Computational functional genomics
(Spring 2005: Lecture 10)

David K. Gifford
(Adapted from a lecture by Tommi S. Jaakkola)
MIT CSAIL
Topics

- **Basic classification approaches**
  - decisions
  - estimation
  - variable selection

- **Examples**

- **More advanced methods**
Classification

• 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
Generative approach to classification

- A mixture of two Gaussians, one Gaussian per class

\[
\begin{align*}
\text{choice of class} & \\
\overleftarrow{P(\text{class} = 1)} & \overrightarrow{P(\text{class} = 0)} \\
X \sim N(\mu_1, \Sigma_1) & X \sim N(\mu_0, \Sigma_0)
\end{align*}
\]

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

- Three basic problems we need to address:
  1. decisions
  2. estimation
  3. variable selection
Mixture classifier cont’d

• Examples $X$ (tissue samples) are classified on the basis of which Gaussian better explains the new sample (cf. likelihood ratio test)

$$\log \frac{P(X|\mu_1, \Sigma_1)P(class = 1)}{P(X|\mu_0, \Sigma_0)P(class = 0)} > 0 \text{ class } = 1 \quad (1)$$

$$\leq 0 \text{ class } = 0 \quad (2)$$

where the prior class probabilities $P(class)$ bias our decisions towards one class or the other.

• Decision boundary

$$\log \frac{P(X|\mu_1, \Sigma_1)P(class = 1)}{P(X|\mu_0, \Sigma_0)P(class = 0)} = 0 \quad (3)$$
Mixture classifier: decision boundary

- Equal covariances

\[
X \sim N(\mu_1, \Sigma), \quad \text{class} = 1 \quad (4)
\]
\[
X \sim N(\mu_0, \Sigma), \quad \text{class} = 0 \quad (5)
\]

- The decision rule is linear
Mixture classifier: decision boundary

- Unequal covariances

\[
X \sim N(\mu_1, \Sigma_1), \quad \text{class} = 1 \quad (6)
\]

\[
X \sim N(\mu_0, \Sigma_0), \quad \text{class} = 0 \quad (7)
\]

- The decision rule is quadratic
Mixture classifier: estimation

- Suppose we are given a set of labeled tissue samples

\[
\begin{align*}
\text{class}=1 & \quad x^{(1)}, \ldots, x^{(n_1)} \\
\text{class}=0 & \quad x^{(n_1+1)}, \ldots, x^{(n)}
\end{align*}
\]  

(8)

- We can estimate the two Gaussians separately.

For example, maximum likelihood estimation gives

\[
\hat{P}(\text{class } = 1) = \frac{n_1}{n} 
\]  

(9)

\[
\hat{\mu}_1 = \text{sample mean of } x^{(1)}, \ldots, x^{(n_1)}
\]

(10)

\[
\hat{\Sigma}_1 = \text{sample covariance of } x^{(1)}, \ldots, x^{(n_1)}
\]

(11)

and similarly for the other class(es)
Mixture classifier: example

- Golub et al. leukemia classification problem
  7130 ORFs (expression levels)
  38 labeled training examples,
  34 test examples

- Our mixture model (assume equal class priors)

\[
X \sim N(\mu_1, \Sigma_1), \quad \text{class} = 1 \quad (12)
\]
\[
X \sim N(\mu_0, \Sigma_0), \quad \text{class} = 0 \quad (13)
\]

Problems?
Mixture classifier: example

• Golub et al. leukemia classification problem
  7130 ORFs
  38 labeled training examples,
  34 test examples

• Our mixture model (assume equal class priors)

\[
X \sim N(\mu_1, \Sigma_1), \; \text{class} = 1 \tag{14}
\]
\[
X \sim N(\mu_0, \Sigma_0), \; \text{class} = 0 \tag{15}
\]

Problems?

• For 6,000 genes we would need to set roughly 18,000,000 parameters in each covariance matrix! (with 38 examples)
Mixture classifier: example cont’d

- The model is too complex. We need to constrain the covariance matrices
  - simple constraints (common diagonal covariance matrix)
  - more general regularization

- Let’s use the simple constraints
  1. common covariance for the two classes $\Sigma_1 = \Sigma_0$
  2. diagonal covariance matrix

$$
\Sigma = \Sigma_1 = \Sigma_2 = \begin{bmatrix}
\sigma_1^2 & \ldots & 0 \\
0 & \ldots & 0 \\
0 & \ldots & \sigma_n^2
\end{bmatrix}
$$

(16)

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?
Mixture classifier: example cont’d

• The model is too complex. We need to constrain the covariance matrices
  – simple constraints (common diagonal covariance matrix)
  – more general regularization

• Let’s use the simple constraints
  1. common covariance for the two classes $\Sigma_1 = \Sigma_0$
  2. diagonal covariance matrix

$$\Sigma = \Sigma_1 = \Sigma_2 = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ 0 & \cdots & 0 \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

(17)

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)
Mixture classifier: variable selection

- Test which genes are predictive of the class distinction
- Why is this important? Is more more information always better?
- We can test the predictive power of genes by testing if the mean expression level is different in the two class populations
- $\sigma$ is the variance of the entire population
- We assume Class 0 and Class 1 have the same variance $\sigma'$
Mixture classifier: variable 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

$H_0: X_1 \sim N(\mu, \sigma^2), X_0 \sim N(\mu, \sigma^2)$

$H_1: X_1 \sim N(\mu', \sigma'^2), X_0 \sim N(\mu'_0, \sigma'^2)$

- We can use a likelihood ratio test for this purpose

Let $\{x_i^{(t)}\}$ denote the observed expression levels for gene $i$

$$T(x_i) = 2 \cdot \log \frac{\prod_{t \in \text{class1}} P(x_i^{(t)}|\hat{\mu}_1', \hat{\sigma}'^2) \prod_{t \in \text{class0}} P(x_i^{(t)}|\hat{\mu}_0', \hat{\sigma}'^2)}{\prod_t P(x^{(t)}|\hat{\mu}, \hat{\sigma}^2)}$$

$$= n \cdot \log \frac{\hat{\sigma}^2}{\hat{\sigma}'^2}$$

(18)

where the parameter estimates are computed from the available populations in accordance with the hypothesis.

- Where does this come from?
Mixture classifier: example cont’d

• We rank the genes in the descending order of the test statistics $T(x_i)$.

• How many genes should we include?
Mixture classifier: example cont’d

• We rank the genes in the descending order of the test statistics $T(x_i)$.

• How many genes should we include?

• We include all the genes for which the associated p-value of the test statistic is less than $1/m$, where $m$ is the number of genes

• This ensures that we get on average only 1 erroneous predictor (gene) after applying the test for all the genes
Mixture classifier: example cont’d

- We rank the genes in the **descending** order of the test statistics $T(x_i)$.

- **How many genes should we include?**

- We include all the genes for which the associated p-value of the test statistic is less than $1/m$, where $m$ is the number of genes.

- This ensures that we get on average only 1 erroneous predictor (gene) after applying the test for all the genes.

  In the Golub et al. problem, we get 187 genes, and only 1 test error (out of 34)

- **How many genes do we really need?**
Mixture classifier: example cont’d

Only a few genes are necessary for making accurate class distinctions.
Mixture classifier: example cont’d

The figure shows the value of the discriminant function

\[ f(X) = \log \frac{P(X | \hat{\mu}_1', \hat{\sigma}^2)}{P(X | \hat{\mu}_0', \hat{\sigma}^2)} \]  

across the test examples

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