GENOMIC SIGNAL PROCESSING

Lecture 2

Classification of disease subtype based on microarray data

1. Analysis of microarray data (see last 15 slides of Lecture 1)
2. Classification methods for microarray data
3. Discrimination analysis by linear discrimination
4. A case study: clasification of ALL/AML Leukemia

2. Classification methods for microarray data

- Unsupervised learning = Learning without a teacher = Cluster analysis
- Supervised learning = Learning with a teacher = Discriminant analysis

Classification is the name used mostly for supervised learning.

Examples of supervised learning:

1. You measure the spectrum of a frame a speech, and compute 24 averages over fixed subbands, and form with them the vector of 24 spectra coefficients, called a feature vector. Each frame is labeled with 1 if it was spoken by a boy, with a 0 if it was spoken by a girl. If enough examples of frames and their corresponding labels are available, one can find a rule by which to guess boy or girl, for each frame, given the vector of 24 spectral coefficients.

2. You have a vector of 1000 gene expressions for each patient suffering of Leukemia. Leukemia has two major subtypes: ALL (acute lymphocytic leukemia) and AML (acute myelogenous leukemia). A number of examples of patients with ALL and AML are given, together with the gene expressions for each patient. It is required to find a rule by which to diagnose ALL or AML based on the expressions of the genes.

The problem of discriminant analysis (or supervised learning):

- The multivariate observations for each individual form the feature vector. Each individual has associated a class label. In the training set all individuals have known class labels, and the classification rules are learned (or designed) over the training examples. In the testing set the class labels are not used for determining the correct rule, only to estimate the performance (errors).
- Evaluation of the performance
  - error rate = proportion of items missclassified
  - optimum error rate= the rate which will hold if all parameters of a statistical model for the class labels and feature vectors are known.

- apparent error rate = the rate we obtain by resubstituting the training samples and determining the missclassifications
- Overall error rate versus group error rate (sometimes we don't like the global optimum, which may be poor for some groups)

Error rates

- Apparent rate: classify the training samples by the rule derived from the training set itself. The estimator is overly optimistic if the sample size is not much larger than the number of variables.
- Leave-one-out Omit one observation, recalculate the classification rule from all other observations, classify the deleted observation and repeat the steps for each observation in turn. Counting the errors of missclassification yields almost unbiased estimate of the error rate. However the variance is high (the missclasifications are correlated).
- Crossvalidation is like leave-one-out, but now split the sample in k groups, use k-1 of them to obtain the classification rule and then classify the remaining group.

Variable selection

- Select a small number of variables relative to the sizes of the two training sets.
- Variables which are known to be highly correlated can be replaced by just one variable.
- Variables included are those whose scaled between group distances are the largest.
- In a forward program variables are included one at a time, based on which of them decreases the error rate the most.
- In a backward program one begins with the entire set and at each step drops the variable that increases the error rate the least.
- A simplistic plan: select first variables based on their ratio BSS/WSS (see linear discrimination part), and then perform a stepwise selection.

3. Discrimination analysis by linear discrimination

We discuss first the two-class problems (only two class labels).

- Let $\underline{X}_i$ be the feature (measurement) vector of cell i, and $D_i \in \{0, 1\}$ be the class of cell i.
- A useful characterization of the two classes, by the average over the class. Let $\underline{m}_0$ be the mean vector of the class 0

$$\underline{m}_0 = \frac{1}{|\{i|D_i = 0\}|} \sum_{i|D_i=0} \underline{X}_i \tag{1}$$

and m_1 be the mean vector of the class 1

$$\underline{m}_1 = \frac{1}{|\{i|D_i = 1\}|} \sum_{i|D_i=1} \underline{X}_i \tag{2}$$

The goal is to find a vector $\underline{a}$ such the scalar $\underline{a}^T \underline{X}_i$ is compared against $\underline{a}^T \underline{m}_0$ and $\underline{a}^T \underline{m}_1$ and is classified according to the nearest of them.

$$Y_i = \begin{cases} 0 & if \quad |\underline{a}^T \underline{X}_i - \underline{a}^T \underline{m}_0|^2 < |\underline{a}^T \underline{X}_i - \underline{a}^T \underline{m}_1|^2 \\ 1 & else \end{cases} \tag{3}$$

The discrimination equation is

$$(\underline{a}^T \underline{X}_i)^2 - 2(\underline{a}^T \underline{X}_i)(\underline{a}^T \underline{m}_0) + (\underline{a}^T \underline{m}_0)^2 < (\underline{a}^T \underline{X}_i)^2 - 2(\underline{a}^T \underline{X}_i)(\underline{a}^T \underline{m}_1) + (\underline{a}^T \underline{m}_1)^2$$

$$2(\underline{a}^T\underline{X}_i)\underline{a}^T(\underline{m}_0-\underline{m}_1)<(\underline{a}^T\underline{m}_1)^2-(\underline{a}^T\underline{m}_0)^2$$
$$2(\underline{a}^T\underline{X}_i)<\frac{(\underline{a}^T\underline{m}_1)^2-(\underline{a}^T\underline{m}_0)^2}{\underline{a}^T(\underline{m}_0-\underline{m}_1)}=\underline{a}^T(\underline{m}_0+\underline{m}_1)$$
$$\underline{a}^T(\underline{X}_i-\frac{\underline{m}_0+\underline{m}_1}{2})<0 \tag{4}$$

Thus the classifier has the equivalent form

$$Y_i=\begin{cases}0 & if \quad \underline{a}^T(\underline{X}_i-\frac{\underline{m}_0+\underline{m}_1}{2})<0\\ 1 & else\end{cases} \tag{5}$$

We would certainly like to maximize the number of correct classifications

$$\max_a Card\{i|\underline{a}^T(\underline{X}_i-\frac{\underline{m}_0+\underline{m}_1}{2})<0;D_i=0\}+Card\{i|\underline{a}^T(\underline{X}_i-\frac{\underline{m}_0+\underline{m}_1}{2})\geq 0;D_i=1\} \tag{6}$$

but this problem has no closed form solution.

Since we cannot maximize directly the number of correct classifications we consider the following, easier, problem.

Heuristic problem: Fisher's linear discriminant

We like to find two centers, $\underline{a}^T\underline{m}_0$ and $\underline{a}^T\underline{m}_1$, such that the data in class 0 is the least scattered around $\underline{a}^T\underline{m}_0$, and data in class 1 is the least scattered around $\underline{a}^T\underline{m}_1$, and moreover the centers are as further apart as possible. All these requirements can be written in one single criterion to be maximized,

$$J(\underline{a}) = \frac{(\underline{a}^T\underline{m}_0 - \underline{a}^T\underline{m}_1)^2}{\sigma_0^2 + \sigma_1^2} \tag{7}$$

where the variances are

$$\begin{aligned}
\sigma_0^2 &= \sum_{i|D_i=0} (\underline{a}^T\underline{X}_i - \underline{a}^T\underline{m}_0)^2 = \sum_{i|D_i=0} \underline{a}^T(\underline{X}_i - \underline{m}_0)(\underline{X}_i - \underline{m}_0)^T\underline{a} = \underline{a}^T S_0 \underline{a} \\
\sigma_1^2 &= \sum_{i|D_i=1} (\underline{a}^T\underline{X}_i - \underline{a}^T\underline{m}_1)^2 = \sum_{i|D_i=1} \underline{a}^T(\underline{X}_i - \underline{m}_1)(\underline{X}_i - \underline{m}_1)^T\underline{a} = \underline{a}^T S_1 \underline{a}
\end{aligned}$$

and therefore

$$J(\underline{a}) = \frac{(\underline{a}^T\underline{m}_0 - \underline{a}^T\underline{m}_1)^2}{\sigma_0^2 + \sigma_1^2} = \frac{\underline{a}^T(\underline{m}_0 - \underline{m}_1)(\underline{m}_0 - \underline{m}_1)^T\underline{a}}{\underline{a}^T(S_0 + S_1)\underline{a}} = (\underline{a}^T B\underline{a})/(\underline{a}^T W\underline{a}) \tag{8}$$

where the matrices $B = (\underline{m}_0 - \underline{m}_1)(\underline{m}_0 - \underline{m}_1)^T$ and $W = S_0 + S_1$ are known.

Case A: Nonsingular Within Class Covariance matrix

To solve the maximization problem, suppose $W = S_0 + S_1 = W^{T/2}W^{1/2}$ is positive definite and make the change of variable $\underline{\theta} = W^{1/2}\underline{a}$ ($\underline{a} = W^{-1/2}\underline{\theta}$) to get

$$J(\underline{a}) = \frac{\underline{a}^T B \underline{a}}{\underline{a}^T W \underline{a}} = \frac{\underline{\theta}^T W^{-T/2} B W^{-1/2} \underline{\theta}}{\underline{\theta}^T \underline{\theta}} \tag{9}$$

The maximizing $\underline{\theta}$ satisfies

$$\begin{aligned} \underline{\theta}^* &= \arg \max_{||\underline{\theta}||=1} \underline{\theta}^T W^{-T/2} B W^{-1/2} \underline{\theta} \\ \underline{\theta}^*, \mu^* &= \arg \max_{\underline{\theta}, \mu} \theta^T Q \underline{\theta} + \mu(1 - \underline{\theta}^T \underline{\theta}) \end{aligned} \tag{10}$$

where we denote $W^{-T/2} B W^{-1/2} = Q$. Now the Lagrangian is minimized when $2Q\underline{\theta} - 2\mu\underline{\theta} = 0$, which shows that $\underline{\theta}$ must be a eigenvector of Q and μ is its corresponding eigenvalue. It is obvious that the maximization is realized by taking $\underline{\theta}$ as the (unit norm) eigenvector corresponding to the largest egenvalue μ^* of $W^{-T/2} B W^{-1/2} = Q$, when the criterion

$$J(\underline{\theta}) = \underline{\theta}^T Q \underline{\theta} = \mu^* \tag{11}$$

Finally, the linear combination vector is

$$\underline{a}^* = W^{-1/2}\underline{\theta}^* = W^{-1/2} \cdot \text{max eigenvector of}(W^{-T/2}BW^{-1/2}) \qquad (12)$$

Simplifications:

We may simplify the solution: $W^{-T/2}BW^{-1/2}\underline{\theta}^* = \mu^*\underline{\theta}^*$ implies $W^{-1/2}W^{-T/2}BW^{-1/2}\underline{\theta}^* = W^{-1/2}\mu^*\underline{\theta}^*$ or $W^{-1}B\underline{a}^* = \mu^*\underline{a}^*$, therefore μ^* is the largest eigenvalue of $W^{-1}B$ and $\underline{a}^*$ its corresponding eigenvector.

For a two class experiment the solution simplifies even more. The eigenvector obeys

$$W^{-1}(\underline{m}_0 - \underline{m}_1)(\underline{m}_0 - \underline{m}_1)^T\underline{a}^* = \mu^*\underline{a}^*$$
$$W^{-1}(\underline{m}_0 - \underline{m}_1) = \frac{\mu^*}{(\underline{m}_0 - \underline{m}_1)^T\underline{a}^*}\underline{a}^*$$

therefore

$$\underline{a}^* = W^{-1}(\underline{m}_0 - \underline{m}_1) \qquad (13)$$

or any scaled version of it.

Case B: Singular Within Class Covariance matrix

To solve the maximization problem when $W = S_0 + S_1$ is singular consider the over-parameterization $\underline{a} = s\underline{a}_s$, such that $\underline{a}_s^T W \underline{a}_s = C$ with a constant scalar C, which represent all possible $\underline{a}$ vectors. Observe that $J(\underline{a}) = \frac{\underline{a}^T B \underline{a}}{\underline{a}^T W \underline{a}} = \frac{s\underline{a}_s^T B s\underline{a}_s}{s\underline{a}_s^T W s\underline{a}_s} = \underline{a}_s^T B \underline{a}_s / C$ and construct the Lagrangian

$$J(\underline{a}_s, \mu) = \underline{a}_s^T B \underline{a}_s + \mu(C - \underline{a}_s^T W \underline{a}_s) \tag{14}$$

which has an extremum when $2B\underline{a}_s - 2\mu W \underline{a}_s = 0$, which tells that all Lagrangian extremum points are at those $\underline{a}_s$ which are scaled generalized eigenvectors of (B, W). Recall that the generalized eigenvector decomposition of (B, W) obeys $BV = WV\Lambda$ where Λ is diagonal and V is a square matrix whose columns are generalized eigenvectors, each eigenvector $\underline{v}_i$ satisfying $B\underline{v}_i = \lambda_i W \underline{v}_i$.

Denote $\underline{a}_s^*$ any scaled generalized eigenvector satisfying $B\underline{a}_s^* - \lambda_1 W \underline{a}_s^* = 0$ and $C = \underline{a}_s^{*T} W \underline{a}_s^*$, then $J(\underline{a}_s^*, \lambda_1) = \underline{a}_s^{*T} B \underline{a}_s^*$, which can be evaluated by left-multiplying $B\underline{a}_s^* = \lambda_1 W \underline{a}_s^*$ with $\underline{a}_s^{*T}$, to get $J(\underline{a}_s^*, \lambda_1) = \underline{a}_s^{*T} B \underline{a}_s^* = \lambda_1 \underline{a}_s^{*T} W \underline{a}_s^* = \lambda_1 C$. Therefore the maximum criterion $J(\underline{a}_s^*, \lambda_1)$ is obtained when λ_{max} is the generalized eigenvalue of (B, W) and $\underline{a}_s^*$ is its corresponding eigenvector. We consider that

the diagonal matrix Λ is ordered such that λ_1 is the maximum generalized eigenvalue, and $\underline{v}_1$ the corresponding eigenvector. Therefore $\underline{a} = \underline{v}_1$ is the linear discrimination vector maximizing the Fischer discrimination criterion.

Summary

1. Given the feature vector $\underline{X}_i$ for each individual i, and the class label for each individual
2. Compute the average over each class $\underline{m}_0$ and $\underline{m}_1$.
3. Compute the covariance matrices

$$\begin{aligned} S_0 &= \sum_{i|D_i=0} (\underline{X}_i - \underline{m}_1)(\underline{X}_i - \underline{m}_1)^T \\ S_1 &= \sum_{i|D_i=1} (\underline{X}_i - \underline{m}_1)(\underline{X}_i - \underline{m}_1)^T \end{aligned}$$

4. Construct the matrix $W = S_0 + S_1$ and take the optimum discriminator as $\underline{a}^* = W^{-1}(\underline{m}_0 - \underline{m}_1)$.
5. Use the discriminator $\underline{a}^*$ to classify a feature vector X_i as follows

$$Y_i = \begin{cases} 0 & if \quad \underline{a}^{*T}(\underline{X}_i - \frac{\underline{m}_0 + \underline{m}_1}{2}) < 0 \\ 1 & else \end{cases} \tag{15}$$

How good is one gene as a feature? A simple indicator: BSS/WSS

- When the feature vector is just a scalar (it is composed of a single gene expression) the discrimination criterion reduces to

$$J(\underline{a}) = (\underline{a}^T B \underline{a})/(\underline{a}^T W \underline{a}) = B/W \tag{16}$$

where the scalar $B = (m_0 - m_1)^2$ is the spread between the centers of the classes and $W = S_0 + S_1 = \Sigma_{i|D_i=0}(X_i - m_0)^2 + \Sigma_{i|D_i=1}(X_i - m_1)^2$ is the within class spread. $J(\underline{a})$ is proportional to the quantity BSS/WSS defined below.

- Let denote N_0 the number of members of class 0, N_1 the number of members of class 1, m the overall mean and N the total number of individuals. The following sums of squares can be defined:

$$\begin{aligned} TSS &= \sum_i (X_i - m)^2 \\ BSS &= N_0(m_0 - m)^2 + N_1(m_1 - m)^2 \\ WSS &= W = \sum_{i|D_i=0} (X_i - m_0)^2 + \sum_{i|D_i=1} (X_i - m_1)^2 \end{aligned} \tag{17}$$

BSS is called between sum of squares; TSS is the total sum of squares; $WSS = W =$ is the within sum of squares. We have $TSS = WSS + BSS$.

- Due to the identity $\frac{N_0 N_1}{N}(m_0 - m_1)^2 = N_0(m_0 - m)^2 + N_1(m_1 - m)^2 = BSS$ we have so $BSS/WSS = B/W\frac{N_0 N_1}{N} = \frac{N_0 N_1}{N}J(\underline{a})$. Thus BSS/WSS gives a good indication of the discrimination capabilities of a scalar feature.

An example: MIT Leukemia data set

The data file

The file 1709.mat contains the matrix M preprocessed with process1.m (floor 100 & ceiling 16000 for each element; exclusion $max/min <= 5$ & $(max - min) <= 500$; log10). The vector v is created such that (1 if ALL and 0 if AML).

Some statistics

```
min(min(M))=0; max(max(M))=4.5791; mean(mean(M))=2.8433;
median(median(M))=2.8156;

v1=M(:); median(v1)=2.8116; imagesc([ones(100,1)*v'*max(max(M)) ; M])
colormap(gray)
```

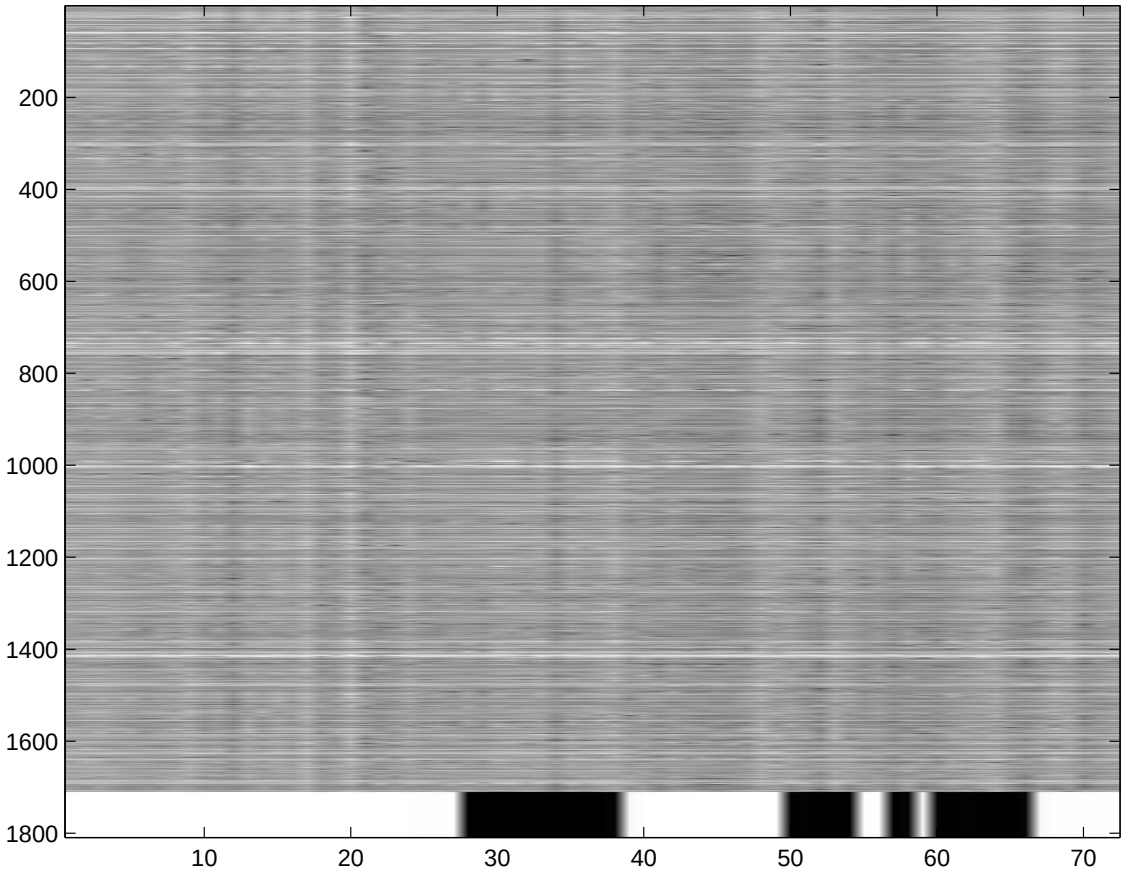

Figure 1: The dataset MitLeukemia preprocessed to keep 1709 gene expressions of 72 cell types. The bottom white/black regions show the known classification of the cells: Cell type 1=ALL (white); Cell type 0=ALM (black).

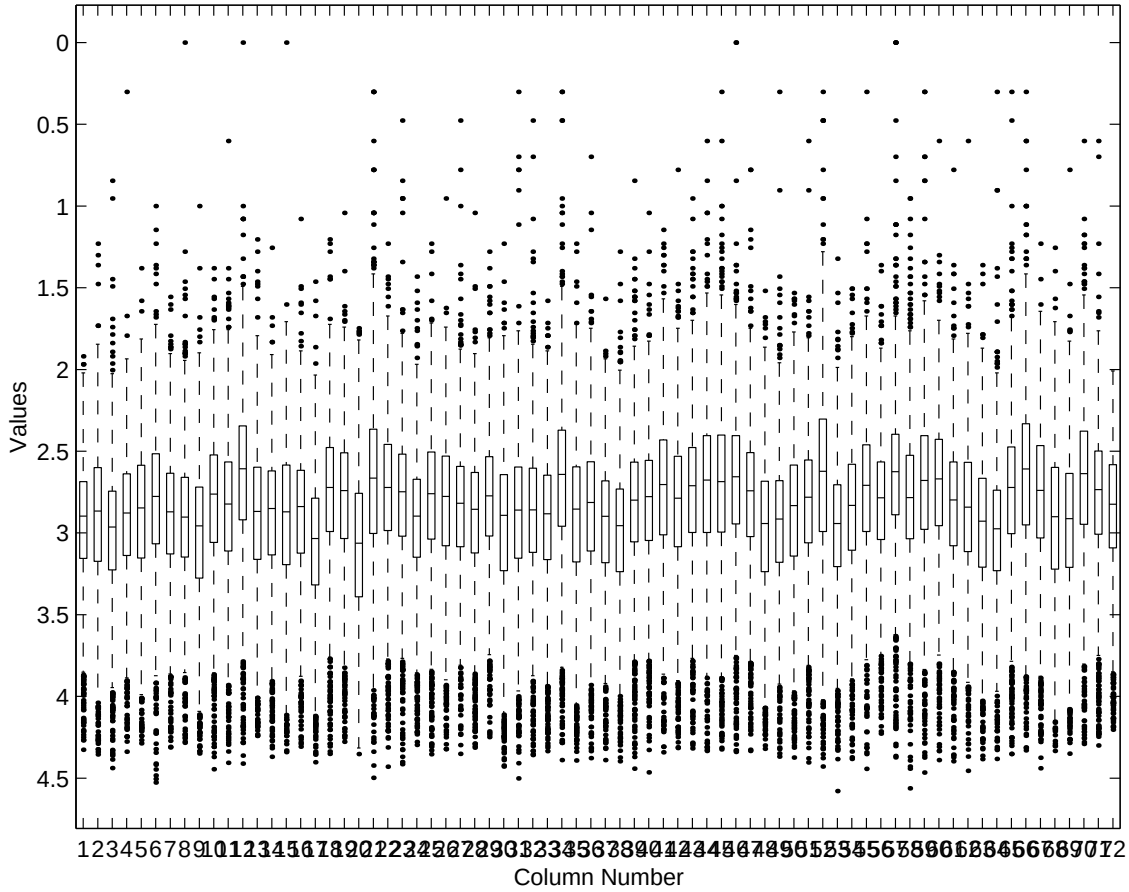

Figure 2: Boxplot of the dataset MitLeukemia preprocessed to keep 1709 gene expressions of 72 cell types. BOXPLOT(X,NOTCH,SYM,VERT,WHIS) produces a box and whisker plot for each column of X. The box has lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers.

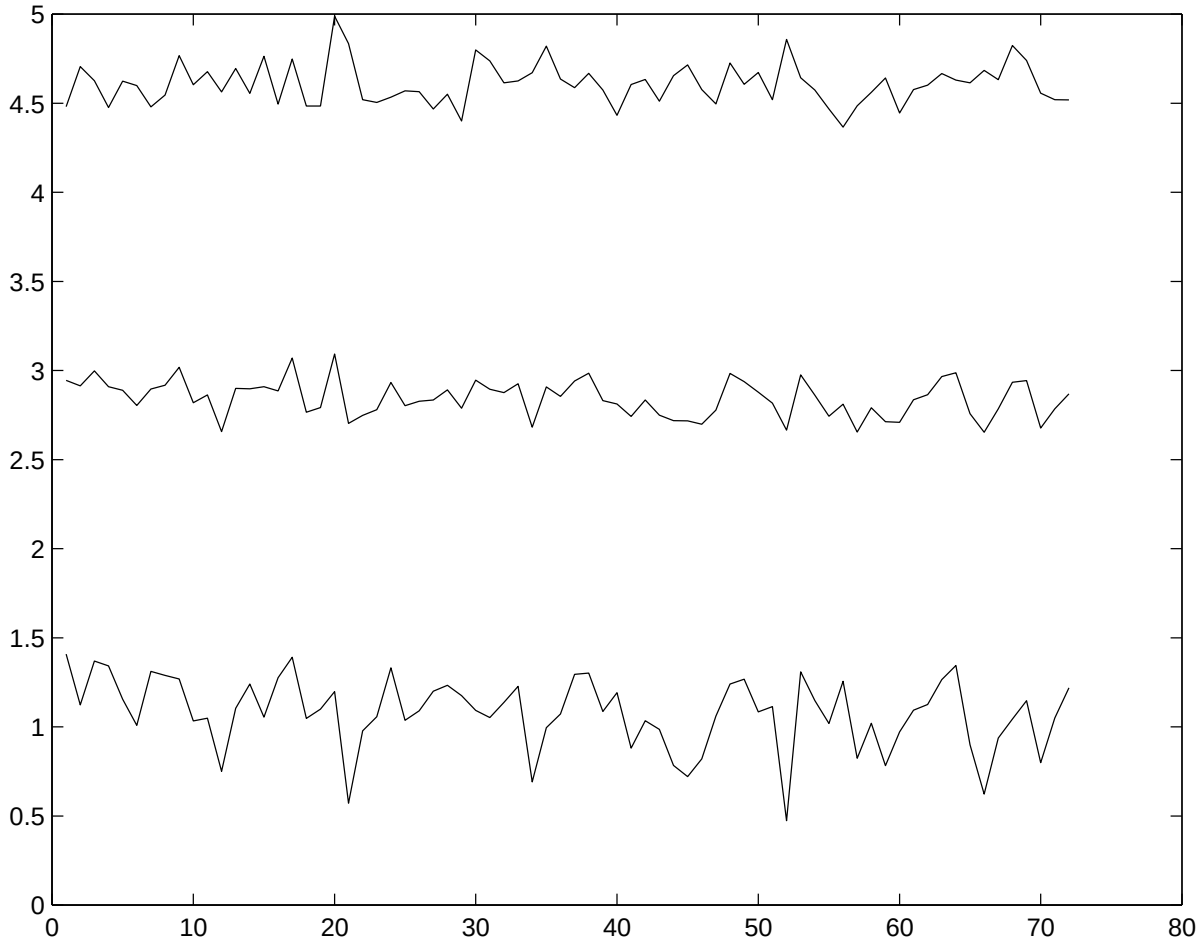

Figure 3: Mean plus/minus four times the standard deviation of the dataset MitLeukemia preprocessed to keep 1709 gene expressions of 72 cell types.

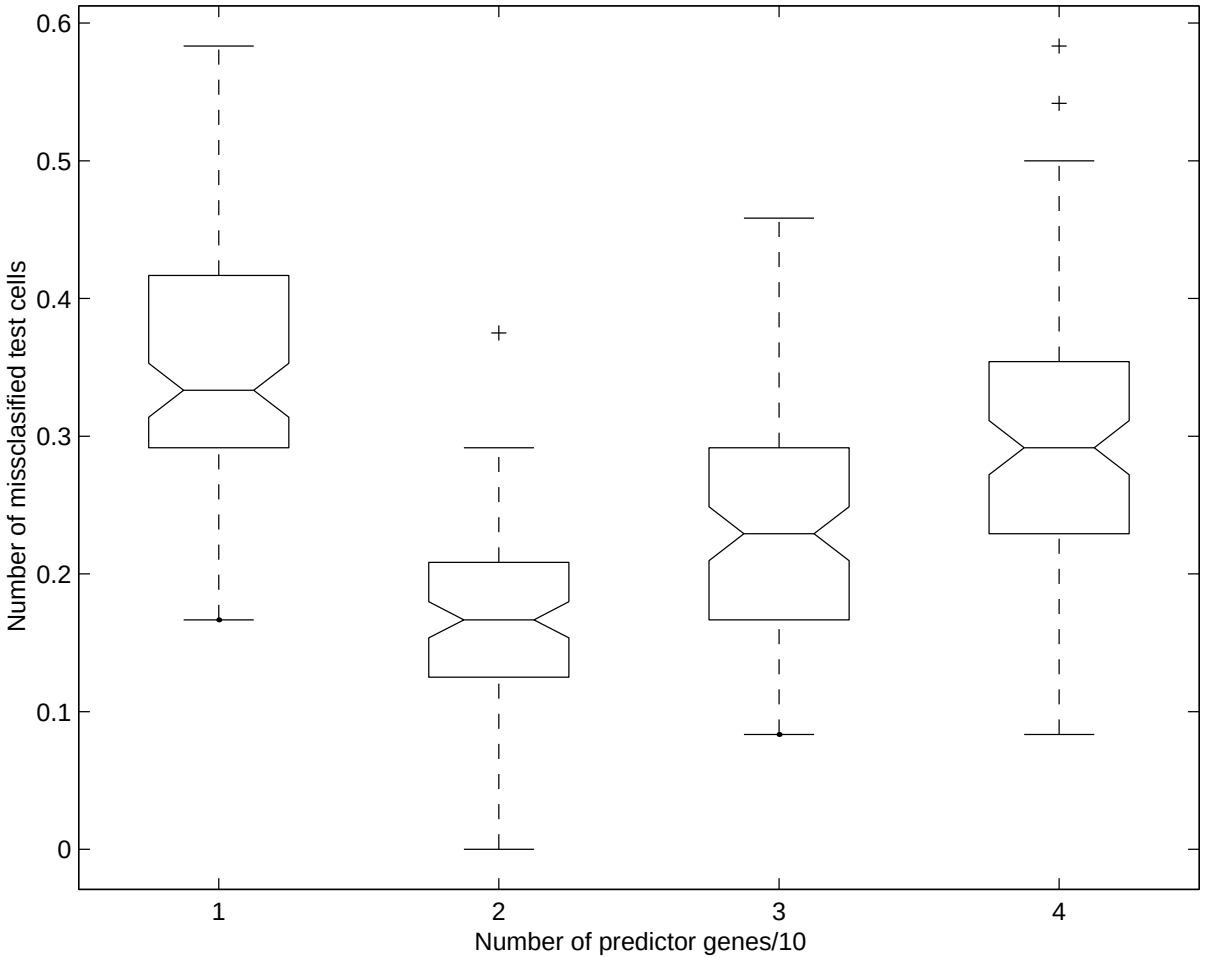

Figure 4: Boxplot of missclassifications in 100 runs of 2:1 sampling scheme (48 training cells and 24 test cells). The predictor genes are the first in the dataset, (1:p), with p=10,20,30,40.

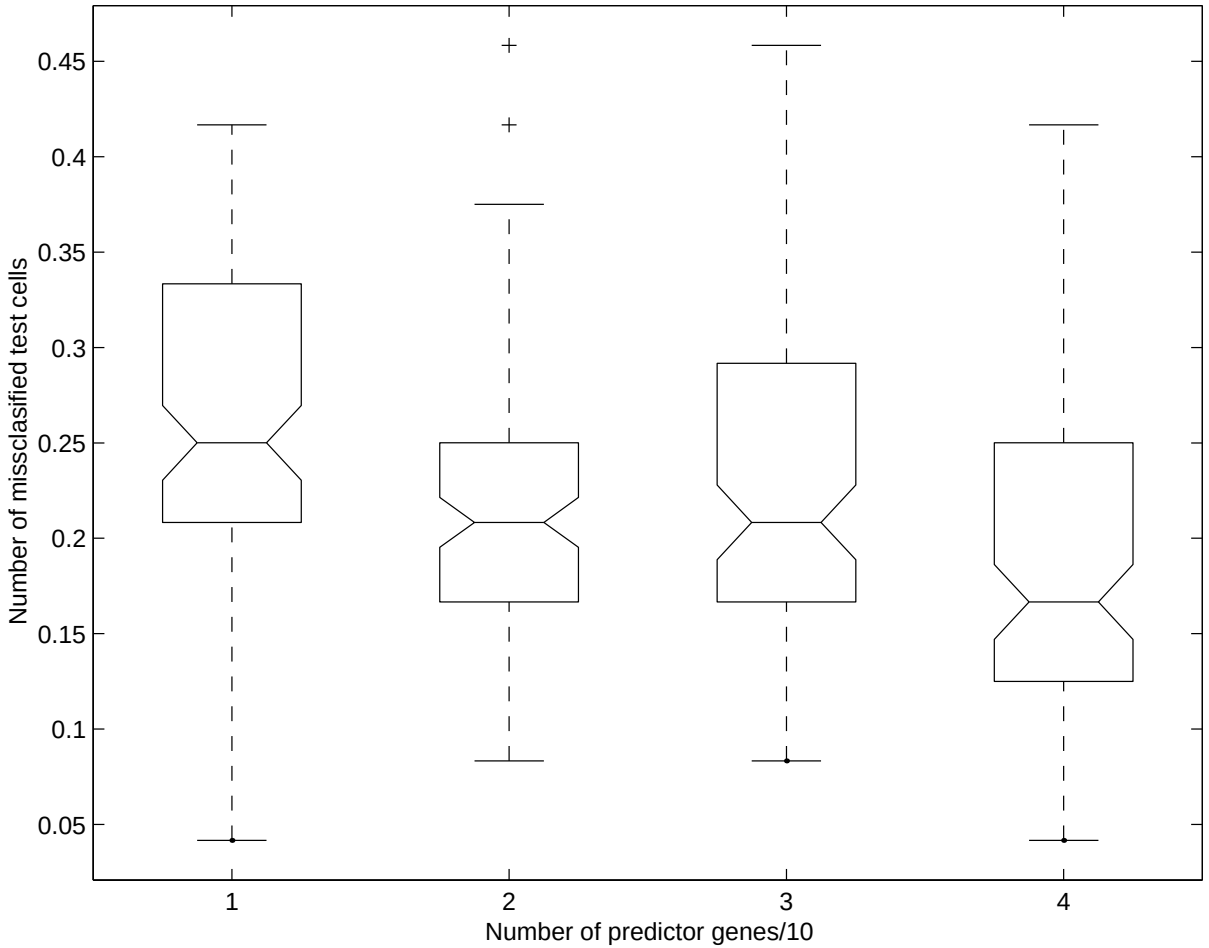

Figure 5: Boxplot of missclassifications in 100 runs of 2:1 sampling scheme (48 training cells and 24 test cells). The predictor genes are taken at the position 100+(1:p) with p=10,20,30,40, to compare random choices (quite similar with the previous figure].

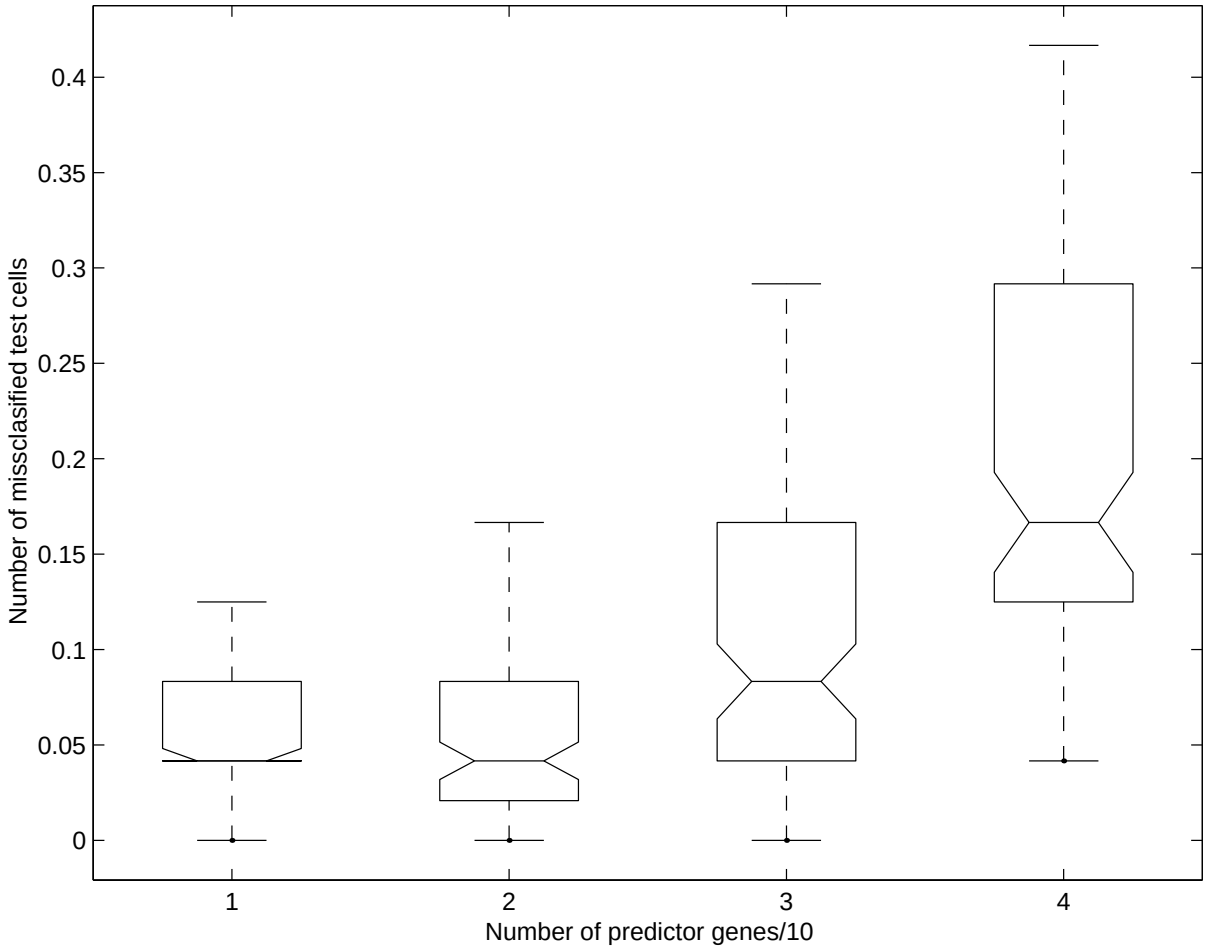

Figure 6: A good feature selection: Boxplot of missclassifications in 100 runs of 2:1 sampling scheme (48 training cells and 24 test cells). The predictor genes are arranged in decreasing order of their BSS/WSS, and p=10,20,30,40.

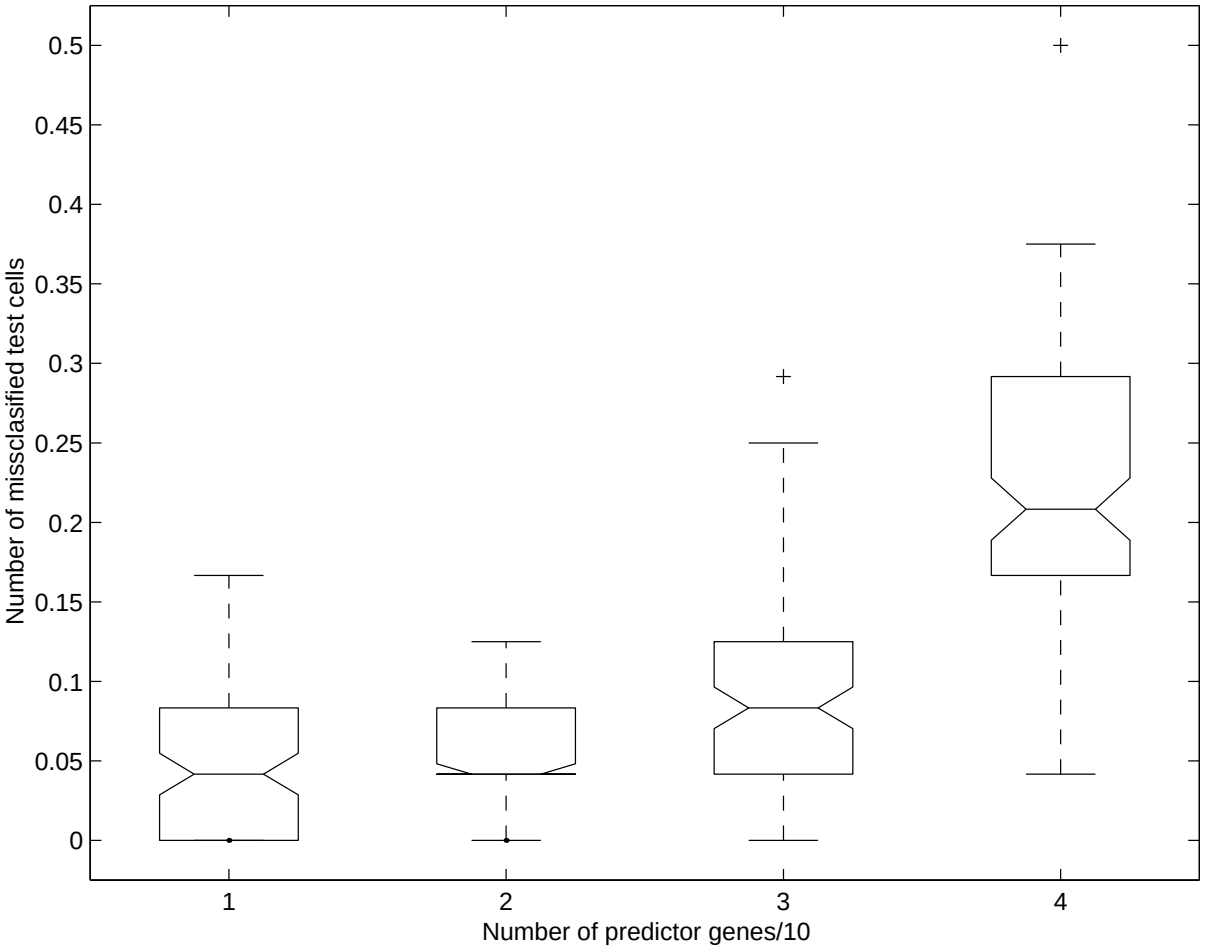

Figure 7: A good feature selection: another realization of the crossvalidation split. Boxplot of missclassifications in 100 runs of 2:1 sampling scheme (48 training cells and 24 test cells). The predictor genes are arranged in decreasing order of their BSS/WSS, and p=10,20,30,40.

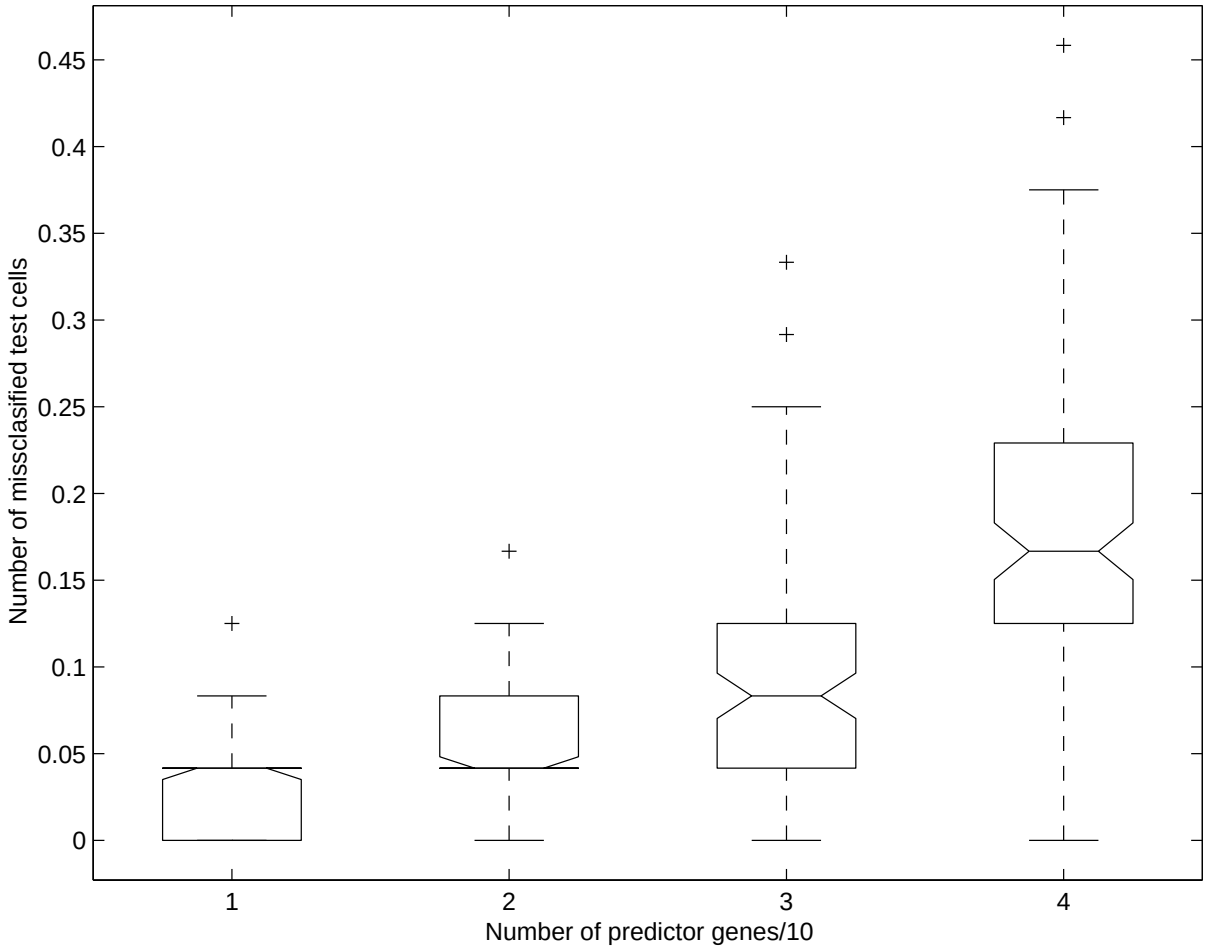

Figure 8: A very good feature selection (without the normalization to unit variance across the genes). Boxplot of missclassifications in 100 runs of 2:1 sampling scheme (48 training cells and 24 test cells). The predictor genes are arranged in decreasing order of their BSS/WSS, and p=10,20,30,40.

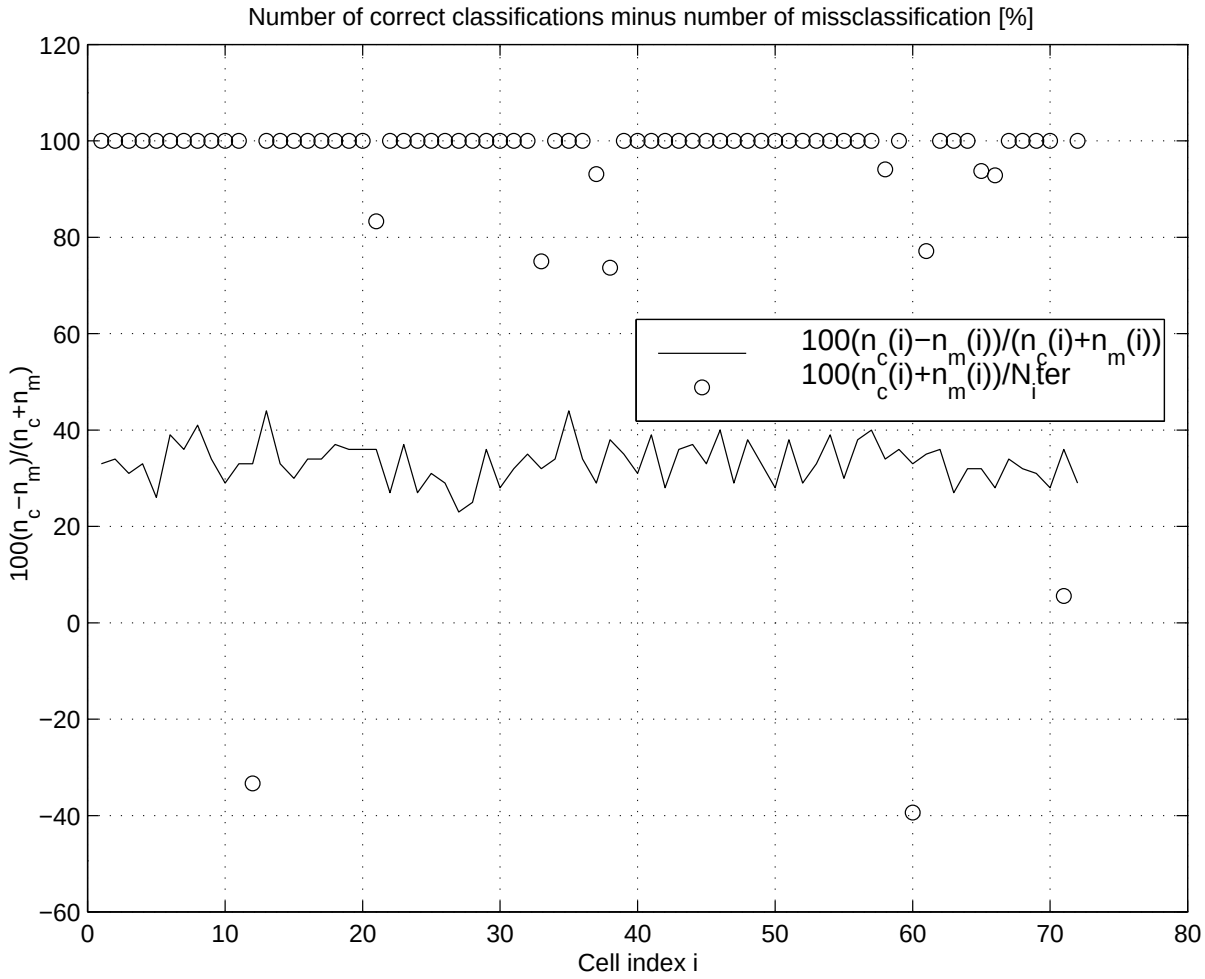

Figure 9: Individual rates of missclassifications for each cell. Three cells are systematically missclassified.

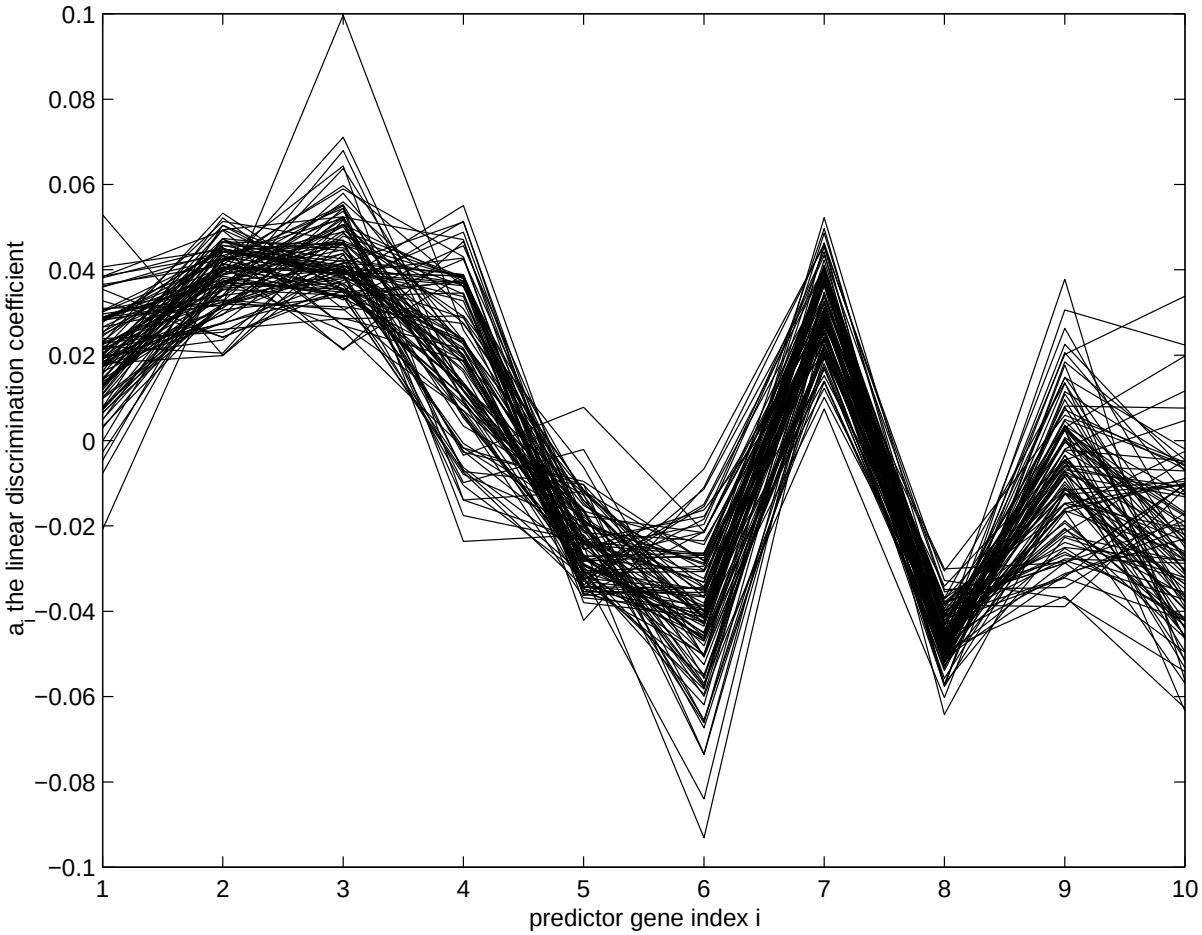

Figure 10: The linear discriminator in 100 runs.