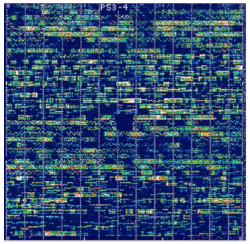
Classification

Methods Course: Gene Expression Data Analysis

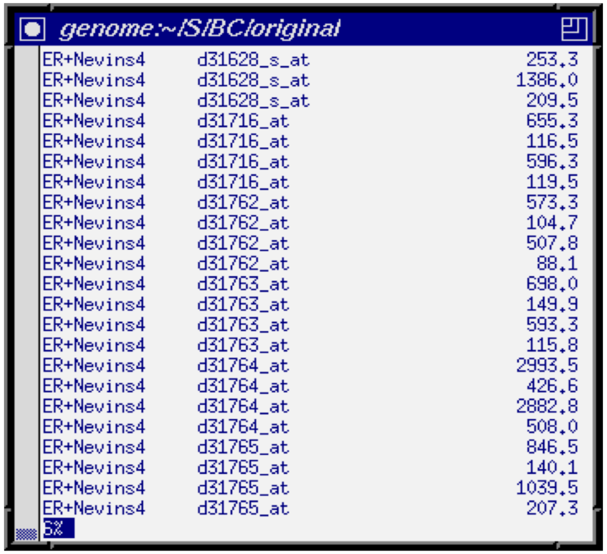
-Day Five –

Rainer Spang

Ms. Smith



DNA Chip of Ms. Smith

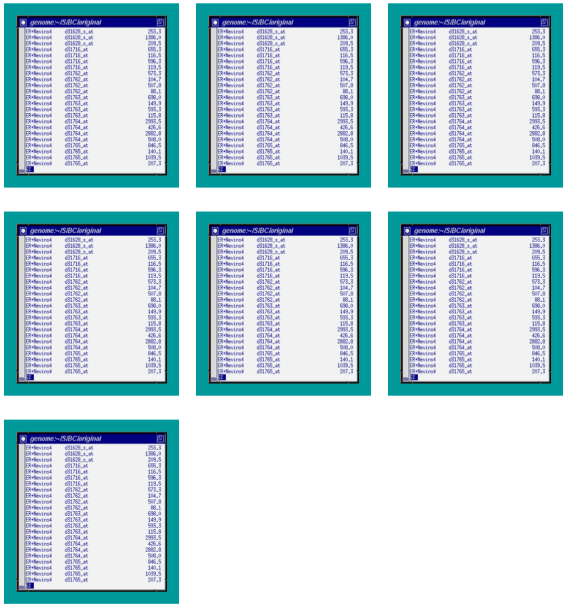


```
ER+Nevins4      d31628_s_at                    253.3
ER+Nevins4      d31628_s_at                   1386.0
ER+Nevins4      d31628_s_at                    209.5
ER+Nevins4      d31716_at                      655.3
ER+Nevins4      d31716_at                      116.5
ER+Nevins4      d31716_at                      596.3
ER+Nevins4      d31716_at                      119.5
ER+Nevins4      d31762_at                      573.3
ER+Nevins4      d31762_at                      104.7
ER+Nevins4      d31762_at                      507.8
ER+Nevins4      d31762_at                       88.1
ER+Nevins4      d31763_at                      698.0
ER+Nevins4      d31763_at                      149.9
ER+Nevins4      d31763_at                      593.3
ER+Nevins4      d31763_at                      115.8
ER+Nevins4      d31764_at                     2993.5
ER+Nevins4      d31764_at                      426.6
ER+Nevins4      d31764_at                     2882.8
ER+Nevins4      d31764_at                      508.0
ER+Nevins4      d31765_at                      846.5
ER+Nevins4      d31765_at                      140.1
ER+Nevins4      d31765_at                     1039.5
ER+Nevins4      d31765_at                      207.3
5%
```

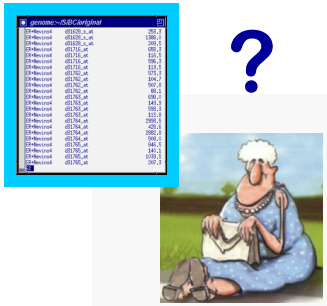
Expression profile of Ms. Smith

Ms. Smith

Looking for similarities

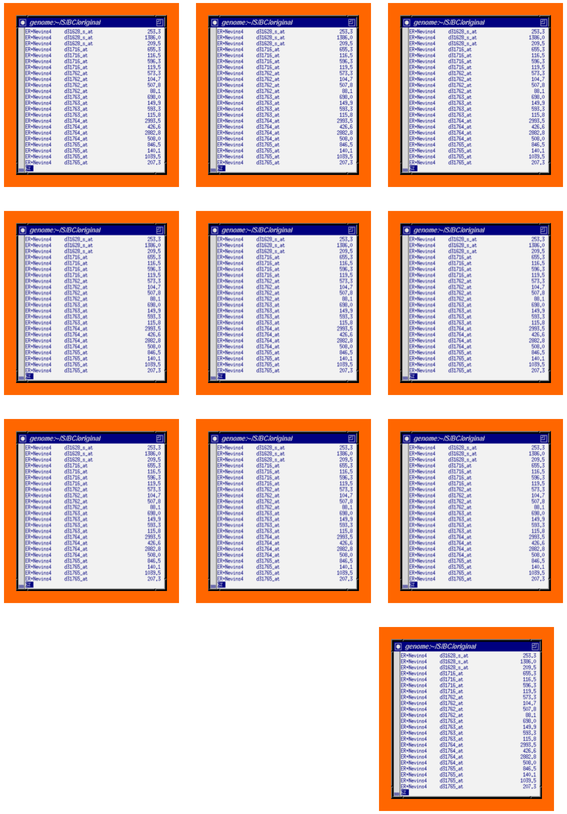


```
ER+Nevins4      d31628_s_at           253.3
ER+Nevins4      d31628_s_at          1386.0
ER+Nevins4      d31628_s_at           209.5
ER+Nevins4      d31716_at             655.3
ER+Nevins4      d31716_at             116.5
ER+Nevins4      d31716_at             596.3
ER+Nevins4      d31716_at             119.5
ER+Nevins4      d31762_at             573.3
ER+Nevins4      d31762_at             104.7
ER+Nevins4      d31762_at             507.8
ER+Nevins4      d31762_at              88.1
ER+Nevins4      d31763_at             698.0
ER+Nevins4      d31763_at             149.9
ER+Nevins4      d31763_at             593.3
ER+Nevins4      d31763_at             115.8
ER+Nevins4      d31764_at            2993.5
ER+Nevins4      d31764_at             426.6
ER+Nevins4      d31764_at            2882.8
ER+Nevins4      d31764_at             508.0
ER+Nevins4      d31765_at             846.5
ER+Nevins4      d31765_at             140.1
ER+Nevins4      d31765_at            1039.5
ER+Nevins4      d31765_at             207.3
$%
```



```
ER+Nevins4      d31628_s_at           253.3
ER+Nevins4      d31628_s_at          1386.0
ER+Nevins4      d31628_s_at           209.5
ER+Nevins4      d31716_at             655.3
ER+Nevins4      d31716_at             116.5
ER+Nevins4      d31716_at             596.3
ER+Nevins4      d31716_at             119.5
ER+Nevins4      d31762_at             573.3
ER+Nevins4      d31762_at             104.7
ER+Nevins4      d31762_at             507.8
ER+Nevins4      d31762_at              88.1
ER+Nevins4      d31763_at             698.0
ER+Nevins4      d31763_at             149.9
ER+Nevins4      d31763_at             593.3
ER+Nevins4      d31763_at             115.8
ER+Nevins4      d31764_at            2993.5
ER+Nevins4      d31764_at             426.6
ER+Nevins4      d31764_at            2882.8
ER+Nevins4      d31764_at             508.0
ER+Nevins4      d31765_at             846.5
ER+Nevins4      d31765_at             140.1
ER+Nevins4      d31765_at            1039.5
ER+Nevins4      d31765_at             207.3
$%
```

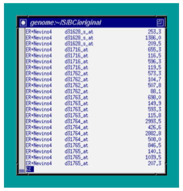
Ms. Smith



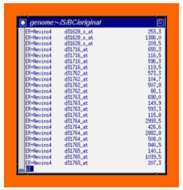
```
ER+Nevins4      d31628_s_at           253.3
ER+Nevins4      d31628_s_at          1386.0
ER+Nevins4      d31628_s_at           209.5
ER+Nevins4      d31716_at             655.3
ER+Nevins4      d31716_at             116.5
ER+Nevins4      d31716_at             596.3
ER+Nevins4      d31716_at             119.5
ER+Nevins4      d31762_at             573.3
ER+Nevins4      d31762_at             104.7
ER+Nevins4      d31762_at             507.8
ER+Nevins4      d31762_at              88.1
ER+Nevins4      d31763_at             698.0
ER+Nevins4      d31763_at             149.9
ER+Nevins4      d31763_at             593.3
ER+Nevins4      d31763_at             115.8
ER+Nevins4      d31764_at            2993.5
ER+Nevins4      d31764_at             426.6
ER+Nevins4      d31764_at            2882.8
ER+Nevins4      d31764_at             508.0
ER+Nevins4      d31765_at             846.5
ER+Nevins4      d31765_at             140.1
ER+Nevins4      d31765_at            1039.5
ER+Nevins4      d31765_at             207.3
$%
```

Compare her profile to profiles of people with tumor type A and to patients with tumor type B

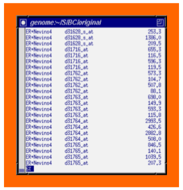
Training sets and test sets



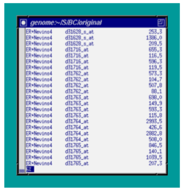
```
ER+Nevins4     d31628_s_at       253.3
ER+Nevins4     d31628_s_at      1386.0
ER+Nevins4     d31628_s_at       209.5
ER+Nevins4     d31716_at         655.3
ER+Nevins4     d31716_at         116.5
ER+Nevins4     d31716_at         596.3
ER+Nevins4     d31716_at         119.5
ER+Nevins4     d31762_at         573.3
ER+Nevins4     d31762_at         104.7
ER+Nevins4     d31762_at         507.8
ER+Nevins4     d31762_at          88.1
ER+Nevins4     d31763_at         698.0
ER+Nevins4     d31763_at         149.9
ER+Nevins4     d31763_at         593.3
ER+Nevins4     d31763_at         115.8
ER+Nevins4     d31764_at        2993.5
ER+Nevins4     d31764_at         426.6
ER+Nevins4     d31764_at        2882.8
ER+Nevins4     d31764_at         508.0
ER+Nevins4     d31765_at         846.5
ER+Nevins4     d31765_at         140.1
ER+Nevins4     d31765_at        1039.5
ER+Nevins4     d31765_at         207.3
$%
```



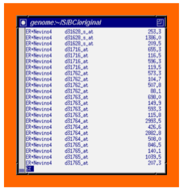
```
ER+Nevins4     d31628_s_at       253.3
ER+Nevins4     d31628_s_at      1386.0
ER+Nevins4     d31628_s_at       209.5
ER+Nevins4     d31716_at         655.3
ER+Nevins4     d31716_at         116.5
ER+Nevins4     d31716_at         596.3
ER+Nevins4     d31716_at         119.5
ER+Nevins4     d31762_at         573.3
ER+Nevins4     d31762_at         104.7
ER+Nevins4     d31762_at         507.8
ER+Nevins4     d31762_at          88.1
ER+Nevins4     d31763_at         698.0
ER+Nevins4     d31763_at         149.9
ER+Nevins4     d31763_at         593.3
ER+Nevins4     d31763_at         115.8
ER+Nevins4     d31764_at        2993.5
ER+Nevins4     d31764_at         426.6
ER+Nevins4     d31764_at        2882.8
ER+Nevins4     d31764_at         508.0
ER+Nevins4     d31765_at         846.5
ER+Nevins4     d31765_at         140.1
ER+Nevins4     d31765_at        1039.5
ER+Nevins4     d31765_at         207.3
$%
```



```
ER+Nevins4     d31628_s_at       253.3
ER+Nevins4     d31628_s_at      1386.0
ER+Nevins4     d31628_s_at       209.5
ER+Nevins4     d31716_at         655.3
ER+Nevins4     d31716_at         116.5
ER+Nevins4     d31716_at         596.3
ER+Nevins4     d31716_at         119.5
ER+Nevins4     d31762_at         573.3
ER+Nevins4     d31762_at         104.7
ER+Nevins4     d31762_at         507.8
ER+Nevins4     d31762_at          88.1
ER+Nevins4     d31763_at         698.0
ER+Nevins4     d31763_at         149.9
ER+Nevins4     d31763_at         593.3
ER+Nevins4     d31763_at         115.8
ER+Nevins4     d31764_at        2993.5
ER+Nevins4     d31764_at         426.6
ER+Nevins4     d31764_at        2882.8
ER+Nevins4     d31764_at         508.0
ER+Nevins4     d31765_at         846.5
ER+Nevins4     d31765_at         140.1
ER+Nevins4     d31765_at        1039.5
ER+Nevins4     d31765_at         207.3
$%
```

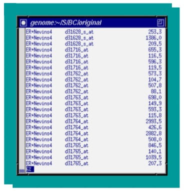


```
ER+Nevins4     d31628_s_at       253.3
ER+Nevins4     d31628_s_at      1386.0
ER+Nevins4     d31628_s_at       209.5
ER+Nevins4     d31716_at         655.3
ER+Nevins4     d31716_at         116.5
ER+Nevins4     d31716_at         596.3
ER+Nevins4     d31716_at         119.5
ER+Nevins4     d31762_at         573.3
ER+Nevins4     d31762_at         104.7
ER+Nevins4     d31762_at         507.8
ER+Nevins4     d31762_at          88.1
ER+Nevins4     d31763_at         698.0
ER+Nevins4     d31763_at         149.9
ER+Nevins4     d31763_at         593.3
ER+Nevins4     d31763_at         115.8
ER+Nevins4     d31764_at        2993.5
ER+Nevins4     d31764_at         426.6
ER+Nevins4     d31764_at        2882.8
ER+Nevins4     d31764_at         508.0
ER+Nevins4     d31765_at         846.5
ER+Nevins4     d31765_at         140.1
ER+Nevins4     d31765_at        1039.5
ER+Nevins4     d31765_at         207.3
$%
```



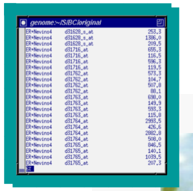
```
ER+Nevins4     d31628_s_at       253.3
ER+Nevins4     d31628_s_at      1386.0
ER+Nevins4     d31628_s_at       209.5
ER+Nevins4     d31716_at         655.3
ER+Nevins4     d31716_at         116.5
ER+Nevins4     d31716_at         596.3
ER+Nevins4     d31716_at         119.5
ER+Nevins4     d31762_at         573.3
ER+Nevins4     d31762_at         104.7
ER+Nevins4     d31762_at         507.8
ER+Nevins4     d31762_at          88.1
ER+Nevins4     d31763_at         698.0
ER+Nevins4     d31763_at         149.9
ER+Nevins4     d31763_at         593.3
ER+Nevins4     d31763_at         115.8
ER+Nevins4     d31764_at        2993.5
ER+Nevins4     d31764_at         426.6
ER+Nevins4     d31764_at        2882.8
ER+Nevins4     d31764_at         508.0
ER+Nevins4     d31765_at         846.5
ER+Nevins4     d31765_at         140.1
ER+Nevins4     d31765_at        1039.5
ER+Nevins4     d31765_at         207.3
$%
```

Use the training samples ...

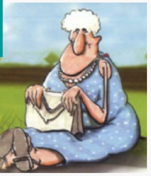


```
ER+Nevins4     d31628_s_at       253.3
ER+Nevins4     d31628_s_at      1386.0
ER+Nevins4     d31628_s_at       209.5
ER+Nevins4     d31716_at         655.3
ER+Nevins4     d31716_at         116.5
ER+Nevins4     d31716_at         596.3
ER+Nevins4     d31716_at         119.5
ER+Nevins4     d31762_at         573.3
ER+Nevins4     d31762_at         104.7
ER+Nevins4     d31762_at         507.8
ER+Nevins4     d31762_at          88.1
ER+Nevins4     d31763_at         698.0
ER+Nevins4     d31763_at         149.9
ER+Nevins4     d31763_at         593.3
ER+Nevins4     d31763_at         115.8
ER+Nevins4     d31764_at        2993.5
ER+Nevins4     d31764_at         426.6
ER+Nevins4     d31764_at        2882.8
ER+Nevins4     d31764_at         508.0
ER+Nevins4     d31765_at         846.5
ER+Nevins4     d31765_at         140.1
ER+Nevins4     d31765_at        1039.5
ER+Nevins4     d31765_at         207.3
$%
```

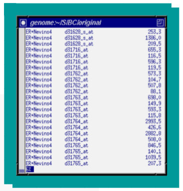
```
ER+Nevins4     d31628_s_at       253.3
ER+Nevins4     d31628_s_at      1386.0
ER+Nevins4     d31628_s_at       209.5
ER+Nevins4     d31716_at         655.3
ER+Nevins4     d31716_at         116.5
ER+Nevins4     d31716_at         596.3
ER+Nevins4     d31716_at         119.5
ER+Nevins4     d31762_at         573.3
ER+Nevins4     d31762_at         104.7
ER+Nevins4     d31762_at         507.8
ER+Nevins4     d31762_at          88.1
ER+Nevins4     d31763_at         698.0
ER+Nevins4     d31763_at         149.9
ER+Nevins4     d31763_at         593.3
ER+Nevins4     d31763_at         115.8
ER+Nevins4     d31764_at        2993.5
ER+Nevins4     d31764_at         426.6
ER+Nevins4     d31764_at        2882.8
ER+Nevins4     d31764_at         508.0
ER+Nevins4     d31765_at         846.5
ER+Nevins4     d31765_at         140.1
ER+Nevins4     d31765_at        1039.5
ER+Nevins4     d31765_at         207.3
$%
```



```
ER+Nevins4     d31628_s_at       253.3
ER+Nevins4     d31628_s_at      1386.0
ER+Nevins4     d31628_s_at       209.5
ER+Nevins4     d31716_at         655.3
ER+Nevins4     d31716_at         116.5
ER+Nevins4     d31716_at         596.3
ER+Nevins4     d31716_at         119.5
ER+Nevins4     d31762_at         573.3
ER+Nevins4     d31762_at         104.7
ER+Nevins4     d31762_at         507.8
ER+Nevins4     d31762_at          88.1
ER+Nevins4     d31763_at         698.0
ER+Nevins4     d31763_at         149.9
ER+Nevins4     d31763_at         593.3
ER+Nevins4     d31763_at         115.8
ER+Nevins4     d31764_at        2993.5
ER+Nevins4     d31764_at         426.6
ER+Nevins4     d31764_at        2882.8
ER+Nevins4     d31764_at         508.0
ER+Nevins4     d31765_at         846.5
ER+Nevins4     d31765_at         140.1
ER+Nevins4     d31765_at        1039.5
ER+Nevins4     d31765_at         207.3
$%
```



Ms. Smith



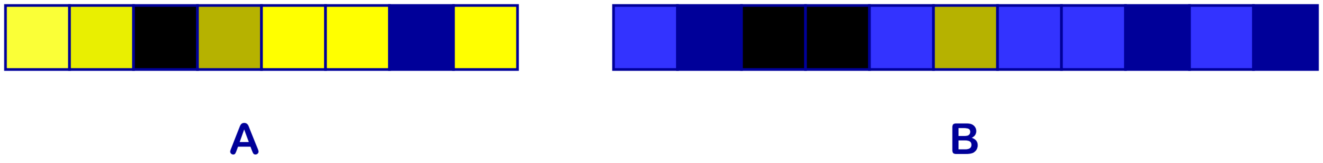
```
ER+Nevins4     d31628_s_at       253.3
ER+Nevins4     d31628_s_at      1386.0
ER+Nevins4     d31628_s_at       209.5
ER+Nevins4     d31716_at         655.3
ER+Nevins4     d31716_at         116.5
ER+Nevins4     d31716_at         596.3
ER+Nevins4     d31716_at         119.5
ER+Nevins4     d31762_at         573.3
ER+Nevins4     d31762_at         104.7
ER+Nevins4     d31762_at         507.8
ER+Nevins4     d31762_at          88.1
ER+Nevins4     d31763_at         698.0
ER+Nevins4     d31763_at         149.9
ER+Nevins4     d31763_at         593.3
ER+Nevins4     d31763_at         115.8
ER+Nevins4     d31764_at        2993.5
ER+Nevins4     d31764_at         426.6
ER+Nevins4     d31764_at        2882.8
ER+Nevins4     d31764_at         508.0
ER+Nevins4     d31765_at         846.5
ER+Nevins4     d31765_at         140.1
ER+Nevins4     d31765_at        1039.5
ER+Nevins4     d31765_at         207.3
$%
```

... to learn how to predict test samples

Biomarker

Prediction with 1 gene (biomarker)

Color coded expression levels of trainings samples

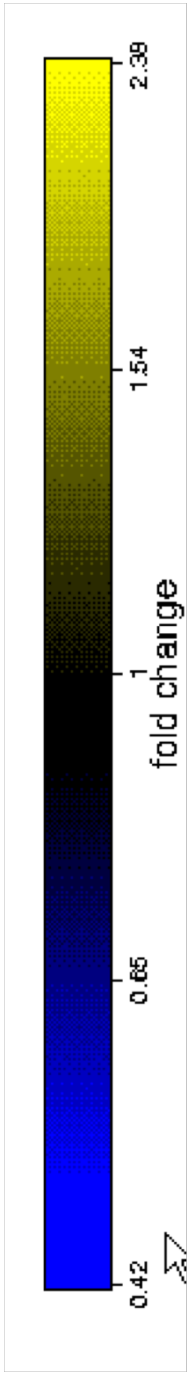


Ms. Smith 🟨 → type A

Ms. Smith 🟦 → type B

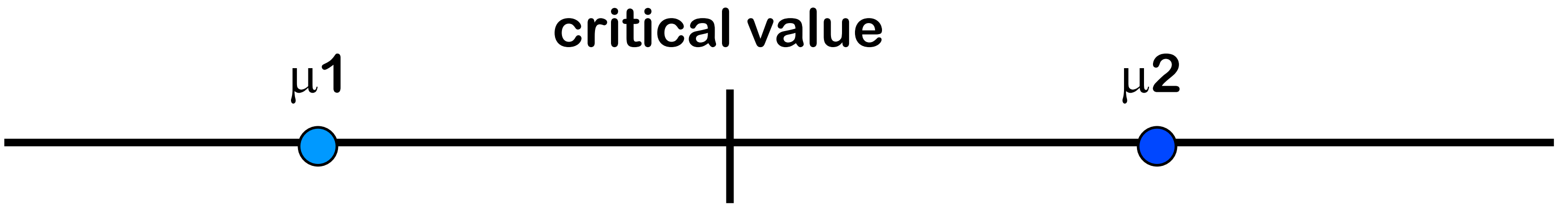
Ms. Smith ⬛ → borderline

Which color shade is a good decision boundary?



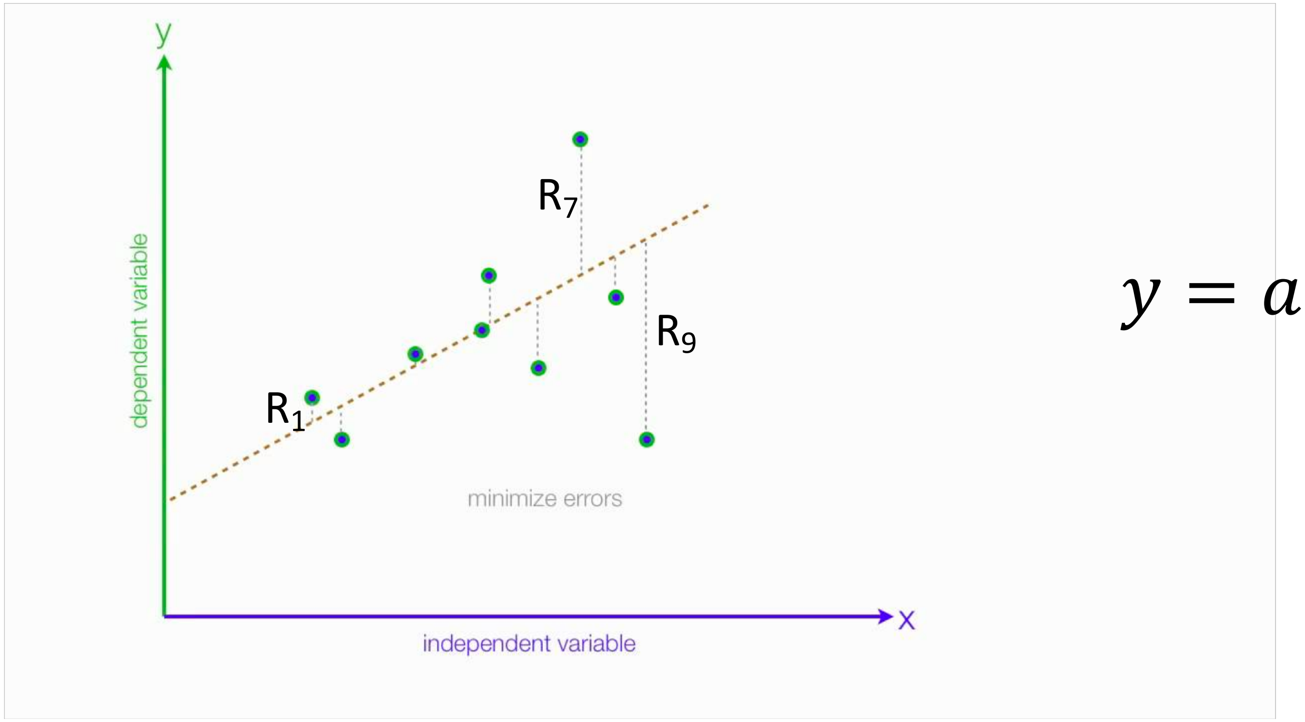
Solution1: Model the classes (Discriminant Analysis)

$$x_0 = \frac{\widehat{\mu_1} + \widehat{\mu_2}}{2}$$



Find the mean in each class and choose the middle as critical value

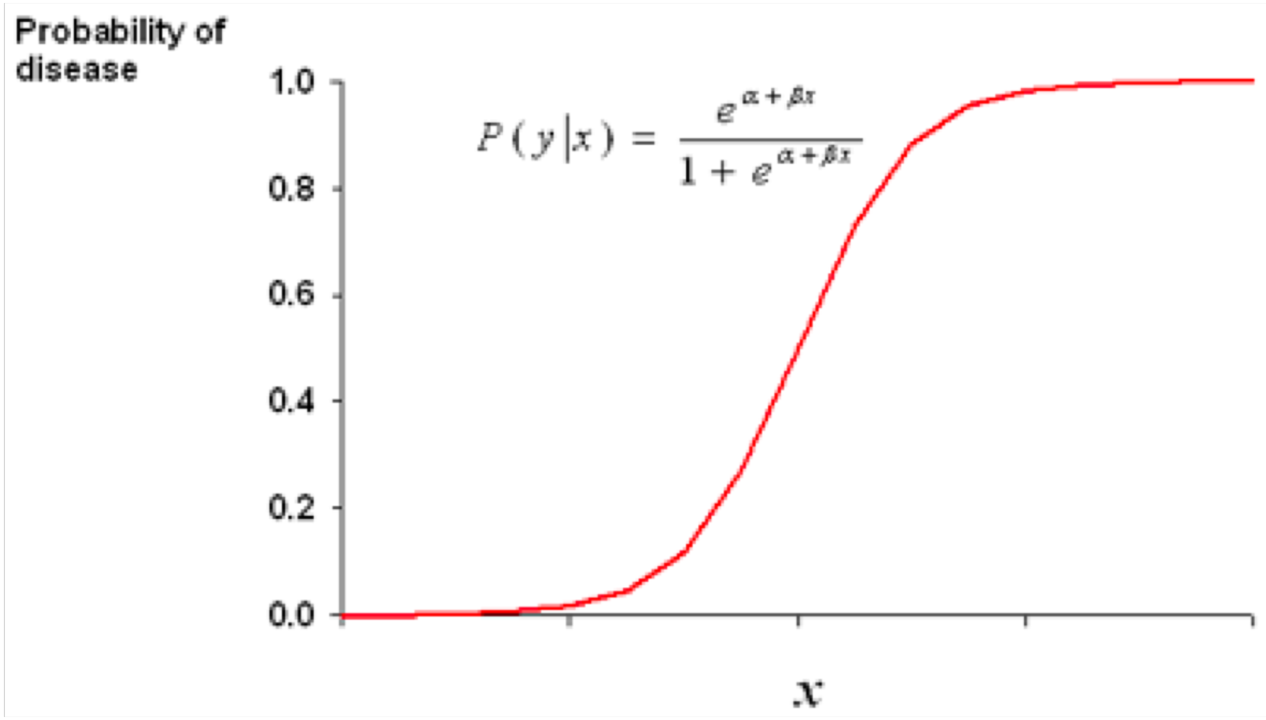
Linear Regression



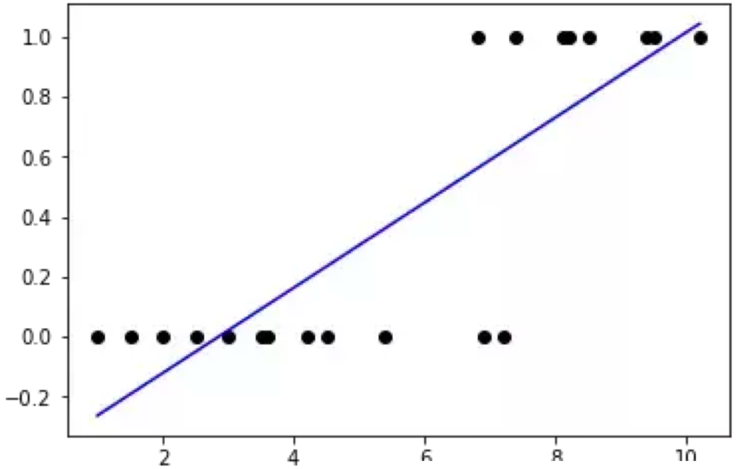
$$y = a + \beta x$$

Find $\beta_{0,}\, \beta_1$ such that the sum of squared residuals R_1^2+… + R_n^2 is minimal

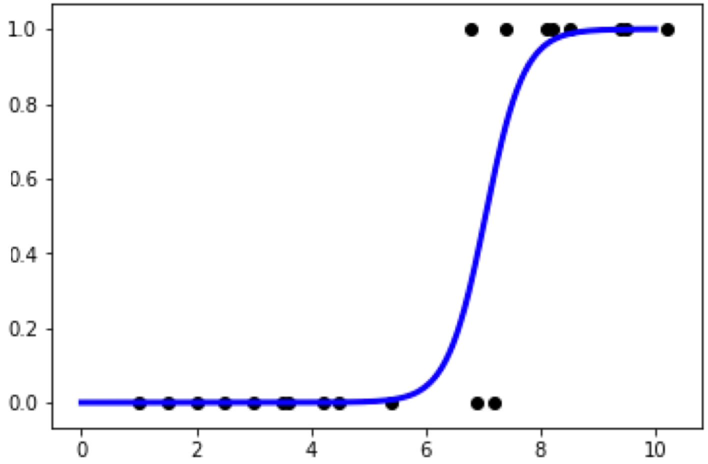
The Logit Function



Solution2: Logistic Regression

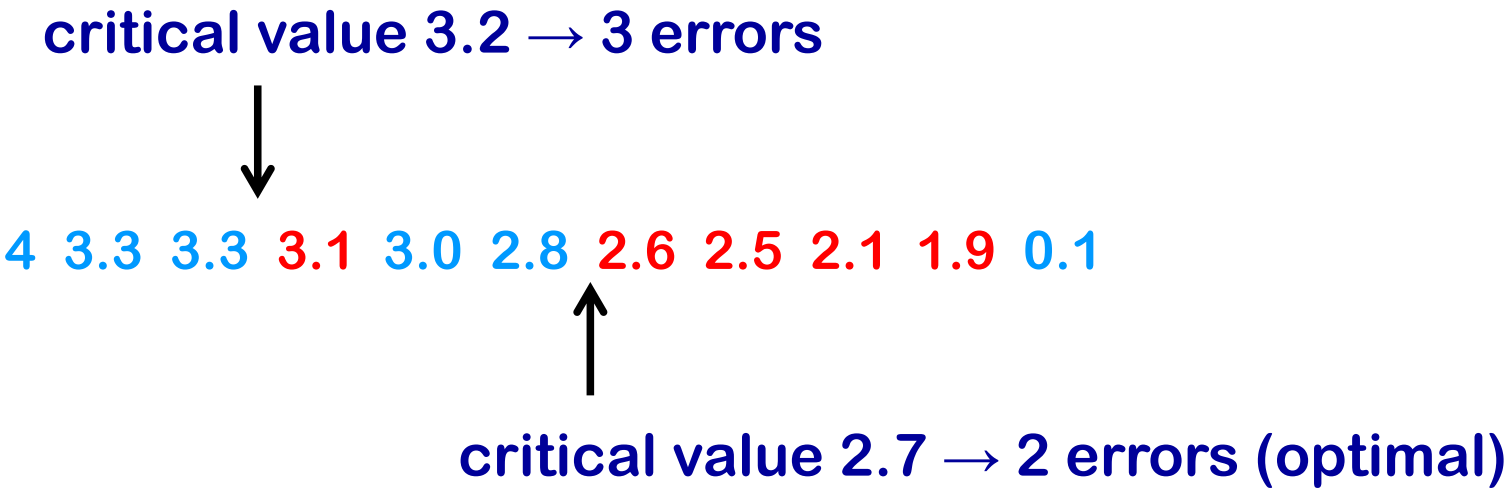


Linear regression on binary data



Logistic regression

Solution3: Go straight for the boundary (Statistical Learning)



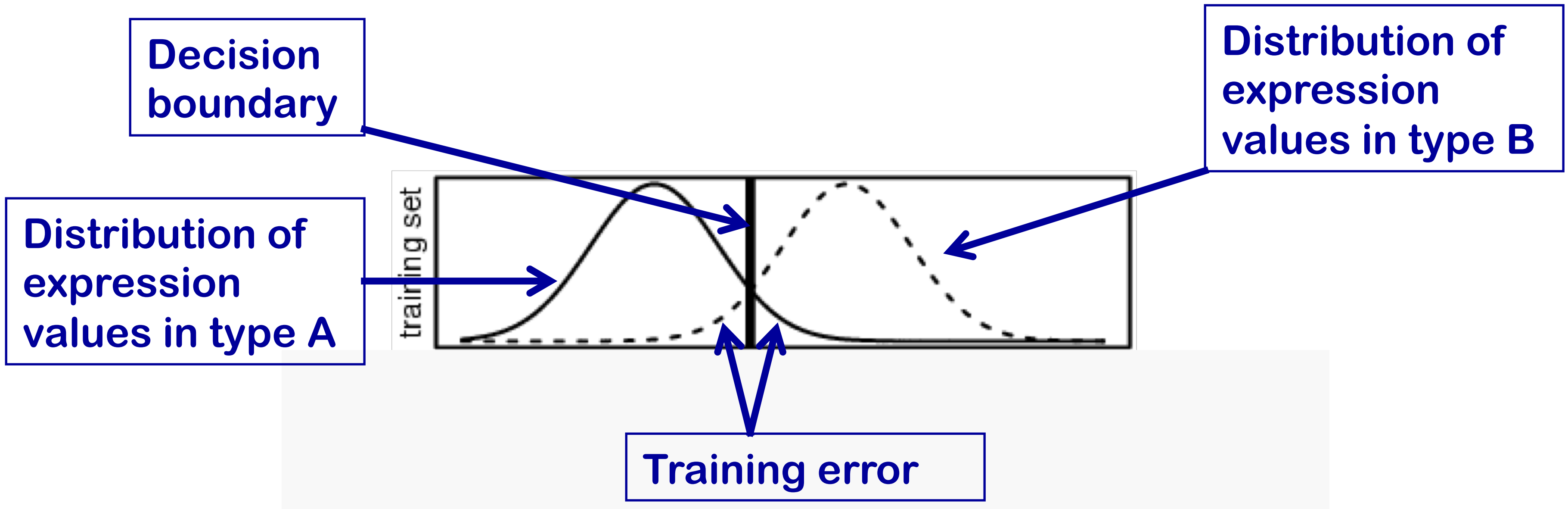
critical value 3.2 → 3 errors

4 3.3 3.3 3.1 3.0 2.8 2.6 2.5 2.1 1.9 0.1

critical value 2.7 → 2 errors (optimal)

The number of misclassifications can be easily optimized on the training data

Minimize the training error



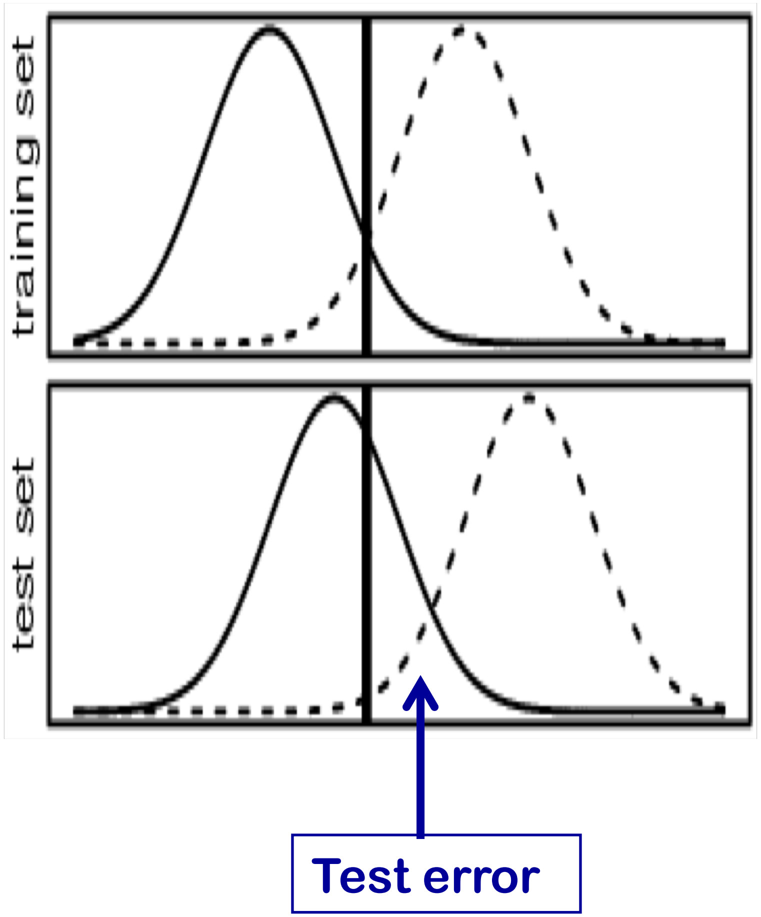
Decision boundary

Distribution of expression values in type A

Distribution of expression values in type B

Training error

Overfitting



The decision boundary was chosen to minimize the training error

The two distributions of expression values for type A and B will be similar but not identical in a set of new cases

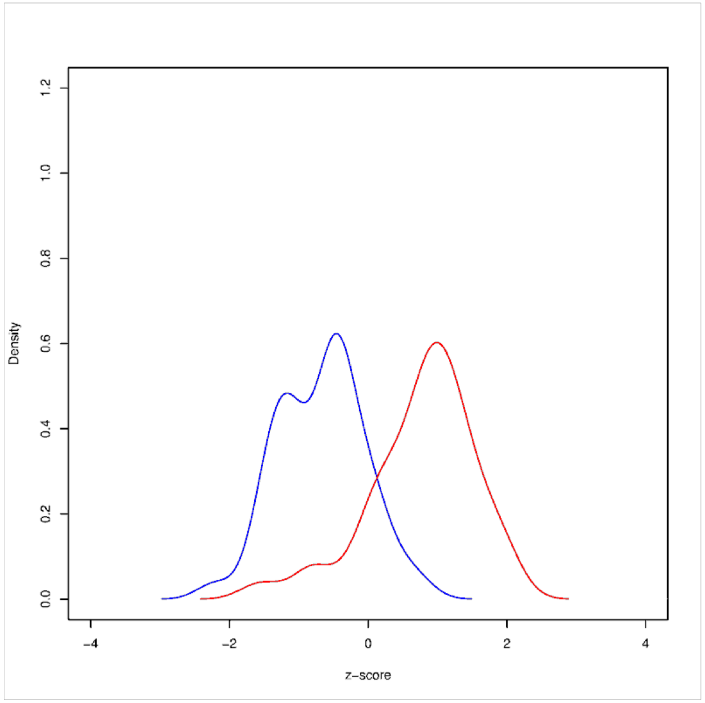
We can not adjust the decision boundary because we do not know the class of the new samples

Test errors are in average bigger then training errors

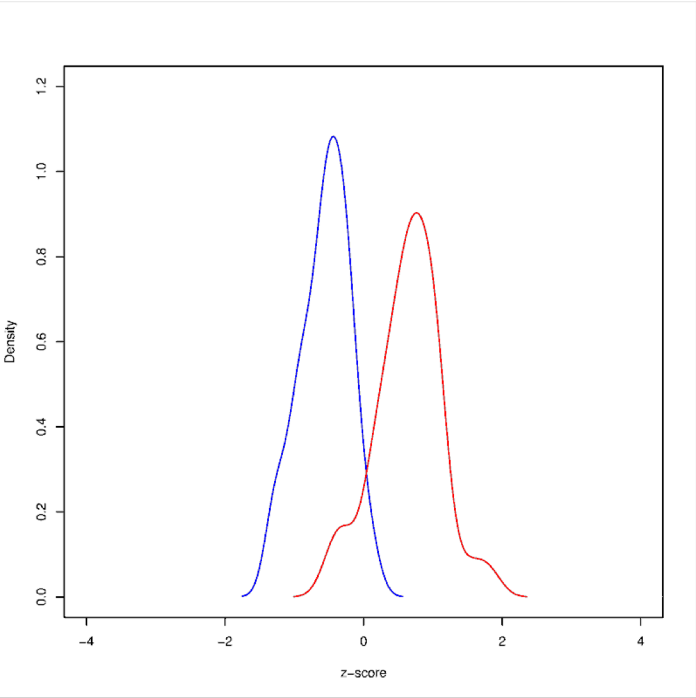
This phenomenon is called *overfitting*

Signatures

Accumulating information across genes



The top gene



The average of
the top 10 genes

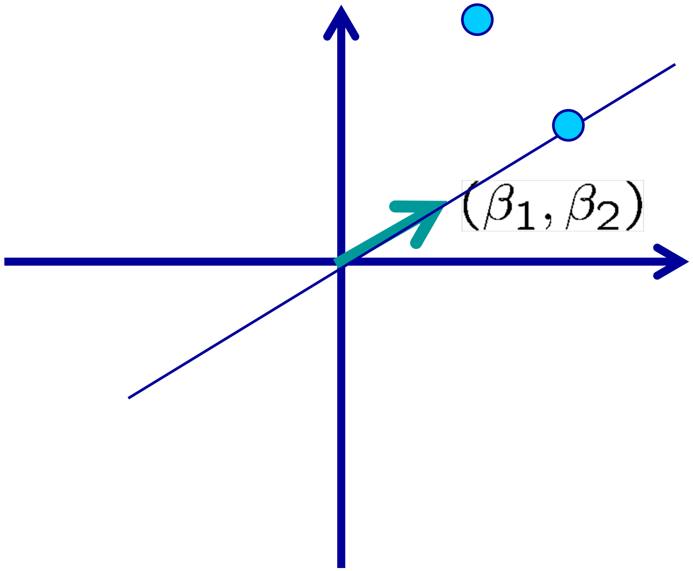
ALL vs. AML Golub et al.

Using a weighted average

$$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

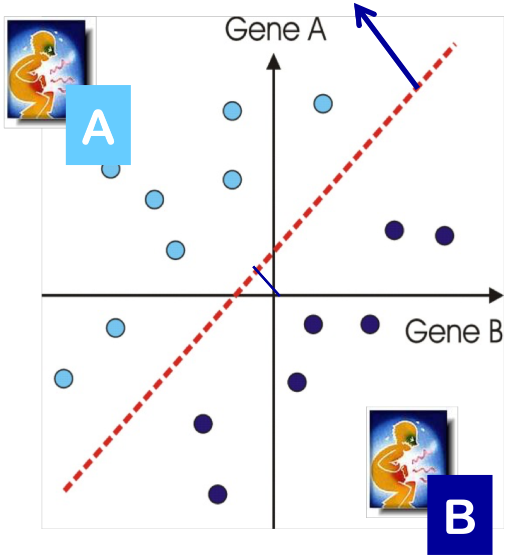
With “good weights” you get an improved separation

The geometry of weighted averages

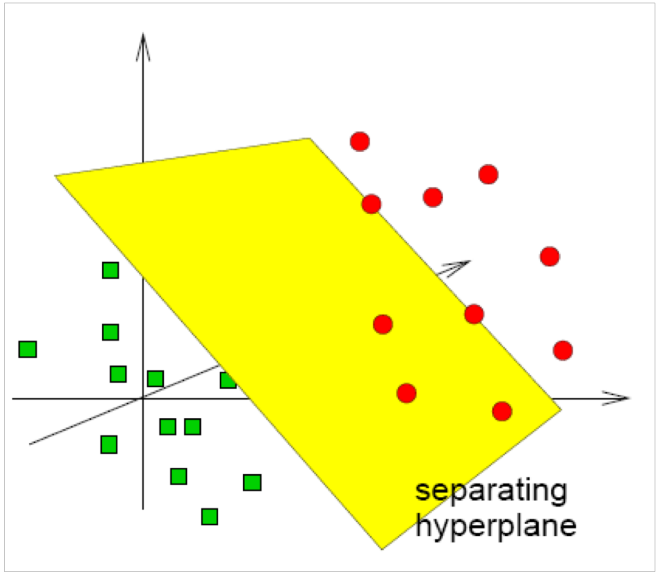


Calculating a weighted average is identical to projecting (orthogonally) the expression profiles onto the line defined by the weights vector

Separating Hyperplanes



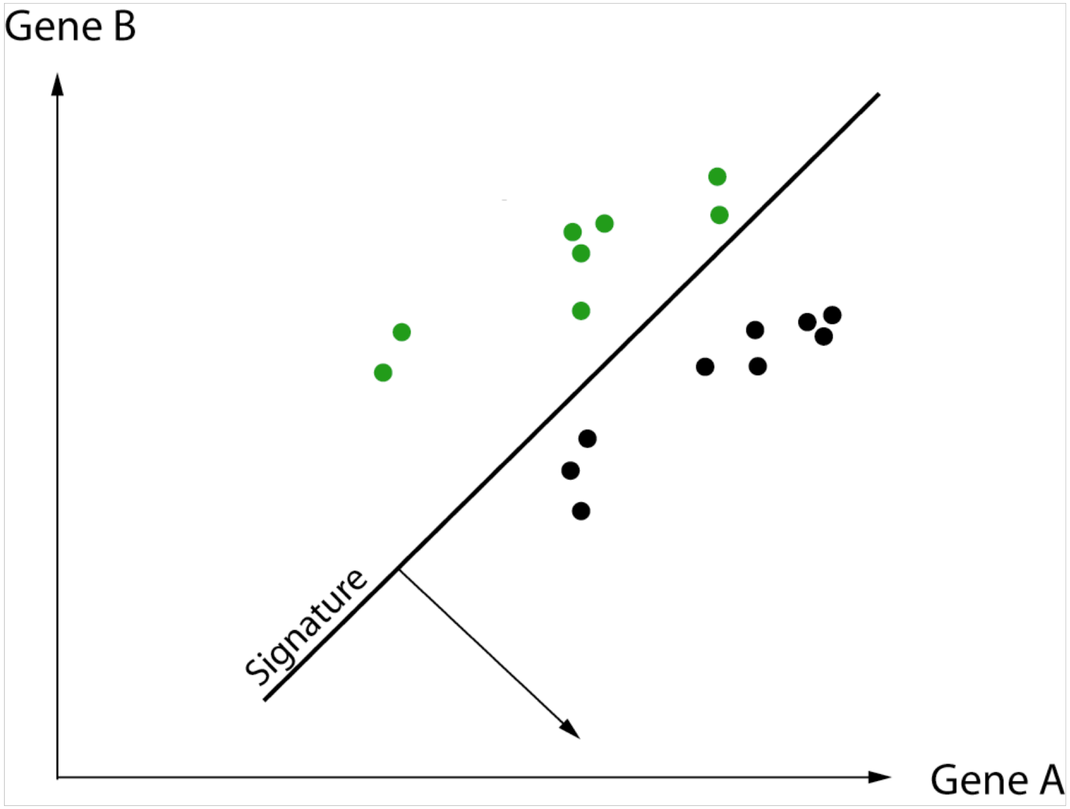
2 genes



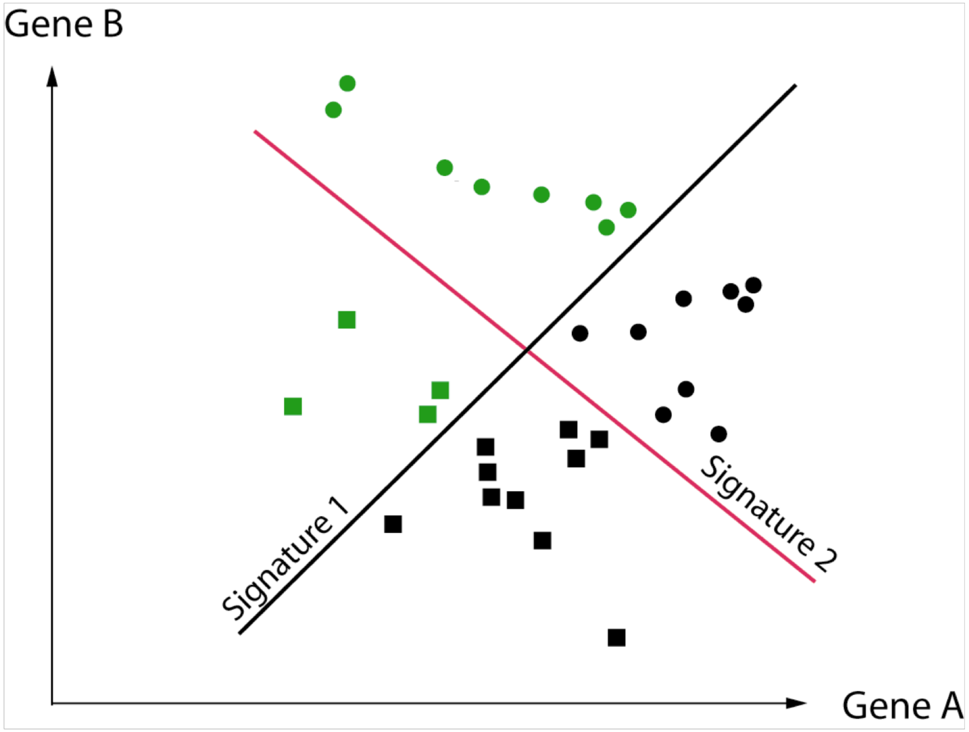
3 genes

Together with an offset the weight vector defines an orthogonal hyperplane that cuts the data in two groups

There are valid signatures without any differently expressed gene

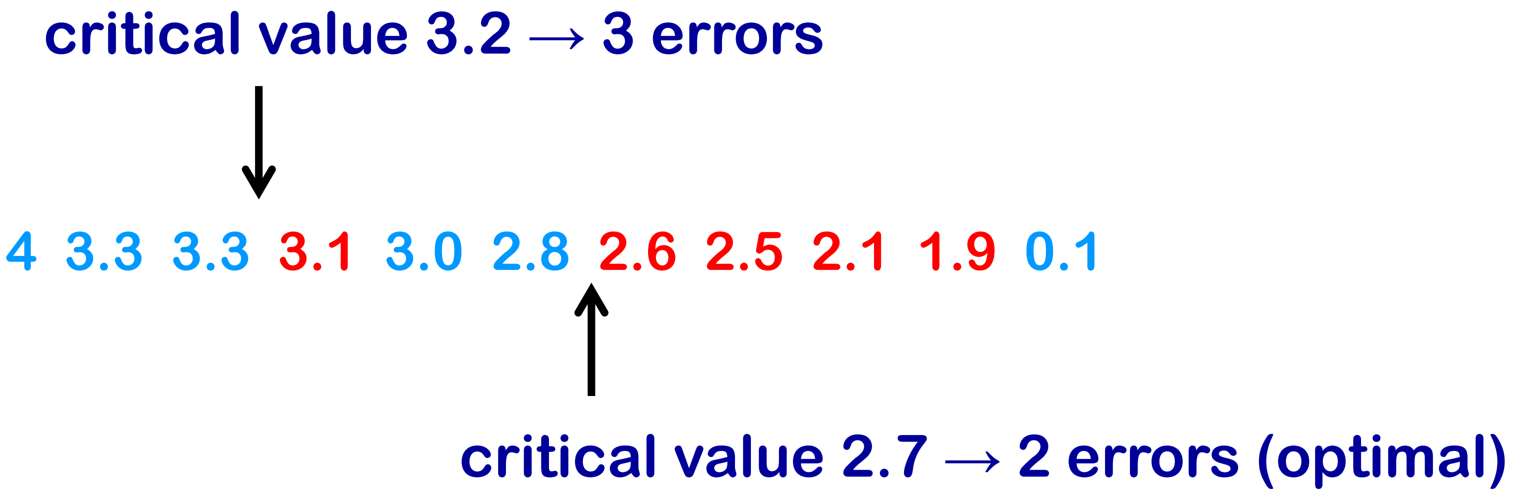


A list of genes does not define a signature yet



Learning Methods

With only one gene the number of misclassifications can be easily optimized on the training data



critical value 3.2 → 3 errors

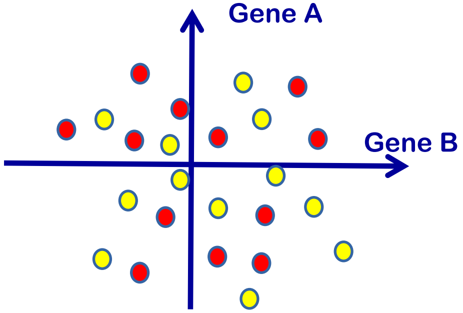
4 3.3 3.3 3.1 3.0 2.8 2.6 2.5 2.1 1.9 0.1

critical value 2.7 → 2 errors (optimal)

Idea: Optimize the weights such that training error becomes minimal

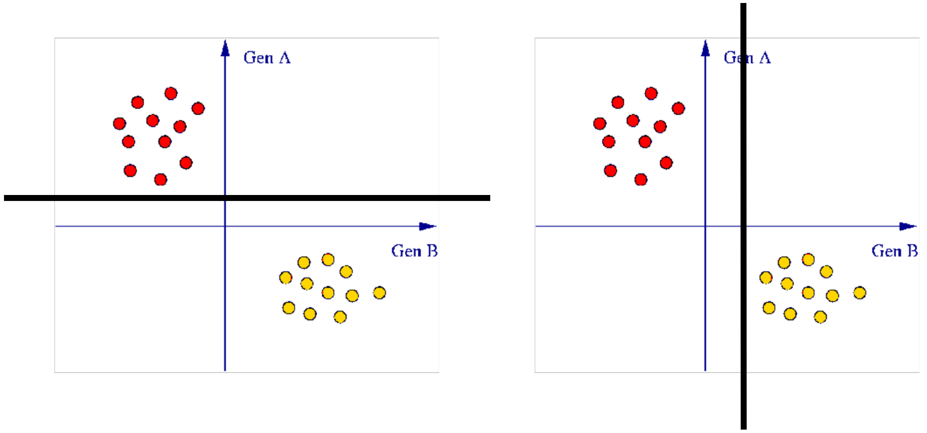
$$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_n x_n$$

In high dimensions only one of these problems exists



Problem 1:

No separating line

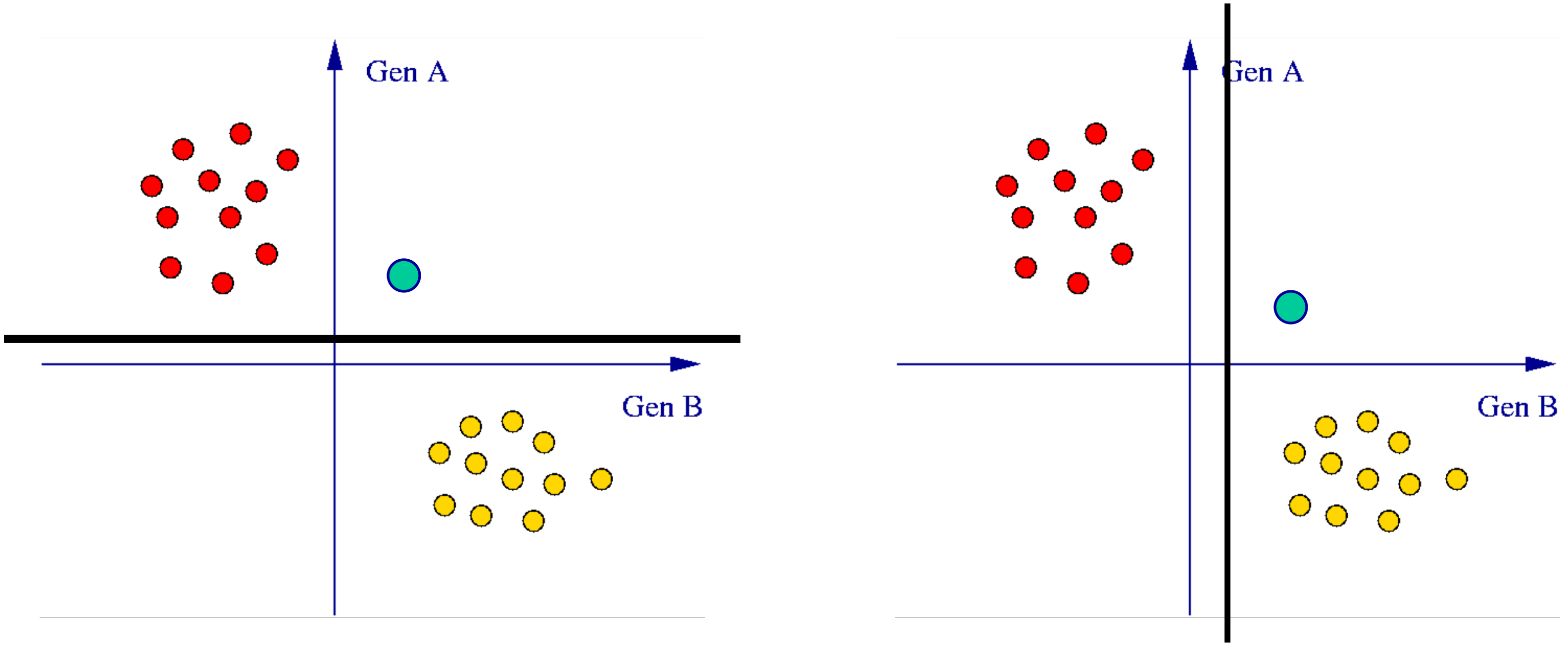


Problem 2:

Many separating lines

Why is this a problem?

What about Ms. Smith ?



This problem is also related to overfitting

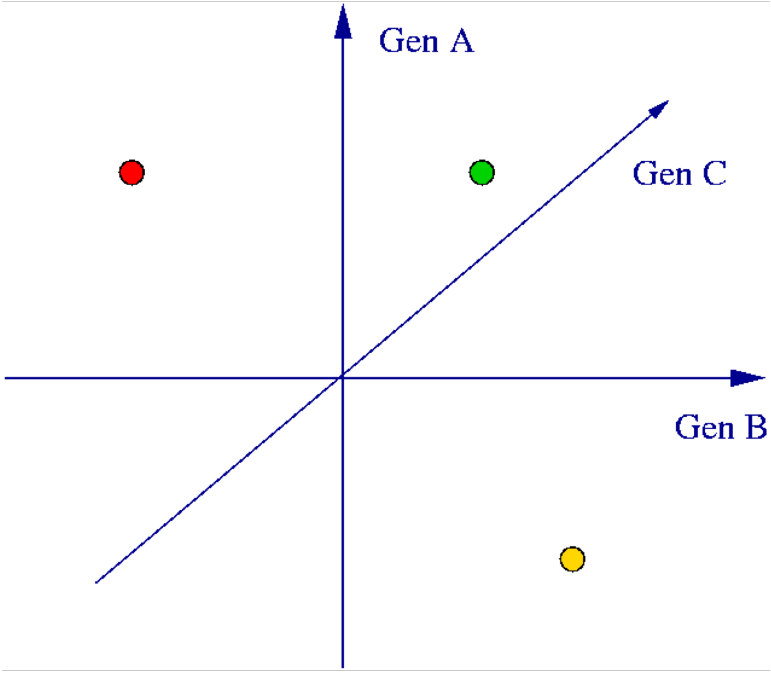
Prediction with 30,000 genes

With the microarray we have more genes than patients

Think about this in three dimensions

There are three genes, two patients with known diagnosis (red and yellow) and Ms. Smith (green)

There is always one plane separating red and yellow with Ms. Smith on the yellow side and a second separating plane with Ms. Smith on the red side

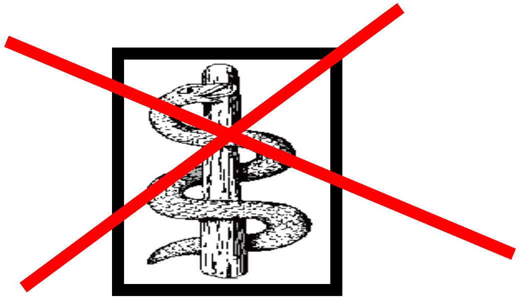


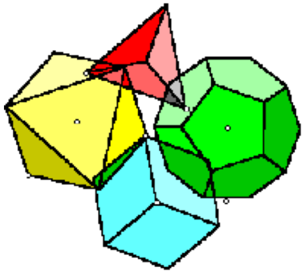
OK! If all points fall onto one line it does not always work. However, for measured values this is very unlikely and never happens in praxis.

The overfitting disaster

From the data alone we can not decide which genes are important for the diagnosis, nor can we give a reliable diagnosis for a new patient

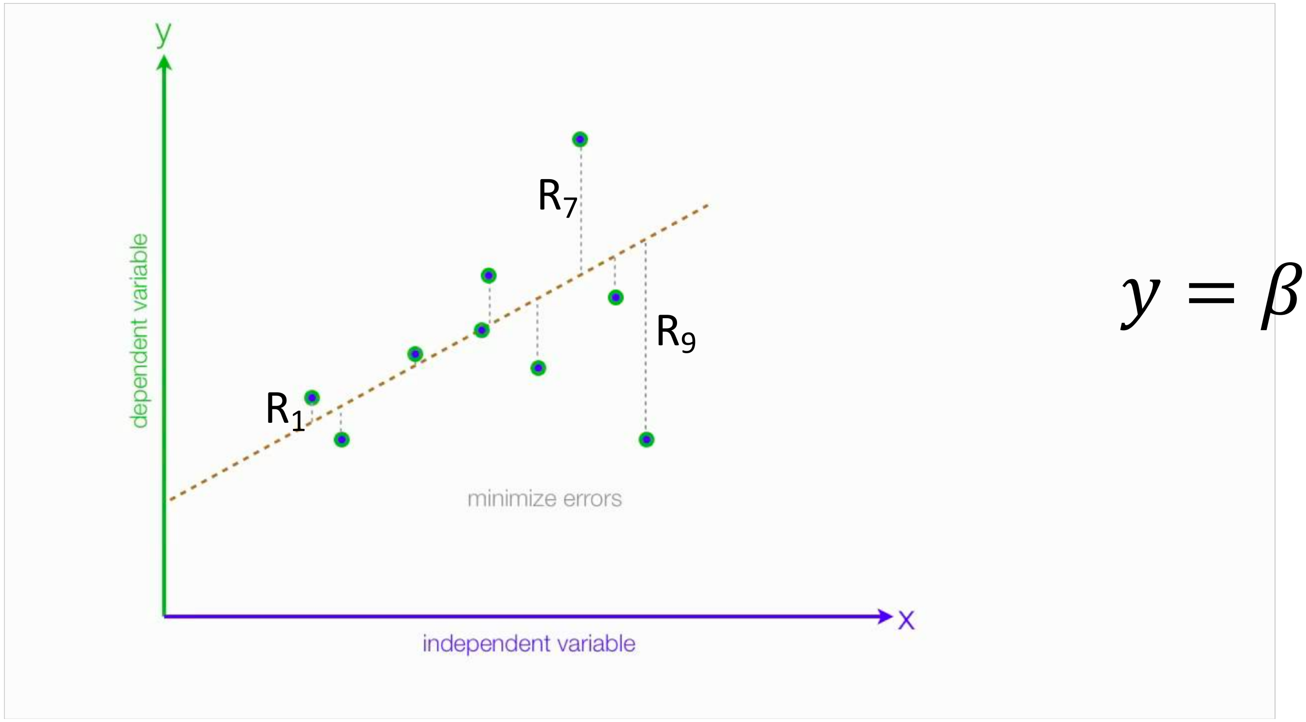
This has little to do medicine. It is a geometrical problem.





Regression and the LASSO

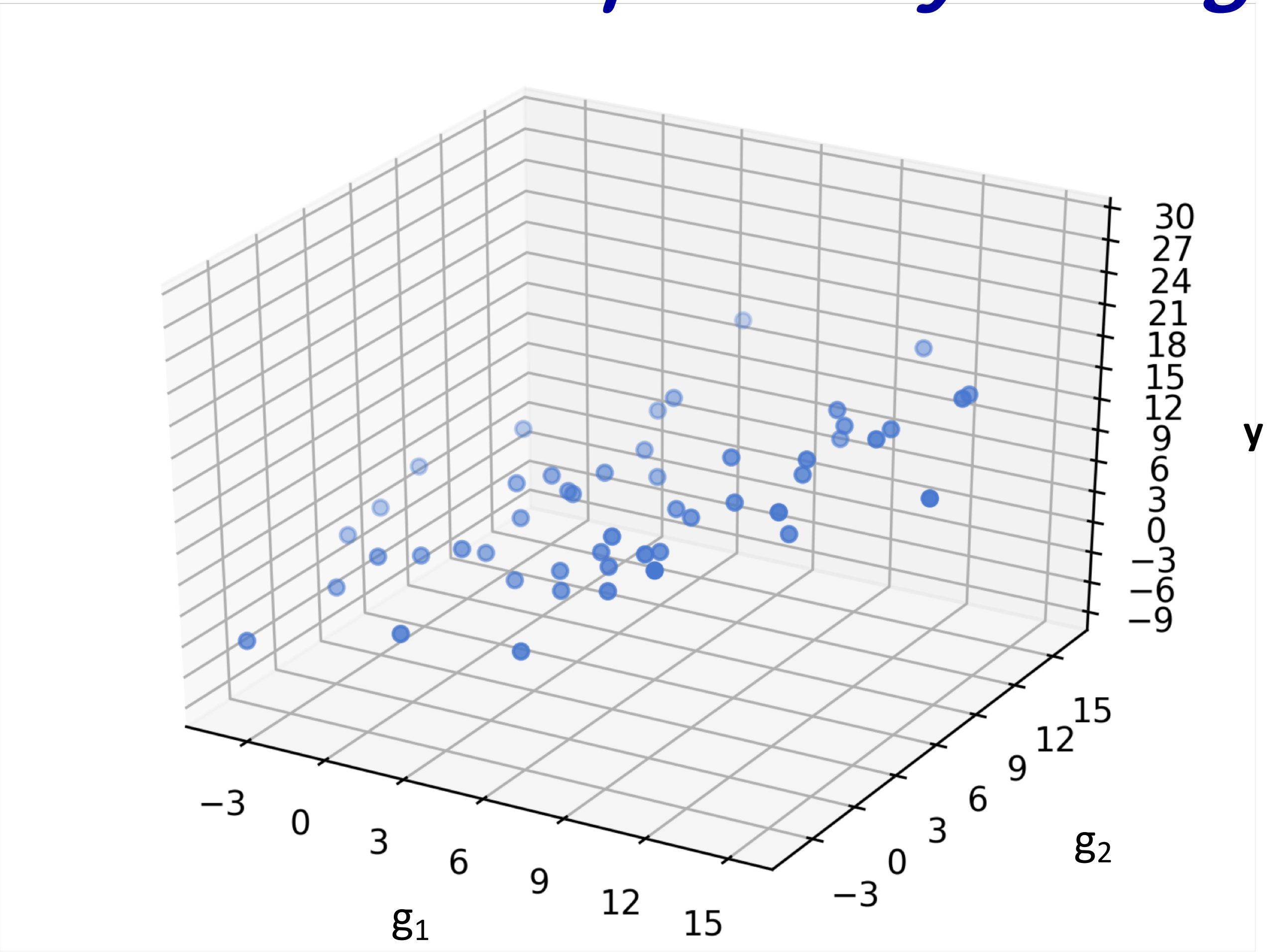
We want to predict y from x



$$y = \beta_0 + \beta_1 x$$

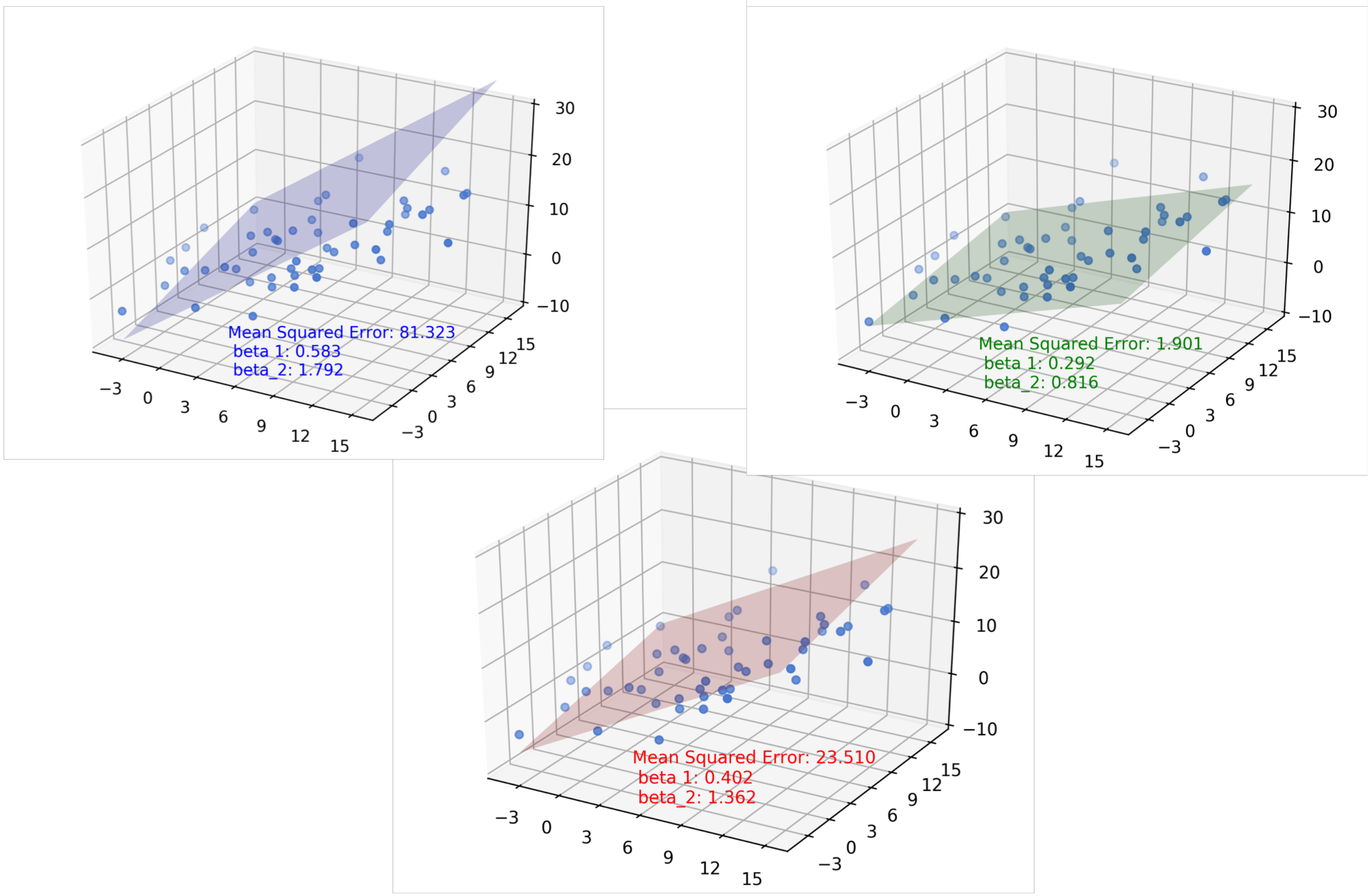
Find $\beta_{0,}\, \beta_1$ such that the sum of squared residuals $R_1^2 + \ldots + R_n^2$ is minimal

We want to predict y from g1 and g2



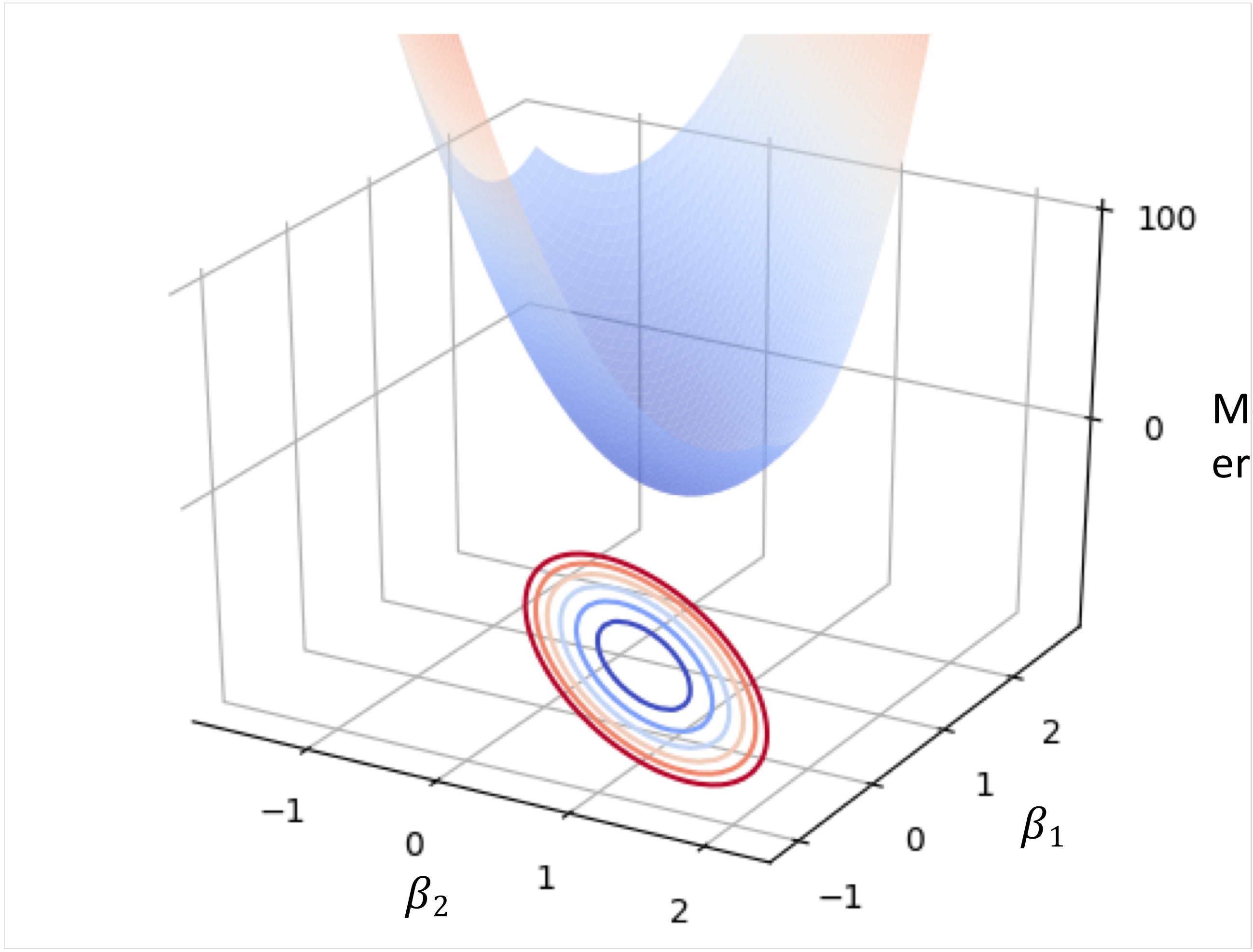
$$y = \beta_1\, g_1 + \beta_2\, g_2$$

Different parameters give different squared residuals



$$y = \beta_1 g_1 + \beta_2 g_2$$

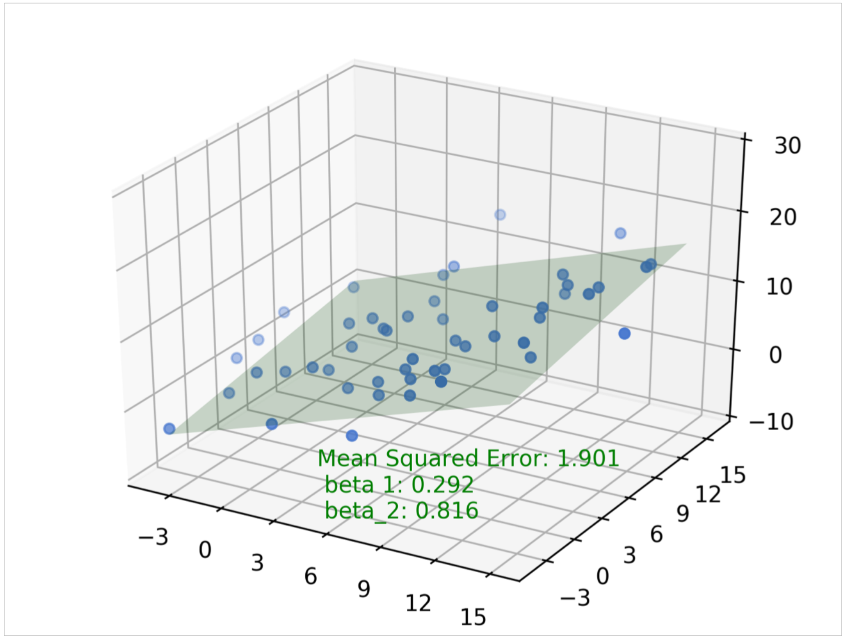
Ordinary least square regression



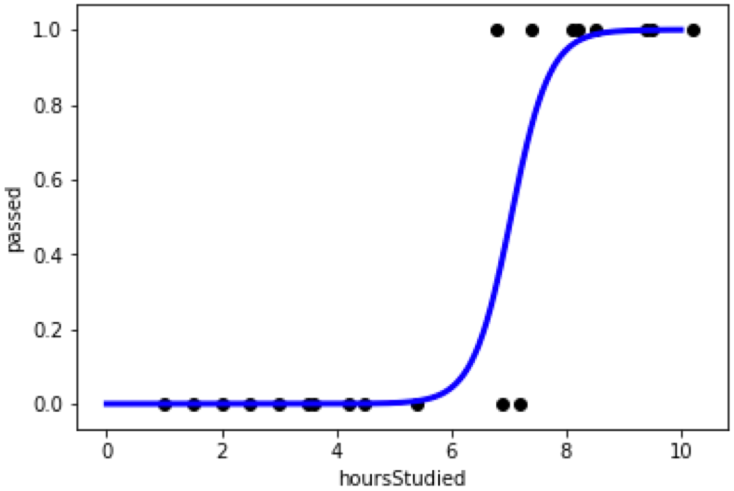
Mean squared error

$$y = \beta_1 g_1 + \beta_2 g_2$$

From a fitting plane to a diagnosis

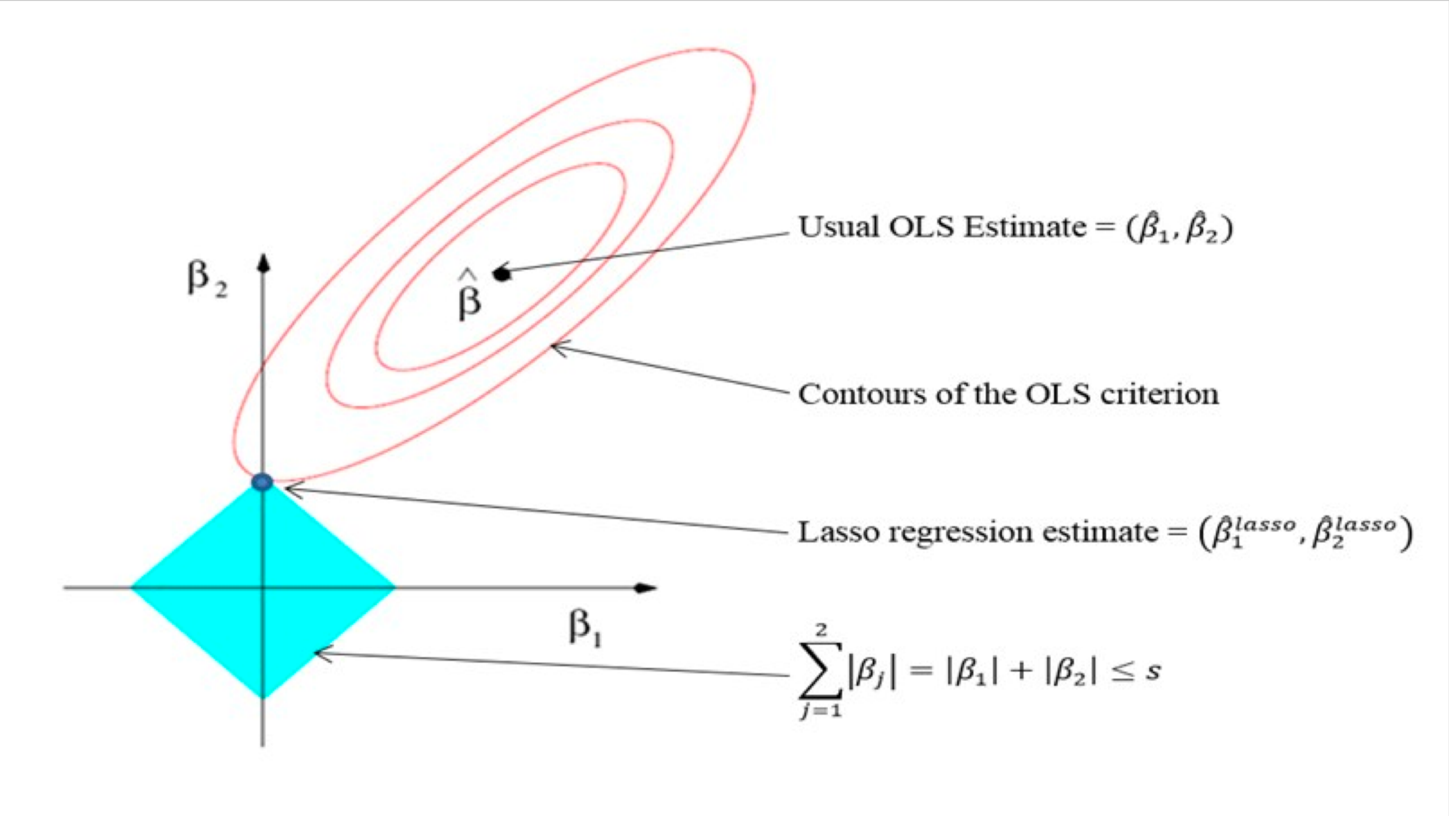


Regression with many features



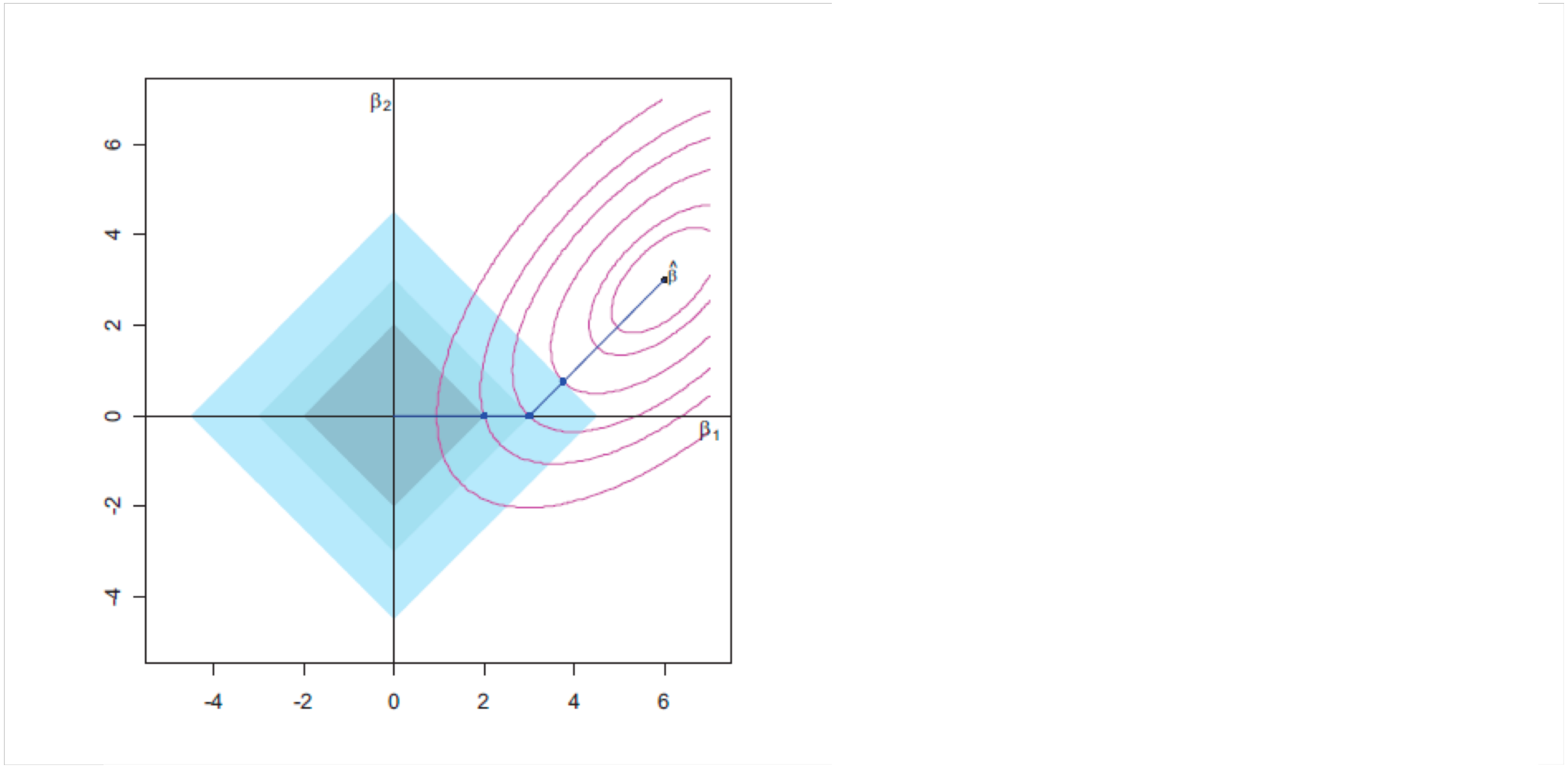
Logistic regression with many features

If we minimize the squared error only in a diamond shaped area, we get sparse models


β_2
$\hat{\beta}$
β_1
Usual OLS Estimate = $(\hat{\beta}_1, \hat{\beta}_2)$
Contours of the OLS criterion
Lasso regression estimate = $(\hat{\beta}_1^{lasso}, \hat{\beta}_2^{lasso})$
$\sum_{j=1}^{2}|\beta_j| = |\beta_1| + |\beta_2| \leq s$

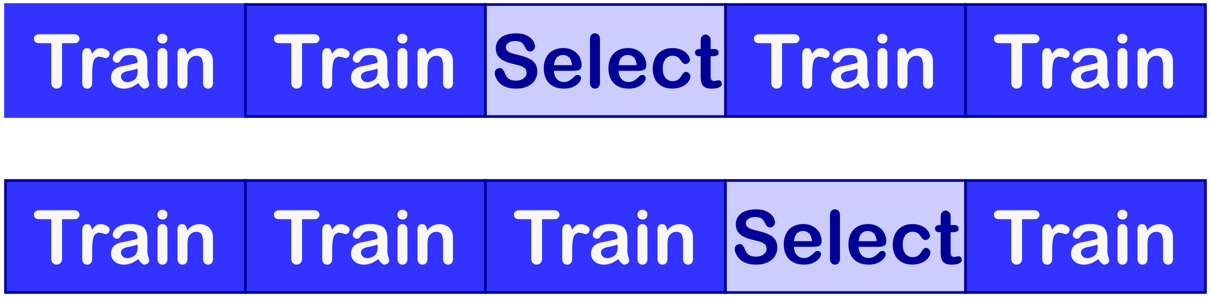
$$(\hat{\alpha}, \hat{\beta}) = \arg\min\left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \qquad \text{subject to} \sum_j |\beta_j| \leqslant t.$$

By making the diamond smaller (shrinkage), we can reduce the number of genes in the model



$$(\hat{\alpha}, \hat{\beta}) = \arg\min\left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \qquad \text{subject to} \sum_j |\beta_j| \leqslant t.$$

How much shrinkage is good ?



Train	Train	Select	Train	Train
Train	Train	Train	Select	Train

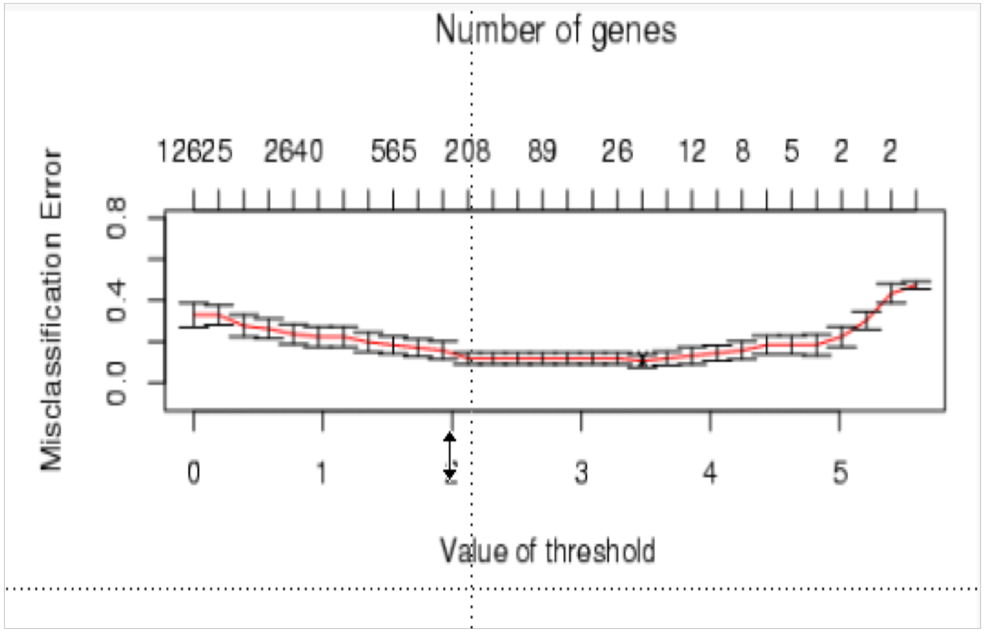
cross validation

Compute the CV-Performance for several values of Δ

Pick the Δ that gives you the smallest number of CV-Misclassifications

PAM does this routinely

Model Selection Output of PAM



Small t, many genes poor performance due to overfitting

High t, few genes, poor performance due to lack of information – *underfitting* -

The optimal t is somewhere in the middle

Validation

How to distinguish a meaningful signature from a meaningless signature?

The meaningless signature might be separating

– *small training error -*

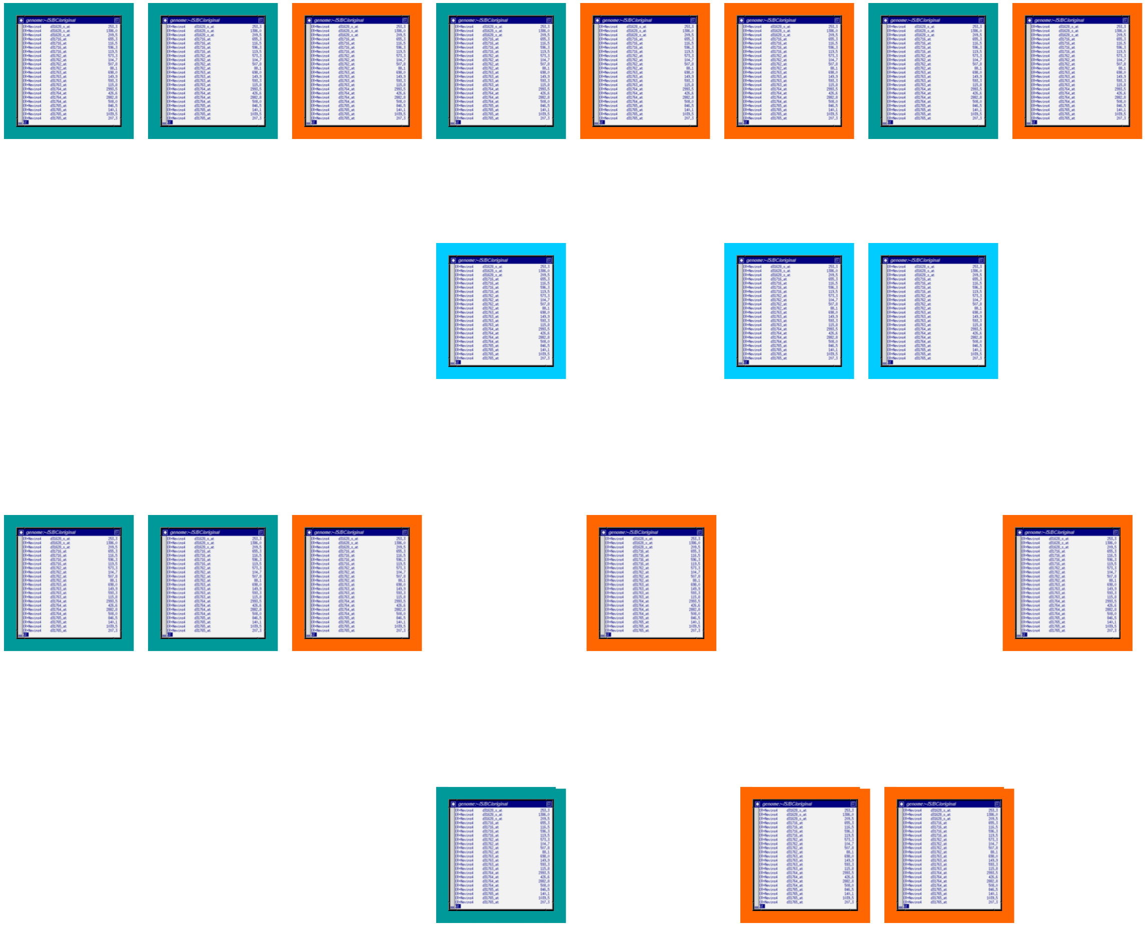
... but it will not be predictive

– *large error in applications* –

The goal is not a separating signature but a predictive signature:

Good performance in clinical practice !!!

Test Sets



Split data into …

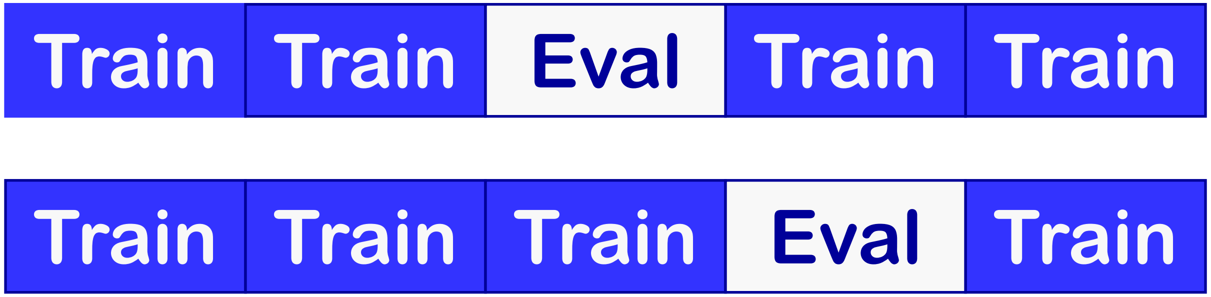
test …

… and training data

ok

ok mistake

Cross Validation



Train Train Eval Train Train

Train Train Train Eval Train

You can not evaluate a fitted classification model (= signature) using cross validation

Cross validation only evaluates the algorithm with which the signature was build

Gene selection must be repeated for every relearning step in the cross validation In the loop gene selection

External Validation and Documentation

Documenting a signature is conceptually different from giving a list of genes, although is is what most publications give you

In order to validate a signature on external data or apply it in practice:

- All model parameters need to be specified

The scale of the normalized data to which the model refers needs to be specified

Establishing a signature

Split Data into Training and Test Data

Training data only:

Machine Learning

- select genes
- find the optimal number of genes
- learn model parameters

Test data only:
Internal validation
Full quantitative specification

External Validations

DOs AND DONTs :

1. Decide on your diagnosis model (PAM,SVM,etc...) and don‘t change your mind later on

2. Split your profiles randomly into a training set and a test set

3. Put the data in the test set away ... far away

4. Train your model only using the data in the training set

(select genes, define centroids, calculate normal vectors for large margin separators, ...)

don‘t even think of touching the test data at this time

5. Apply the model to the test data ...

don‘t even think of changing the model at this time

6. Do steps 1-5 only once and accept the result ...

don‘t even think of optimizing this procedure

Questions

?