

Double Cross Validation for Model-Based Classification

Romain François, Florent Langrognet

INRIA Futurs / Université de Franche-Comté



Machine Learning Focus Session

- Data :

$$X_1, \dots, X_n \in \mathbb{R}^d \stackrel{iid}{\sim} f(\cdot|\theta) = \sum_{k=1}^K p_k \phi(\cdot|\mu_k, \Sigma_k)$$

- Parameter :

$$\theta = ((p_k)_{k=1, \dots, K}, (\mu_k)_{k=1, \dots, K}, (\Sigma_k)_{k=1, \dots, K})$$

Celeux & Govaert (1995)

$$\Sigma_k = \lambda_k D_k A_k D_k'$$

Spherical Family



(2 models)

Diagonal Family



(4 models)

General Family



(8 models)



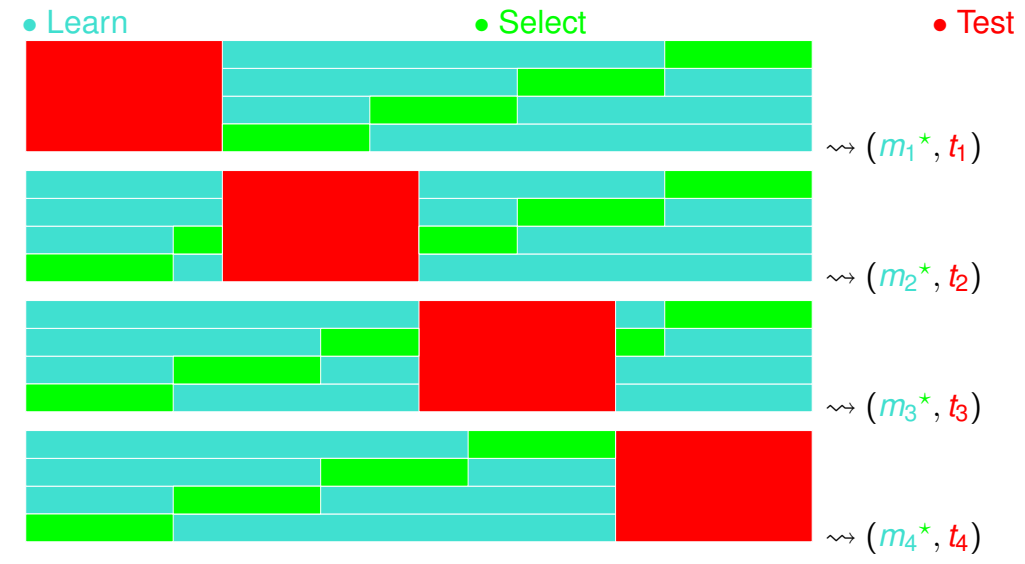
We want to calculate the error rate of the whole procedure

Model the data with one of the many models

- Select a model ⇨ CV
- Find an error rate ⇨ CV

```
R> mixmod(Species~., data=iris, criterion="DCV")
```

```
R> models(orientation="different", shape="equal")
[1] "Gaussian_pk_Lk_Dk_A_Dk" "Gaussian_pk_L_Dk_A_Dk"
[3] "Gaussian_p_Lk_Dk_A_Dk" "Gaussian_p_L_Dk_A_Dk"
```



```
R> mixmod(Species~., data=iris, criterion="DCV", quiet=FALSE)
...running
|-----|
... selecting
DCV |-----| ----> p_L_C (k=3) : 0.02
DCV block 0, n=135, k=3, model : p_L_C , CV ER : 2.22 %, DCV ER :
DCV block 1, n=135, k=3, model : p_L_Dk_A_Dk , CV ER : 1.48 %, DCV ER :
DCV block 2, n=135, k=3, model : p_L_Ck , CV ER : 0.74 %, DCV ER : 13.33 %
DCV block 3, n=135, k=3, model : p_L_Dk_A_Dk , CV ER : 1.48 %, DCV ER :
DCV block 4, n=135, k=3, model : pk_L_Dk_A_Dk , CV ER : , DCV ER : 6.67 %
DCV block 5, n=135, k=3, model : p_L_C , CV ER : 2.22 %, DCV ER :
DCV block 6, n=135, k=3, model : p_L_Dk_A_Dk , CV ER : 1.48 %, DCV ER : 6.67 %
DCV block 7, n=135, k=3, model : p_L_C , CV ER : 1.48 %, DCV ER : 6.67 %
DCV block 8, n=135, k=3, model : p_Lk_Dk_A_Dk , CV ER : 1.48 %, DCV ER :
DCV block 9, n=135, k=3, model : p_L_Dk_A_Dk , CV ER : 1.48 %, DCV ER :

DCV average error rate : 0.0333333
DCV error rate standard error : 0.0447214
```

Thanks for your attention

Romain.Francois@inria.fr

MIXMOD :
<http://www-math.univ-fcomte.fr/mixmod/index.php>