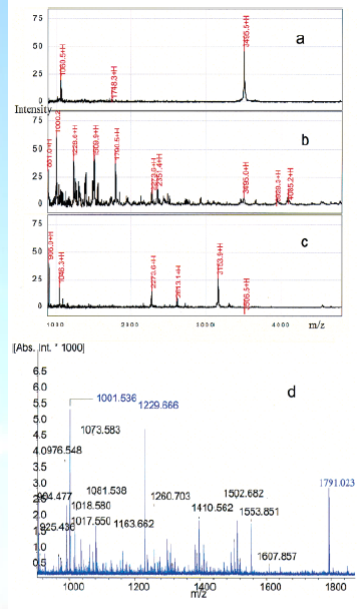


Figure 6. On-chip digestion of the purified 3495 Da protein. (A) Mass spectrum before trypsin digestion on WCX chip. (B) Mass spectrum after 3 h of on-chip trypsin digestion at 37°C. (C) Mass spectrum of trypsin alone after 3 h of on-chip digestion on WCX chip. PMF of tryptic digest after transferring to an AnchorChip and analyzed using an Ultraflex MALDI TOF/TOF mass spectrometer.

Table 4. The CV results on the rat cirrhosis data

Level [a]	R-SVM		SVM-RFE	
	CV2 [b]	AveSV [c]	CV2 [b]	AveSV [c]
93	4.2%	14.75	4.2%	14.75
80	4.2%	11.91	4.2%	14.74
70	4.2%	9.95	4.2%	14.73
60	3.2%	9.22	4.2%	13.91
50	3.2%	9.03	4.2%	13.82
40	3.2%	9.02	4.2%	14.65
30	3.2%	8.95	4.2%	13.65
20	3.2%	8.93	4.2%	9.98
18	4.2%	8.14	4.2%	9.97
16	4.2%	8.08	3.2%	7.26
15	4.2%	7.60	3.2%	7.15
14	4.2%	7.54	3.2%	7.94
13	6.3%	7.58	4.2%	7.98
12	6.3%	7.41	4.2%	8.05
11	6.3%	7.65	4.2%	8.02
10	6.3%	7.64	3.2%	9.83
9	5.3%	6.50	3.2%	8.83
8	4.2%	5.97	4.2%	7.01
7	4.2%	6.73	4.2%	6.05
6	4.2%	5.98	3.2%	5.97
5	5.3%	5.94	4.2%	5.05

The R-SVM LOOCV feature selection and classification results on the rat cirrhosis data

Table 6. The CV results on the human breast cancer dataset

Level [a]	R-SVM		SVM-RFE	
	CV2 [b]	MeanSV [c]	CV2 [b]	MeanSV [c]
98	28.7%	54.65	28.70%	54.65
88	27.9%	50.10	29.40%	55.25
79	29.4%	49.28	30.10%	52.21
71	29.4%	47.48	30.90%	50.88
63	27.9%	44.65	27.90%	48.42
56	27.2%	42.50	27.90%	46.02
50	27.9%	40.04	26.50%	40.13
45	25.7%	38.65	26.50%	40.25
40	24.3%	37.04	27.90%	34.88
36	23.5%	35.16	27.90%	34.51
32	22.1%	33.26	27.90%	30.75
28	22.8%	32.04	27.20%	27.77
25	22.1%	31.24	30.90%	24.61
22	22.1%	31.15	34.60%	23.93
19	22.8%	32.10	30.10%	26.79
17	25.7%	33.26	29.40%	31.28
15	23.5%	35.68	25.70%	35.10
13	19.9%	37.40	26.50%	42.15
11	22.1%	37.83	25.00%	46.03
9	21.3%	42.01	24.30%	50.18
8	17.6%	44.07	22.10%	49.93
7	23.5%	50.29	20.60%	51.43
6	22.1%	54.73	20.60%	52.39
5	22.1%	57.98	20.60%	52.18
4	22.8%	59.75	25.00%	58.92
3	27.2%	78.90	32.40%	77.46

R-SVM to find proteomics markers for breast cancer

[Q. Shi et al, 2005]

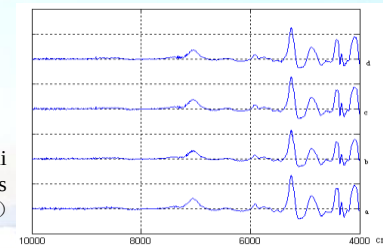
Example:

Discrimination and feature selection of geographic origins of traditional Chinese medicine herbs with NIR spectroscopy

S. Liu, X. Zhang, S. Sun, *Chinese Science Bulletin*, **50**(2): 179-184, 2005

Background:

- The efficiency of some traditional herbal medicines depends on the geographic origin and the growth condition of the herbs.
- Herbs are mixtures of many unknown compounds
- Infrared spectrometry is a key technique in identifying medical compounds, but success have not been widely reported on TCM.



The NIR derivative spectrums of Baizhi from different origins
(a. Henan, b. Hebei, c. Sichuan, d. Zhejiang）

Data and Method

- Pre-processing
- Re-sampling the spectrums
- Scanning for the effective frequency range according to CV performance on training set
- SVM with Gaussian kernel

$$f(\mathbf{x}) = \mathrm{sgn}\{\sum_{i=1}^{n} \alpha_i^* y_i K(\mathbf{x}_i \cdot \mathbf{x}) + b^*\}$$

$$DQ(k) = (1/2)\sum_{i,j=1}^{n} \alpha_i^* \alpha_j^* y_i y_j [K(\mathbf{x}_i, \mathbf{x}_j) - K(\mathbf{x}_{i(-k)}, \mathbf{x}_{j(-k)})]$$

- One-vs-all scheme for multi-category classification
- Test on independent data set

Table 1 The geographic origins of Baizhi samples

Origin	Henan	Hebei	Sichuan	Zhejiang	Total
Number of samples in Set-A	64	60	75	70	269
Number of samples in Set-B	20	18	24	22	84
Total	84	78	99	92	353

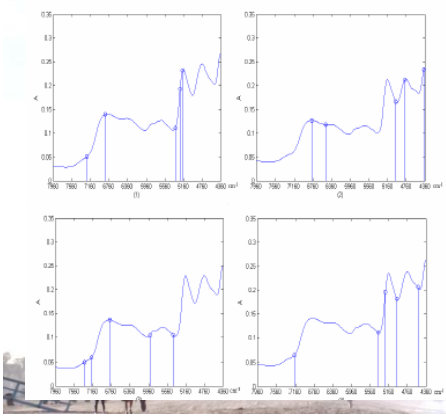
Table 2 The geographic origins and growth conditions of the Danshen samples

	Growth Condition	Geographic origin						Total
		Shandong	Shanxi	Henan	Sichuan	Zhejiang	Hebei	
Number of samples in Set-A	Wild	50	30	30	0	0	0	110
	Cultivated	70	40	25	30	35	40	240
	Total	120	70	55	30	35	40	350
Number of samples in Set-B	Wild	14	9	10	0	0	0	33
	Cultivated	22	13	9	9	10	13	76
	Total	36	22	19	9	10	13	109
Total		156	92	74	39	45	53	459

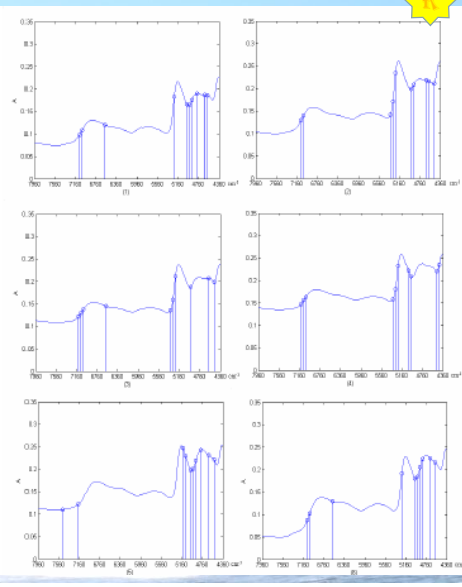
Results

Table 4 The classification accuracy with recursive SVM with different number of features, for Danshen

Number of features	Accuracy for Classification of Origins							Accuracy for wild/cultivated discrimination
	with classifier1	with classifier2	with classifier3	with classifier4	with classifier5	with classifier6	**Final accuracy**	
72	98.3%	88.6%	89.1%	100%	100%	97.5%	95.1%	95.1%
50	98.3%	90%	85.5%	100%	100%	100%	95.1%	95.7%
30	98.3%	88.6%	87.3%	100%	100%	100%	95.1%	95.1%
20	98.3%	82.9%	81.8%	100%	100%	97.5%	92.9%	93.1%
10	94.2%	74.3%	70.9%	100%	100%	97.5%	91.4%	91.1%
5	95%	42.9%	63.6%	100%	94.3%	100%	80.6%	91.7%



The 10 features for discriminating the 6 origins of Danshen (*salviae miltiorrhizae radix*)



The 5 features for discriminating the 4 origins of Baizhi (*angelicae dahuricae radix*)

单选题 1分 设置

5-min break

A Yes, I'm back to my computer.

B Sorry, not yet.

提交

8.5
Feature Extraction and K-L Transform

8.6.1 Feature Extraction to Maximize Separability

$$\boldsymbol{x} = W^T \boldsymbol{y}$$

m_2 $\left\{ \left[\quad \right] \right.$ $\left. \left[\begin{matrix} \\ \\ \\ \end{matrix} \right] \right\} m = \left[\quad \right] \} m_2$

FLD as a special case of 1D feature extraction

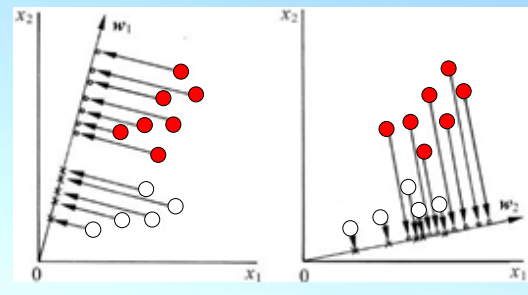
- $\boldsymbol{x} \to y$: $y_i = w^T x_i, \ i = 1, \cdots, N$

$$\max \ J_F(\boldsymbol{w}) = \frac{\boldsymbol{w}^T \boldsymbol{S}_b \boldsymbol{w}}{\boldsymbol{w}^T \boldsymbol{S}_w \boldsymbol{w}}$$

$$\boldsymbol{w}^* = \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

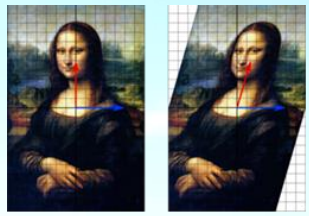
- Multiclass FLD

$$W: \max \ J(W) = \frac{|W^{\mathrm{T}} \mathrm{S_b} W|}{|W^{\mathrm{T}} \mathrm{S_w} W|}$$



- Solution: the $\leq c-1$ eigenvector solutions of $\mathrm{S_b}\boldsymbol{w} = \lambda \mathrm{S_w}\boldsymbol{w}$

- Eigenvector, characteristic vector（本征向量，eigen=own）
- Eigenvalue, characteristic value （本征值）



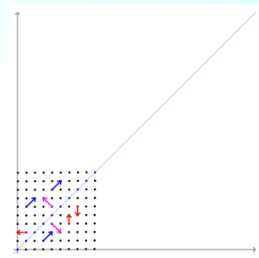
$$\boldsymbol{Av} = \lambda \boldsymbol{v}$$

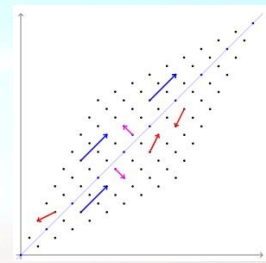
In this shear mapping the red arrow changes direction but the blue arrow does not. The blue arrow is an eigenvector of this shear mapping because it doesn't change direction, and since its length is unchanged, its eigenvalue is 1.

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$$\boldsymbol{v}_{\lambda=3} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\boldsymbol{v}_{\lambda=1} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$





https://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors

Separability

$$J_1 = \mathrm{tr}(S_w + S_b)$$

$$J_2 = \mathrm{tr}(S_w^{-1} S_b)$$

$$J_3 = \ln \frac{|S_b|}{|S_w|}$$

$$J_4 = \frac{\mathrm{tr} S_b}{\mathrm{tr} S_w}$$

$$J_5 = \frac{|S_b - S_w|}{|S_w|}$$

$$\boldsymbol{x} = W^T \boldsymbol{y}$$

Separability

$$J_1(W) = \mathrm{tr}(W^T (S_w + S_b) W)$$

$$J_2(W) = \mathrm{tr}[(W^T S_w W)^{-1} (W^T S_b W)]$$

$$J_3(W) = \ln \frac{|W^T S_b W|}{|W^T S_w W|}$$

$$J_4(W) = \frac{\mathrm{tr}(W^T S_b W)}{\mathrm{tr}(W^T S_w W)}$$

$$J_5(W) = \frac{|W^T \Sigma W|}{|W^T S_w W|}, \Sigma = S_w + S_b$$

To find W^T:

$$J(\boldsymbol{x}) = \max_{\{W\}} J(W^T \boldsymbol{y})$$

- Find W^T so that $J(\boldsymbol{x}) = \max_{\{W\}} J(W^T y)$
- Result:
  - If the eigenvalues of $S_w^{-1}S_b$ are

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$$

The best feature extraction matrix for $J_1,\ J_2,\ J_3,\ J_4$ and J_5 should be

$$W = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_d]$$

composed of the eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_d$ of $S_w^{-1}S_b$ corresponding to the first d largest eigenvalues.

8.6.2 Karhunen-Loève Transform (KLT)

Kari Karhunen (1915-1992), Loève (1907-1979)

- Basic Idea: expansion of functions
  - Function expansion with an orthogonal or orthonormal basis
  - Using expansion coefficients to represent original functions
- Discrete K-L expansion:
  - Expansion of random vectors with a group of orthogonal basis vectors
  - Using expansion coefficients to represent original vectors
    - New feature space: the space of spanned by the basis vectors
    - New feature vector: the vector of the expansion coefficients

- Discrete K-L Expansion

For random vector $\boldsymbol{x} \in R^D$, use a deterministic set of complete orthonormal vectors $\boldsymbol{u}_j,\ j = 1,2,\cdots,\infty$ as the basis to expand as

$$\boldsymbol{x} = \sum_{j=1}^{\infty} c_j \boldsymbol{u}_j$$

where $\boldsymbol{u}_i^T \boldsymbol{u}_j = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$. Multiply both side with $\boldsymbol{u}_j^T$, we get $c_j = \boldsymbol{u}_j^T \boldsymbol{x}$.

To use only $d < D$ dimension to approximate $\boldsymbol{x}$, we have

$$\hat{\boldsymbol{x}} = \sum_{j=1}^{d} c_j \boldsymbol{u}_j \ .$$

The error is $\xi = E[(\boldsymbol{x} - \hat{\boldsymbol{x}})^T (\boldsymbol{x} - \hat{\boldsymbol{x}})] = E\left[\left(\sum_{j=d+1}^{\infty} c_j \boldsymbol{u}_j\right)^T \left(\sum_{j=d+1}^{\infty} c_j \boldsymbol{u}_j\right)\right] = E\left[\sum_{j=d+1}^{\infty} c_j^2\right]$

$$= E\left[\sum_{j=d+1}^{\infty} \boldsymbol{u}_j^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{u}_j\right] = \sum_{j=d+1}^{\infty} \boldsymbol{u}_j^T E[\boldsymbol{x}\boldsymbol{x}^T]\ \boldsymbol{u}_j$$

Denote $\psi = E[\boldsymbol{x}\boldsymbol{x}^T]$, we have

$$\xi = \sum_{j=d+1}^{\infty} \boldsymbol{u}_j^T E[\boldsymbol{x}\boldsymbol{x}^T]\ \boldsymbol{u}_j = \sum_{j=d+1}^{\infty} \boldsymbol{u}_j^T \psi\ \boldsymbol{u}_j$$

To find the orthonormal basis that minimizes the error

$$\min\ \xi$$
$$\text{s.t.}\ \ \boldsymbol{u}_j^T \boldsymbol{u}_j = 1$$

we define the Lagrange function

$$g(\boldsymbol{u}) = \sum_{j=d+1}^{\infty} \boldsymbol{u}_j^T \psi \boldsymbol{u}_j - \sum_{j=d+1}^{\infty} \lambda_j \left[\boldsymbol{u}_j^T \boldsymbol{u}_j - 1\right]$$

Let $\frac{\partial}{\partial \boldsymbol{u}_j} g(\boldsymbol{u}) = 0,\ j = d+1,\cdots,\infty$, we get

$$\left(\psi - \lambda_j I\right)\ \boldsymbol{u}_j = 0, \qquad j = d+1,\cdots,\infty$$

Take $d = 0$, we have

$$\psi \boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j, \qquad j = 1,2,\cdots,\infty$$

Solution: eigenvectors of $\psi = E[\boldsymbol{x}\boldsymbol{x}^T]$

and

$$\xi = \sum_{j=d+1}^{\infty} \lambda_j$$

K-L Transform (KLT):

- Using the d eigenvectors of $\psi = E[\boldsymbol{x}\boldsymbol{x}^T]$ corresponding to the first d largest eigenvalues to express the original vectors $\boldsymbol{x}$, so that the error is the smallest among all orthogonal vector expansions.

$$\psi \boldsymbol{u}_j = \lambda_j \boldsymbol{u}_j, \qquad j = 1,2,\cdots,\infty$$

- $\boldsymbol{u}_j, j = 1,2,\cdots,d$ are the vectors that span the new feature space
- Coefficients $c_j = \boldsymbol{u}_j^T \boldsymbol{x}, j = 1,2,\cdots,d$ compose the new d-dimensional feature

Kari Karhunen, 1947; Michel Lo ève, 1948
a.k.a. Hotelling transform (Harold Hotelling, discrete formulation 1933)
a.k.a. PCA

Properties of K-L Transform

1. Best compressive representation of original data —— minimal squared error
   - Eigenvalue λ_j corresponding to each new feature $\boldsymbol{u}_j$ reflects its relative contribution in representing the original information (information magnitude), ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$
2. The new features are uncorrelated:

$$E[c_i c_j] = E[\boldsymbol{u}_i^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{u}_j] = \lambda_i \boldsymbol{u}_i^T \boldsymbol{u}_j = \lambda_i \delta_{ij}$$

$$E[\boldsymbol{c}\boldsymbol{c}^T] = U^T \psi U = \Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \lambda_d \end{bmatrix}$$

   where $U = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_D]$, $\psi = E[\boldsymbol{x}\boldsymbol{x}^T]$.

3. Minimal representation entropy $H_R = -\sum_{j=1}^{D} \tilde{\lambda}_j \log \tilde{\lambda}_j$ $\tilde{\lambda}_j$: normalized λ_j
   - Best centralizes variances in original random vectors to the first few new features
4. Minimal total entropy $H_p = -E[\log p(\boldsymbol{x})]$
   - If select from the smallest eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d \leq \cdots \leq \lambda_D$
   - Meaning: smallest variance → mean best representing the set

KLT for pattern recognition

- Do eigenvalue decomposition of the matrix

$$\psi = E[\boldsymbol{x}\boldsymbol{x}^T] \quad \text{(2nd-order moment)}$$

or $$\sum = E[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T] \quad \text{(covariance matrix)}$$

or $$S_w = \sum_{i=1}^{c} P_i \Sigma_i \quad \text{(within-class scatter matrix)}$$

where $\Sigma_i = E[(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^T]$

- Select a subset of coordinates best for classification (according to some criteria or assumption for the classification)

Case 1: Extracting classification information in class means

- Use KLT to remove correlation in features
- Choose KLT features with small variances and large differences in class means as extracted features

- **Algorithm**
  - Compute $S_w = \sum_{i=1}^{c} P_i \Sigma_i$
  - Do KLT on S_w to get $\lambda_i, \boldsymbol{u}_i, i = 1, \cdots, D$ and new features $\boldsymbol{y}_i = \boldsymbol{u}_i^T \boldsymbol{x},\ i = 1, \cdots, D$ with variances λ_i
  - Compute $J(y_i) = \frac{\boldsymbol{u}_j^T S_b \boldsymbol{u}_j}{\lambda_i},\ j = 1, \cdots, D$

$$S_b = \sum_{i=1}^{c} P(\omega_i)(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

to assess the separability of each new feature and rank them as

$$J(y_1) \geq J(y_2) \geq \cdots \geq J(y_d) \geq \cdots \geq J(y_D)$$

  - Choose the first d new features $U = [\boldsymbol{u}_1\ \ \boldsymbol{u}_2\ \ \cdots\ \ \boldsymbol{u}_d]$ as the extracted features

Case 2: Optimal extraction of class-mean information

- Make features uncorrelated, and
- Use the minimal dimensions to extract classification information in class means

- **Algorithm**
  - Use S_w to do KLT: $U^T S_w U = \Lambda$

    Let $B = U\Lambda^{-\frac{1}{2}}$ so that $B^T S_w B = I$ (whitening transformation)

    Computer the between-class scatter matrix $S_b' = B^T S_b B$
  - Do KLT again on S_b', to compress the information in class means
  - Since $rank(S_b') \leq c - 1$, there are at most $d = c - 1$ non-zero eigenvalues, and there corresponding eigenvectors are
    $$V' = [v_1, \cdots v_d]$$
  - The overall transform is: $W = U\Lambda^{-\frac{1}{2}}V'$

Case 3: Extracting classification information from de-centralized samples

- Remove class means
- Only consider classification information in class covariances

- **Algorithm**
  - Do KLT on S_w, the variance of the jth new feature in class i is r_{ij}
  - Normalize $\tilde{r}_{ij} = P_i \frac{r_{ij}}{\lambda_j}, i = 1, \cdots, c,\ j = 1, \cdots, D$

    λ_j: jth eigenvalue of S_w, i.e. total variance on jth feature $\lambda_j = \sum_i P_i r_{ij}$

    → $\sum_{i=1}^{c} \tilde{r}_{ij} = 1$ (like a PDF)
  - Use entropy $J(x_j) = -\sum_{i=1}^{c} \tilde{r}_{ij} \log \tilde{r}_{ij}$ or $J(x_j) = \prod_{i=1}^{c} \tilde{r}_{ij}$ to represent diversity on jth component
  - Rank $J(x_1) \leq J(x_2) \leq \cdots \leq J(x_d) \leq \cdots \leq J(x_d)$ and choose the first d

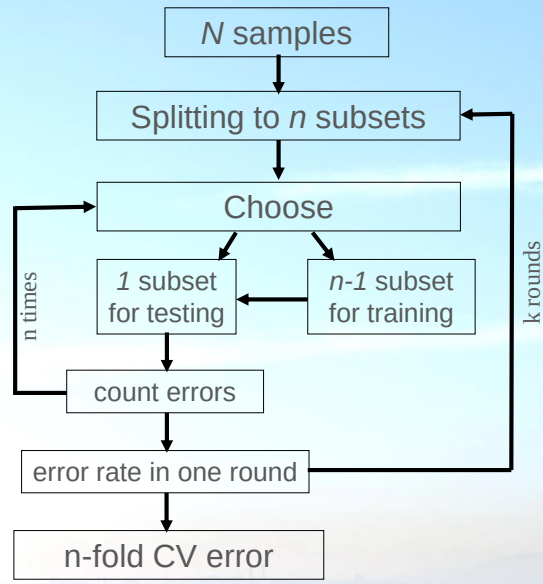
- Case 4: Combining classification information in means and variances
  - Choose the $d' \leq c-1$ features that best extracts information in class means
  - Choose the other $d-d'$ new features to extract in class variances

1-minute break

8.6
Assessment of
Feature-engineered Classifiers

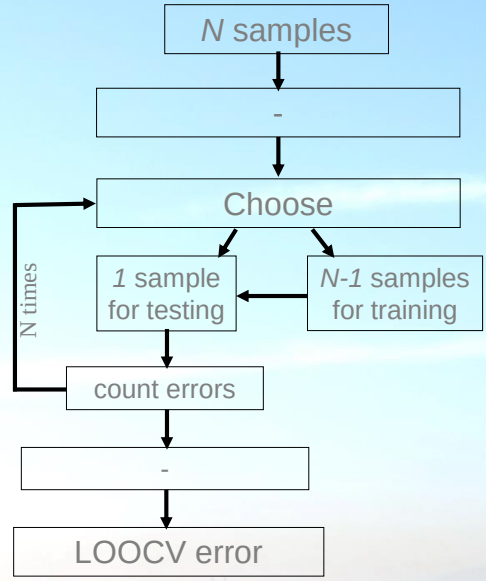
Recap: Cross Validation (CV)

- For a reliable estimation of error, we need large test sets
  - which can be expensive
- A compromise: Cross Validation
  - n-fold cross validation
  - Statistical properties of the CV error estimation



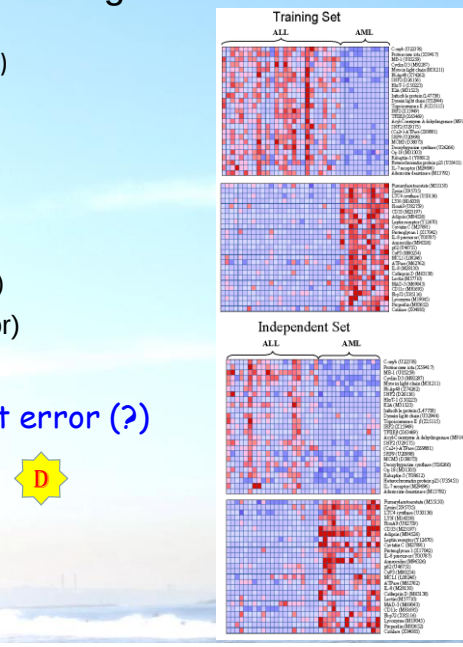
Recap: LOOCV

- The extreme case: Leave-one-out cross validation (LOOCV)
- Discussion
  - Widely used
  - Cross-validation provides almost unbiased estimation of the performance of a machine trained on $N-1$ or $\left(1-\frac{1}{n}\right)N$ samples.
  - Slightly conservative



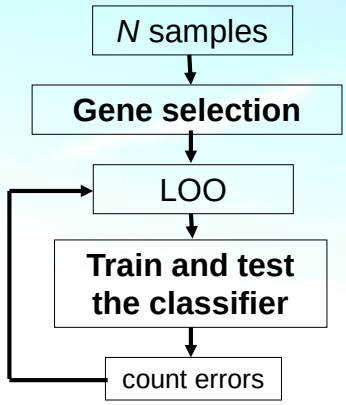
Caution:
CV error can be very biased with improper timing of cross validation

- e.g. ALL/AML classification (T. Golub et al, 1999)
  - Gene selection (with SNR they defined)
  - Classification (weighted voting)
  - CV and independent test
- Result:
  - CV error 5.26% (2/38 no-calls that would be error)
  - Test error 14.7% (5/34 no-calls that would be error)
- Observation:
  - The CV error is so different with the test error (?)

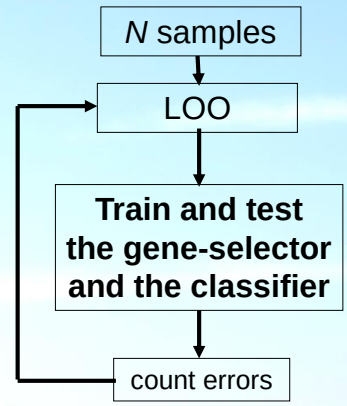


Two schemes of Cross Validation

- Scheme 1 (CV1)



- Scheme 2 (CV2)



Zhang et al, 2001, unpublished
Zhang et al, BMC Bioinformatics, 2006

Example: R-SVM on ALL/AML classification

- Method:
  - R-SVM
  - CV2 and independent test
- Result:
  - CV2 error: 1/38 error (2.63%)
  - Test error: 1/34 error (2.94%)
- Observation:
  - R-SVM reaches a better accuracy
  - CV2 error is a better estimate of the test error

Golub et al:
CV1 error: 5.26%
Test error: 14.7%

Zhang et al, 2001, unpublished

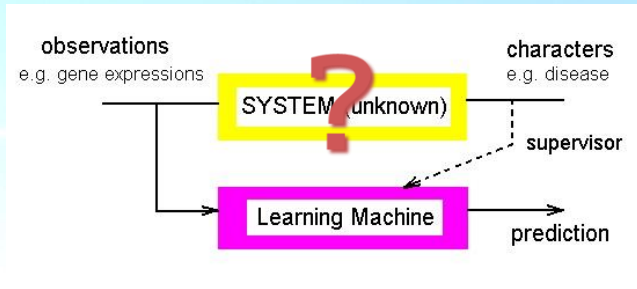
Artificial “Fake-class” Data

- Two fake "classes" of 20 cases each, generated from the same Gaussian model of 1000 simulated genes.

# selected genes	CV1 error	CV2 error
1000	0.5	0.5
500	0.275	0.5
200	0.1	0.575
100	0.025	0.475
50	0.025	0.5
30	0.025	0.475
20	0	0.475

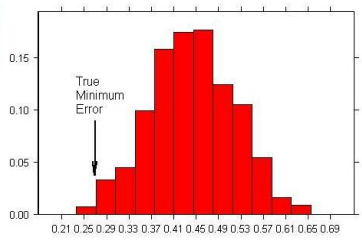
Zhang et al, 2001, unpublished

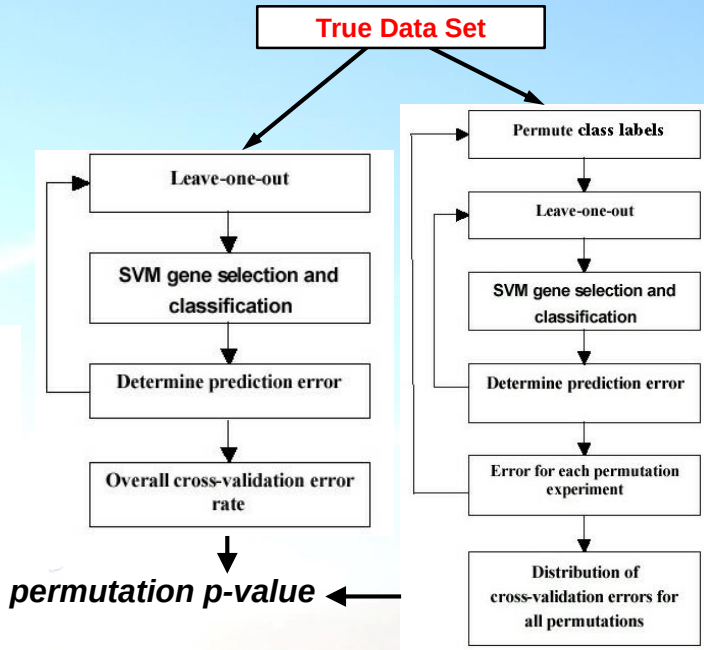
Recall: The significance of an accuracy



Zhang et al, 2001, unpublished

Permutation





Zhang et al, 2001, unpublished

Example: ALL/AML classification

- Method: R-SVM (CV and independent test)
- Result:
  - training (CV2): 1 error (2.63%)
  - test: 1 error (2.94%)
- Difference of CV schemes:
  - R-SVM CV1 error: 0 Permutation p-value: 0.168
  - R-SVM CV2 error: 0.0263 Permutation p-value: 0.000

Zhang et al, 2001, unpublished

Homework

- Computer exercises (Ex4a)
  - Feature selection
  - Lasso
- Deadline:
  - Nov. 10 (Wednesday), 23:00

单选题 1分 设置

5-min break

A Yes, I'm back to my computer.

B Sorry, not yet.

提交

About the mid-term test

- Nov. 4, 9:50-10:30
- Venue: I-205
- Open-book test
  - To check your understanding of the basic concepts and way of study
- TAs will work with each on-line student for the test
- Questions in English, answers in Chinese or English
- Suggestions for your review:
  - Look back on what we have learned so far, including the homework. Ask yourself on the basic principles and ideas behind the methods.