10-810 /02-710

Computational Genomics

Classification

Types of classifiers

- We can divide the large variety of classification approaches into roughly two main types

1. Generative:
   - build a generative statistical model
   - e.g., mixture model

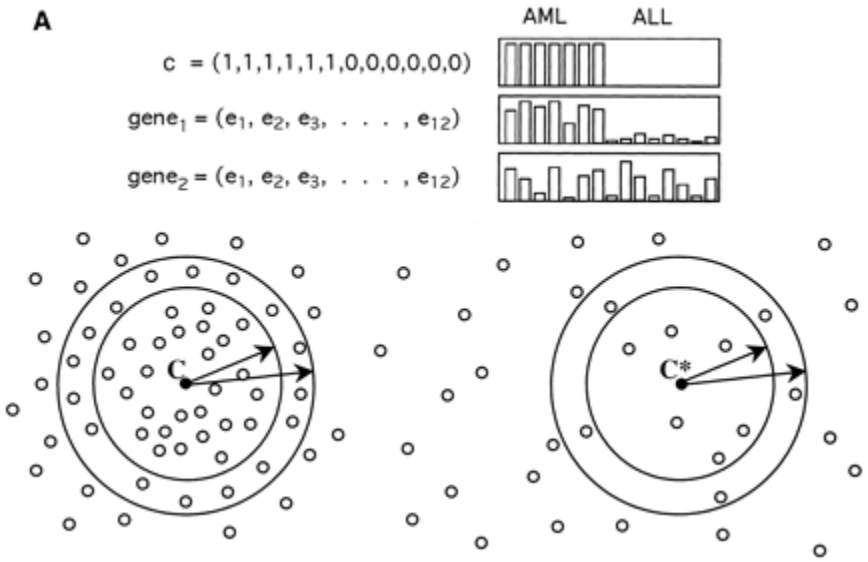
2. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., logistic regression

Golub et al

- 38 test samples (27 ALL 11 AML)
- Each gene was initially compared to an idealized expression pattern: 11111111111111000000000000000000 for class 1 and similarly 00000000000000000000011111111111111 for the second class.
- The actual selection was done by setting:

$$p(g,c) = \frac{\mu_1(g) - \mu_2(g)}{\sigma_1(g) + \sigma_2(g)}$$

- Large values of |p(g,c)| indicate strong correlation between the gene and the classes, and the sign of p(g,c) depends on the class in which this gene is expressed.


A
AML
ALL
c = (1,1,1,1,1,1,1,0,0,0,0,0,0)
gene1 = (e1, e2, e3, . . . , e12)
gene2 = (e1, e2, e3, . . . , e12)
C
C*

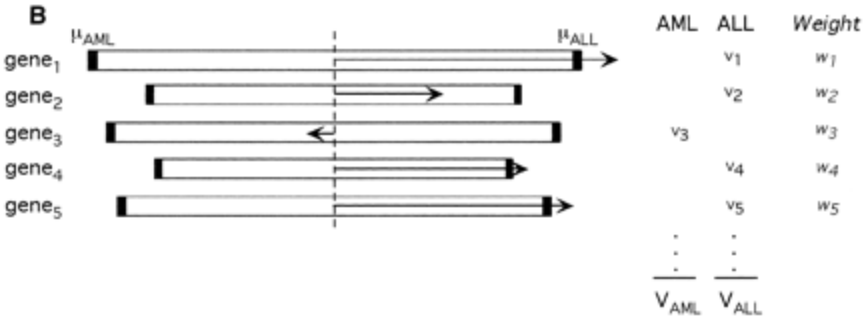
Weighted voting

- Use a subset of the selected genes (50).
- Set $a_g = p(g,c)$ and $b_g = (\mu_1(g)+\mu_2(g))/2$
- Given a new sample X, we set the vote of gene g to:

$$v_g = a_g(x_g - b_g)$$

- A positive value is a vote for class 1 and a negative for the second class

Weighted voting



Voting strength

- The votes are summed for each of the two classes.
- The decision is made by using:

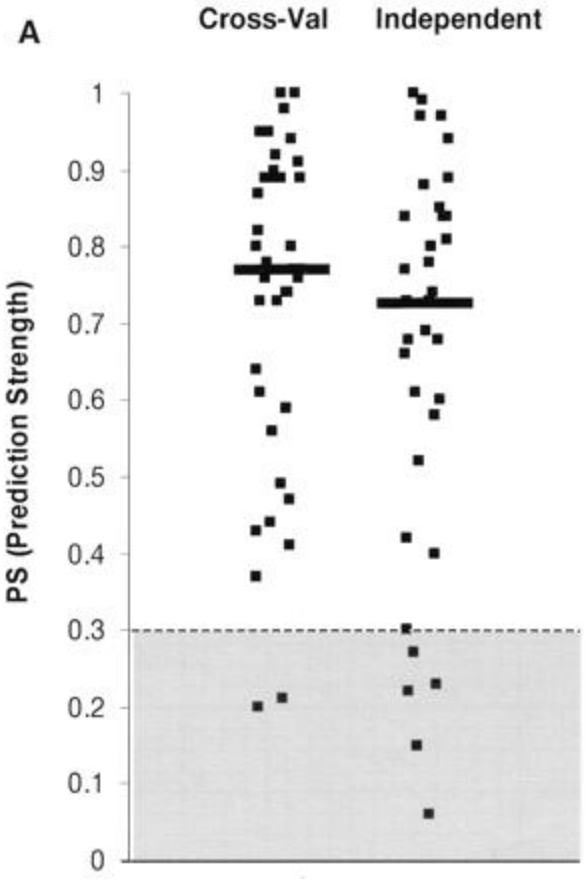
$$PS = \frac{v_{win} - v_{lose}}{v_{win} + v_{lose}}$$

- PS determines our confidence in the classification result.
- How do we chose PS ?

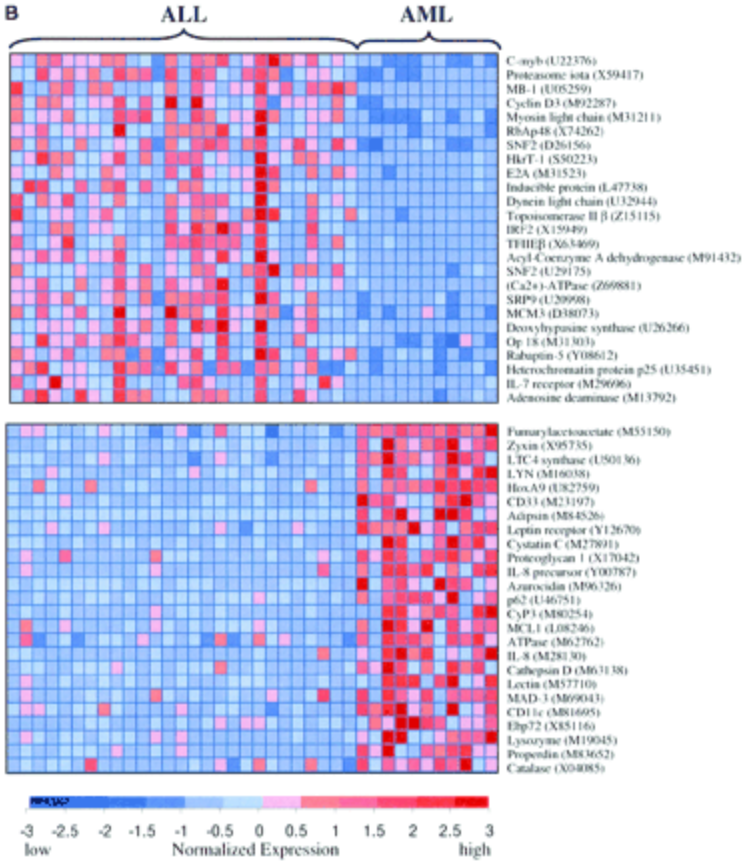
Testing the classifier

- Cross validation.
- Test set: 38 samples:
  - 20 ALL
  - 14 AML
- 29 of 34 had a classification value higher than the threshold and all were predicted correctly.

Classification results



Selected genes



Can we do better?

Generative classifiers: Bayes classification

- A mixture of two Gaussians, one Gaussian per class choice of class:

$$X \in class \quad 1 \Rightarrow X \sim (\mu_1, \sigma_1)$$
$$X \in class \quad 0 \Rightarrow X \sim (\mu_0, \sigma_0)$$

- where X corresponds to, e.g., a tissue sample (expression levels across the genes).

- Three basic problems we need to address:
  - decisions
  - estimation
  - variable (feature) selection

Decision: Bayesian classifiers

- Given a probabilistic model and an unlabeled data vector **X**, we can use Bayes rule to determine the class:

$$p(class=1|X)=\frac{P(X|class=1)P(class=1)}{P(X|class=1)P(class=1)+P(X|class=0)P(class=0)}$$

- We compute *p(class=1|X)* and *p(class =0|X)* and chose the class with the highest probability
- This method can be easily extended to multiple classes

Decision boundary

- Given a probabilistic model and an unlabeled data vector **X**, we can use Bayes rule to determine the class:

$$p(class = 1 | X) = \frac{P(X | class = 1)P(class = 1)}{P(X | class = 1) + P(X | class = 0)}$$

- Using Bayes classifiers, the decision comes down to the following (log) likelihood ratio:

$$\log \frac{p(X | \mu_1, \sigma_1)\, p(class = 1)}{p(X | \mu_0, \sigma_0)\, p(class = 0)} > 0 \Rightarrow class = 1$$

Decision boundaries

- Equal covariances

$$X \sim (\mu_1, \Sigma\ ); class = 1$$

$$X \sim (\mu_0, \Sigma\ ); class = 0$$

The decision rule is linear

Decision boundaries

- Unequal covariances

$$X \sim (\mu_1, \sigma_1); class = 1$$
$$X \sim (\mu_0, \sigma_0); class = 0$$

- The decision rule is quadratic

Estimation

- Suppose we are given a set of labeled tissue samples

  $X^1 \ldots X^k$ – class = 1

  $X^{k+1} \ldots X^n$ – class =0

- We can estimate the two Gaussians separately.
- For example, using maximum likelihood estimation we get

  $P(class=1) = k/n$

  μ_1 = sample mean of $X^1 \ldots X^k$

  Σ_1 = sample covariance of $X^1 \ldots X^k$
- And similarly for the other class(es)

Golub et al

- Leukemia classification problem

- 7130 ORFs (expression levels)
- 38 labeled training examples,
- 34 test examples

Our mixture model (assume equal class priors)

$$X \sim (\mu_1, \Sigma\ ); class = 1$$

$$X \sim (\mu_0, \Sigma\ ); class = 0$$

Problems?

Golub et al

•Leukemia classification problem

•7130 ORFs (expression levels)

•38 labeled training examples,

•34 test examples

Our mixture model (assume equal class priors)

$$X \sim (\mu_1, \Sigma\ ); class = 1$$

$$X \sim (\mu_0, \Sigma\ ); class = 0$$

Problems?

For 7000+ genes we would need to set roughly 18,000,000 parameters in each covariance matrix! (with 38 examples)

Naïve Bayes classifiers

- This full covariance model is too complex, we need to constrain the covariance matrices
- The simplest constraint we can use is a diagonal covariance matrix instead of a full covariance
- When using such a matrix we make the (implicit) assumption that the genes are *independent* given the class labels
- In other words, we assume that:

$$p(X \mid class = 1) = \prod_i p(X_i \mid class = 1)$$

$$X_i \sim N(\mu_i^1, \sigma_i^2)$$

where X_i is the expression value for gene i

Naïve Bayes classifiers

- Lets further assume equal variance for a specific gene across the two sets of samples (that is, noise is independent of the sample condition)
- As a result, we need to only estimate class-conditional means and a common variance for each gene
- How well might we do in the Golub et al. task?

3 test errors (out of 34)

Feature selection

- Test which genes are predictive of the class distinction
- Why is this important? Is more information always better?

Feature selection

- H_0 is that a gene is not predictive of the class label
- H_1 is that a gene can predict the class label

$$H_0 = X_1 \sim N(\mu, \sigma^2), X_2 \sim N(\mu, \sigma^2)$$
$$H_1 = X_1 \sim N(\mu_1, \sigma^2), X_2 \sim N(\mu_2, \sigma^2)$$

- We can use a likelihood ratio test for this purpose Let x^t_i denote the observed expression levels for gene i
- The parameter estimates are computed from the available populations in accordance with the hypothesis.

Gene selection (cont.)

- We rank the genes in the descending order of the test statistics $T(x_i)$.
- How many genes should we include?
- As discussed, we can use various multiple hypothesis correction methods here, for example FDR.
- For 187 genes we have a FDR < 1%

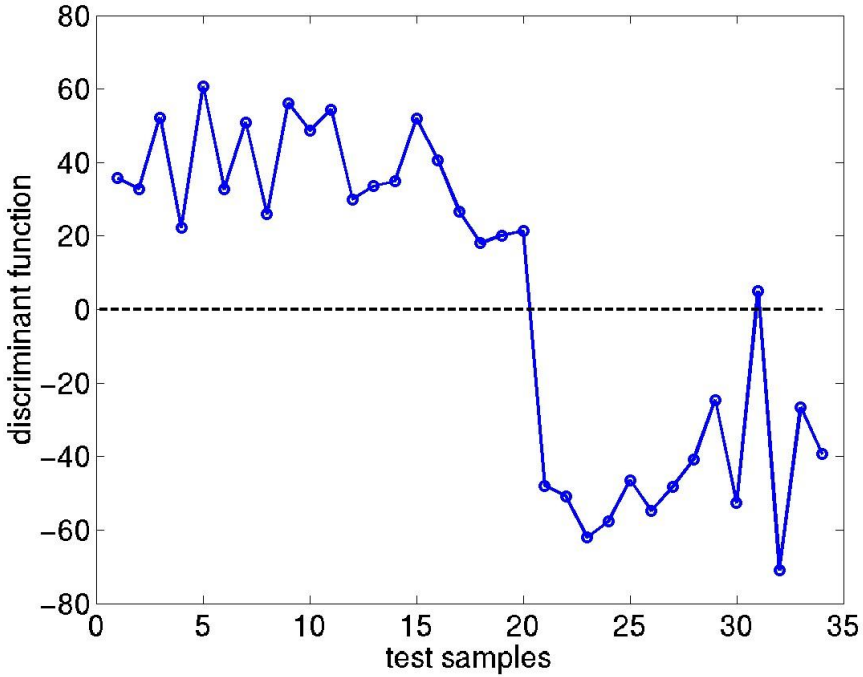
Golub cont.

- The figure shows the value of the discriminant function

$$f(x) = \log \frac{p(X \mid \mu_1, \sigma_1)}{p(X \mid \mu_0, \sigma_0)}$$

across the test examples

- The only test error is also the decision with the lowest confidence



Types of classifiers

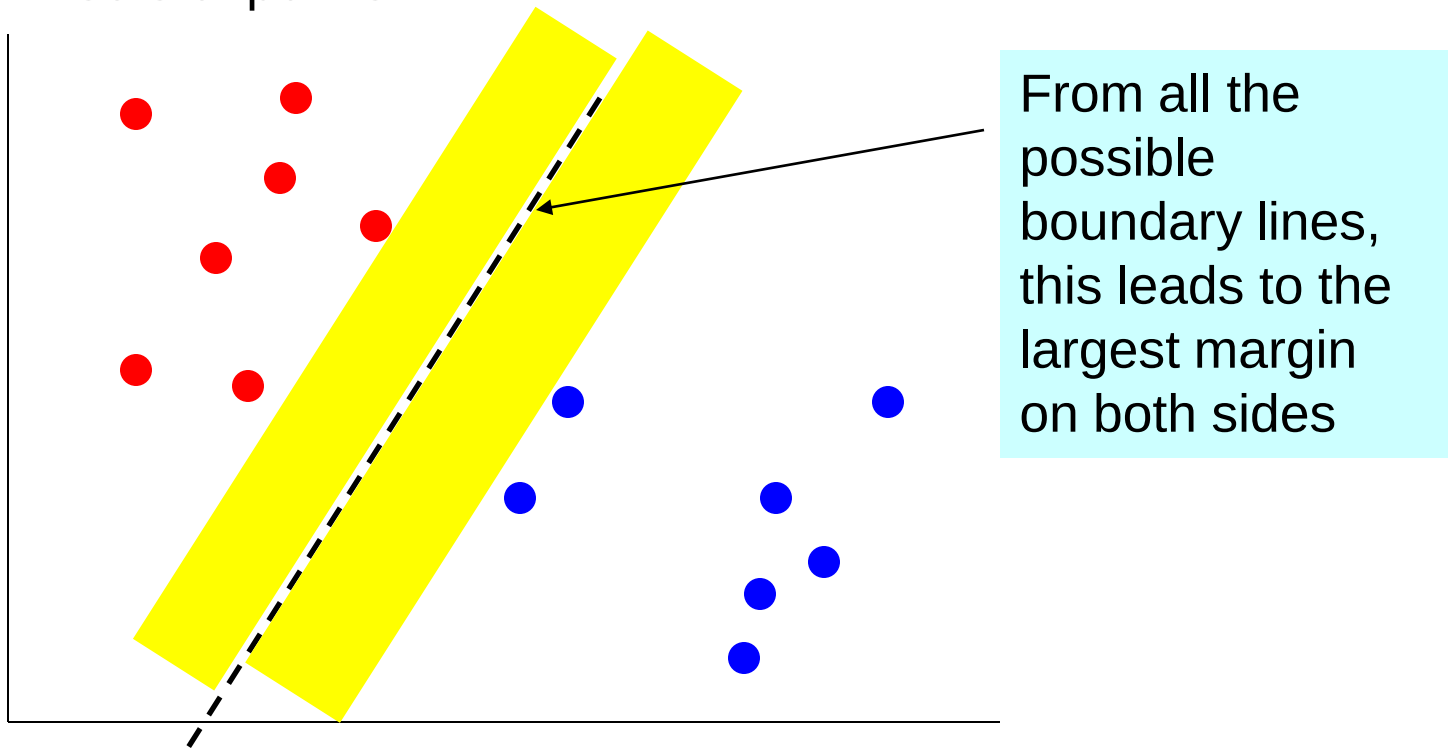
- We can divide the large variety of classification approaches into roughly two main types

1. Generative:
   - build a generative statistical model
   - e.g., mixture model

2. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., logistic regression

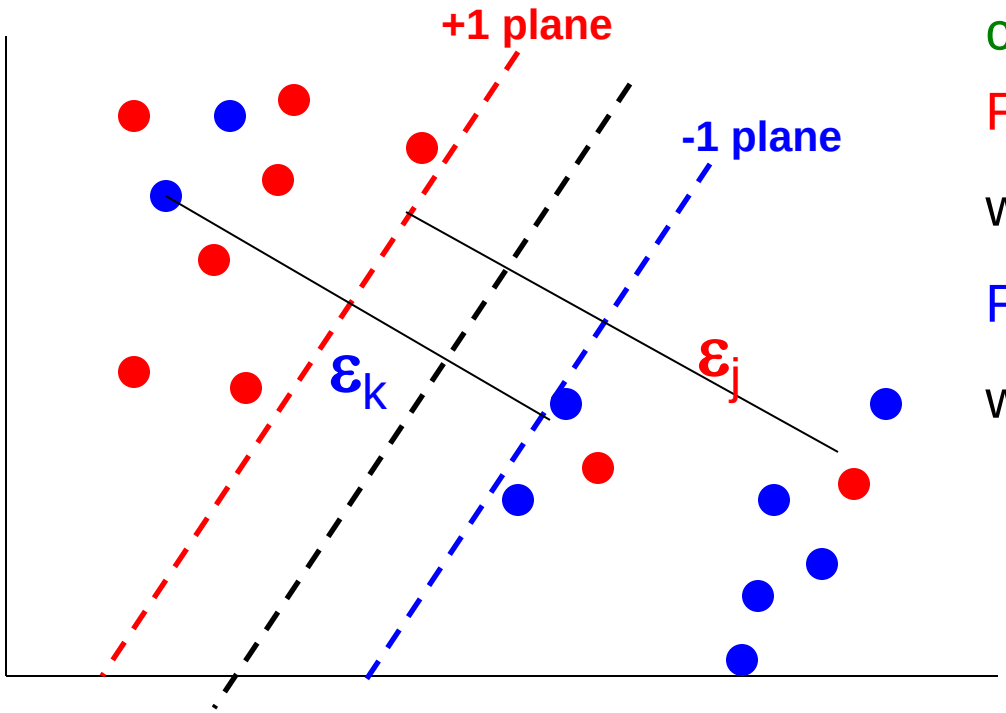
SVM: A max margin classifier

- Instead of fitting all points, focus on boundary points
- Learn a boundary that leads to the largest margin from both sets of points



From all the possible boundary lines, this leads to the largest margin on both sides

SVM for non linearly separable data



SVM optimizes the following:

$$\min_w \frac{w^T w}{2} + \sum_{i=1}^{n} C\varepsilon_i$$

subject to the following inequality constraints:

For all x_i in class + 1

$w^T x+b \geq 1- \varepsilon_i$

For all x_i in class - 1

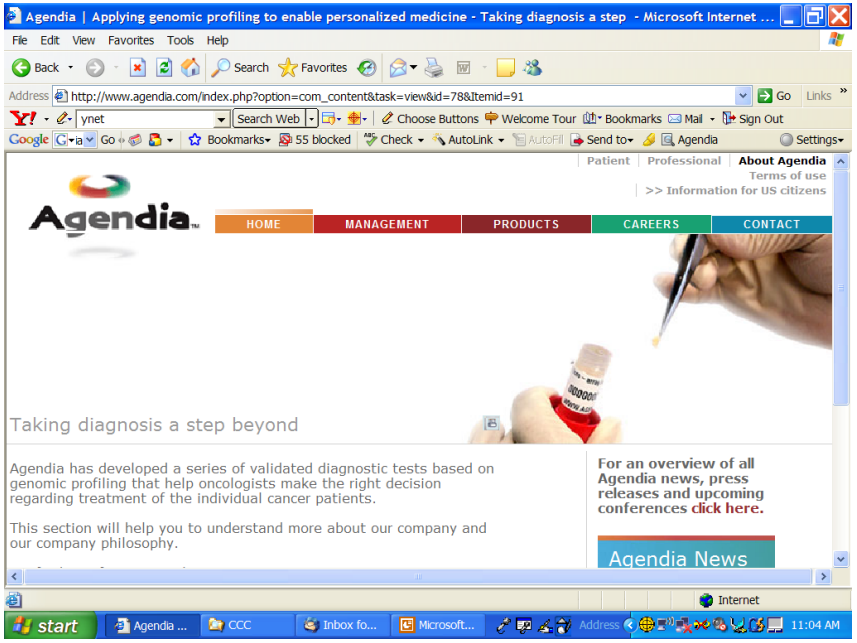
$w^T x+b \leq -1+ \varepsilon_i$

For all i

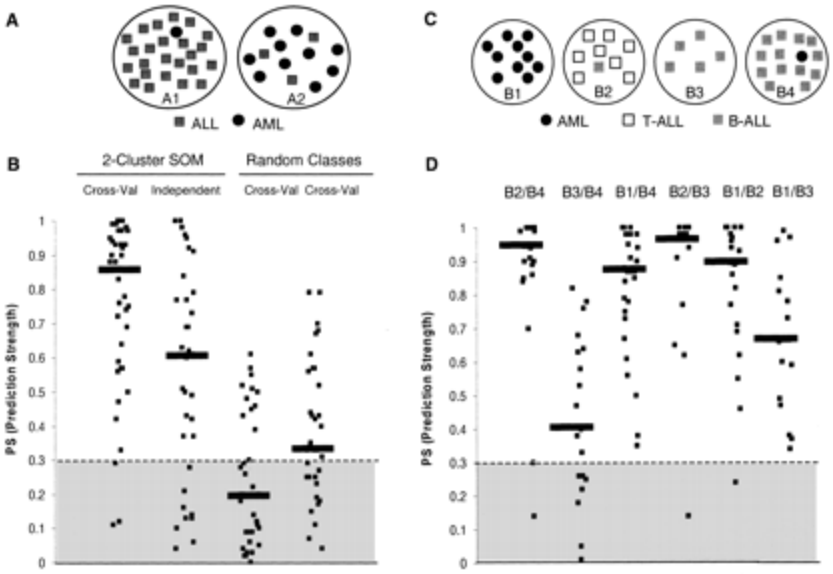
$\varepsilon_i \geq 0$

MammaPrint FDA Approved Gene-Based Breast Cancer Test

- Actual classifier used is proprietary.
- But based on work that led to this diagnostic tool it is likely based on SVMs
- The researchers also performed some feature selection since only 70 genes are used by the classifier.



Unsupervised



- Build a class predictor using the clustering algorithm
- Use cross validation to determine class membership
- Problems ?

What you should know

- Optimal ordering can help interpreting expression results
- Different classifier types
- Cross validation, feature selection