Lecture 15: Classification

IST5573

統計方法 Statistical methods

2016/12/21

Classification: R software

Type	Packages	Functions	Description
Supervised classification and discriminant analysis	MASS	lda	Linear discrimination
		qda	Quadratic discrimination
	mda	mda	Mixture discriminant analysis
		fda	Flexible discriminant analysis
		mars	Multivariate adaptive regression splines
		bruto	Adaptive spline backfitting
	rda		Classification for high dimensional data by means of shrunken centroids regularized discriminant analysis
	class		Package contains functions for classification
		knn	k-nearest neighbours
Recursive partitioning	rpart		Recursive partitioning and regression trees
	tree		Classification and regression trees
	Rweka		Package provides an interface to Weka (a rich toolbox of partitioning algorithms)
	maptree		Graphical tools for the visualization of trees
Random forests	randomForest		The reference implementation of the random forest algorithm for regression and classification
Regularized and shrinkage methods	lars		LASSO
	penalizedLDA		Fisher's LDA projection with an optional LASSO penalty to produce sparse solutions
Support vector machines	e1071	svm	Support vector machines
	kernlab		Kernel-based machine learning methods for classification, regression, clustering, novelty detection, quantile regression and dimensionality reduction.

http://ghuang.stat.nctu.edu.tw/course/statmethods16/files/lectures/classification_in_R.xlsx

部落格分類

	A	B	C	D	E	F	G
1	blogid	commentscount	articlescount	usagedays	subscribercount	cateid	
2	1	15648	2756	2638	161	0	
3	2	18876	3075	3545	2254	0	
4	3	24614	1465	229	417	0	
5	4	18426	583	1494	101	0	
6	5	15854	1390	2639	757	0	
7	6	14624	1856	2285	487	0	
8	7	7608	1422	2374	1375	0	
9	8	2866	789	2500	657	0	
10	9	2973	1118	1591	1134	0	
11	10	8320	366	979	3650	0	
12	11	11339	1822	1453	99	0	
13	12	7269	1374	2707	805	0	
14	13	12138	2318	2566	50	0	
15	14	2657	565	1791	916	0	
16	15	1654	1795	2519	385	0	

(RMD_example 15.1)

Variable	Description
blogid	部落格 ID
commentscount	累積留言數
articlescount	總發表文章數
usagedays	使用痞客邦天數
subscribercount	訂閱數
cateid	部落格分類編號：0=美食情報，1=休閒旅遊，2=職場甘苦

Classification: What is the task?

- Given the sample profile, predict the class
- Mathematical representation: find function D that maps the data matrix $\boldsymbol{X} = [X_1, X_2, \cdots, X_p]$ to $\{1, \cdots, K\}$
- Can we use clustering algorithms?
  - Not appropriate for this tasks. We are ignoring useful information in our prototype data: We know the classes!
- Many methods for class prediction
  - Linear and quadratic discriminant analysis (LDA, QDA)
  - k-nearest neighbor (KNN)
  - Classification and regression tree (CART)

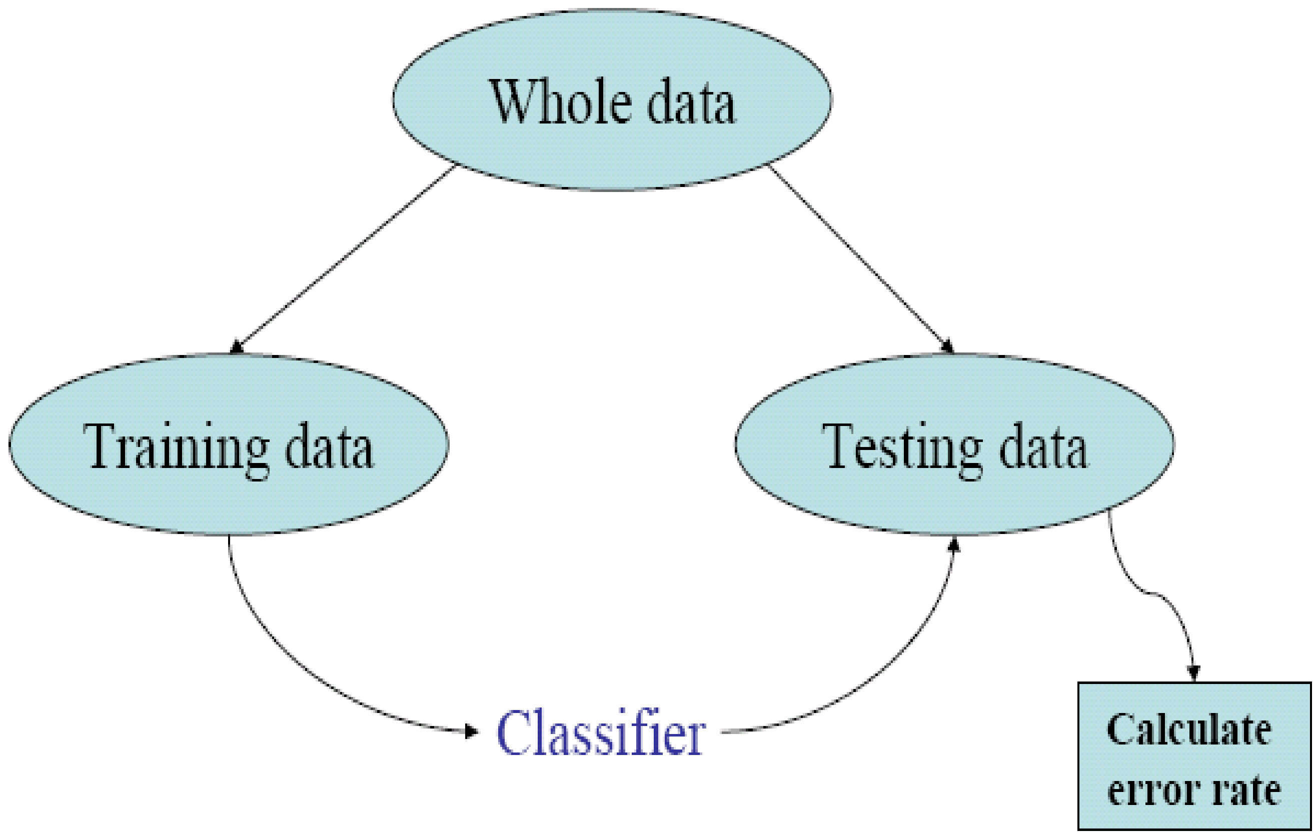
部落格分類

Class label	Measured variable $\boldsymbol{X}$
0 = 美食情報	$\boldsymbol{X}$ = [
1 = 休閒旅遊	X_1(commentscount),
2 = 職場甘苦	X_2(articlescount),
	X_3(usagedays),
	X_4(subscribercount)]

Classification

- Data: Objects $\{\boldsymbol{X}_j, Y_j\}$ $(j = 1, \cdots, n)$. Each object $\boldsymbol{X}_j$ is associated with a class label $Y_j \in \{1, \cdots, K\}$.
- Method: Develop a classification rule $D(\boldsymbol{X})$ that predicts the class label Y well.
- How does the classifier learned from the training data generalize to (predict) a new example?
- Goal: Find a classifier $D(\boldsymbol{X})$ with high "generalization" ability.

Classification methods



Classification for two classes

- Separating **two** classes of objects, or assigning a new object to one of **two** classes.
- π_1, π_2: labels of two classes
- $\boldsymbol{X} = [X_1, X_2, \cdots, X_p]$: measurements on p associated random variables of objects
- 1st class - the population of $\boldsymbol{x}$ values for π_1
  2nd class - the population of $\boldsymbol{x}$ values for π_2
- $f_1(\boldsymbol{x}), f_2(\boldsymbol{x})$: probability density functions for π_1 and π_2, respectively

How to build classification rules?

1. Measured characteristics of randomly selected objects **known** to come from each of the two classes are examined for differences.
2. The set of all possible sample outcomes is divided into two regions R_1 and R_2.
3. A **new** object falls in $R_1 \rightarrow$ class π_1

   A **new** object falls in $R_2 \rightarrow$ class π_2

What should a good or optimal classification procedure be?

- There may not be a clear distinction between measured characteristics of the classes. The groups may **overlap**. It is then possible to **misclassify** new objects.
- A good of optimal classification procedure should
  1. result in few misclassification
  2. take the prior probabilities of occurrence into account (e.g., one class has a greater likelihood of occurrence than another)
  3. account for the costs associated with misclassification (e.g., classifying a π_1 object as belonging to π_2 represents a more serious error than classifying a π_2 object as belonging π_1)

Notations

- $f_1(\boldsymbol{x})$: the probability density of $\boldsymbol{X}$ for class π_1
  $f_2(\boldsymbol{x})$: the probability density of $\boldsymbol{X}$ for class π_2
- R_1: the set of $\boldsymbol{x}$ values for class π_1
  R_2: the set of $\boldsymbol{x}$ values for class π_2
- $\Omega = R_1 \cup R_2$ and $R_1 \cap R_2 = \emptyset$

- $P(2|1)$: the conditional probability of classifying a π_1 object as belonging to π_2

$$P(2|1) = P(\boldsymbol{X} \in R_2|\pi_1) = \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x}$$

- $P(1|2)$: the conditional probability of classifying a π_2 object as belonging to π_1

$$P(1|2) = P(\boldsymbol{X} \in R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x}$$

- p_1: the prior probability of π_1
  p_2: the prior probability of π_2
  $p_1 + p_2 = 1$
- *P*(observation is correctly classified as π_1) = *P*(observation comes from π_1 and is correctly classified as π_1) =
$$P(\boldsymbol{X} \in R_1|\pi_1) \times P(\pi_1) = P(1|1)p_1$$
- *P*(observation is misclassified as π_1) =
$$P(\boldsymbol{X} \in R_1|\pi_2) \times P(\pi_2) = P(1|2)p_2$$
- *P*(observation is correctly classified as π_2) =
$$P(\boldsymbol{X} \in R_2|\pi_2) \times P(\pi_2) = P(2|2)p_2$$
- *P*(observation is misclassified as π_2) =
$$P(\boldsymbol{X} \in R_2|\pi_1) \times P(\pi_1) = P(2|1)p_1$$

Expected cost of misclassification (ECM)

- Cost matrix

		Classify as	
		π_1	π_2
True class	π_1	0	$c(2\|1)$
	π_2	$c(1\|2)$	0

- Expected cost of misclassification (ECM):

$$\text{ECM} = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$$

- A reasonable classification rule should have ECM as small as possible.

Results

- The regions R_1 and R_2 that minimize the ECM are defined by the value $\boldsymbol{x}$ for which the following inequalities hold:

$$R_1\text{:}\ \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_2\text{:}\ \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

- Based on the result, if $\boldsymbol{x}_0$ is a new observation and
$$\frac{f_1(\boldsymbol{x}_0)}{f_2(\boldsymbol{x}_0)} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$
we assign $\boldsymbol{x}_0$ to π_1. If
$$\frac{f_1(\boldsymbol{x}_0)}{f_2(\boldsymbol{x}_0)} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$
we assign $\boldsymbol{x}_0$ to π_2.

Classification with two multivariate normal populations

- $\boldsymbol{X} = \left[X_1, X_2, \cdots, X_p\right]$
- $f_1(\boldsymbol{x}) \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
  $f_2(\boldsymbol{x}) \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$

Classification of normal populations when $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

- If $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are known, the allocation rule that minimizes the ECM is as follows:

  Allocate $\boldsymbol{x}_0$ to π_1 if

  $$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

  Allocate $\boldsymbol{x}_0$ to π_2 if otherwise

- In most practical situations, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are unknown. Suggest replacing by **sample mean and covariance matrices**.
- Note that R_1 and R_2 are defined by **linear** function of $\boldsymbol{x}_0$ (i.e., $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_0$). We thus call this classification rule as the **linear discriminant analysis** (LDA).
- $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}_0$ is called the **linear discriminant**, which can be used for classifying objects.

Fisher’s approach

- Fisher’s approach results in the same discrimination rules as LDA.
- Fisher’s idea was to **transform** the **multivariate** observations x to **univariate** observations y such that the y’s derived from class π_1 and π_2 were **separated** as much as possible.
- Fisher suggested
  1. taking linear transformation,
  2. not assuming that populations are normal,
  3. implicitly assuming that the population covariances are equal.

Classification of normal populations when $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

- Allocate $\boldsymbol{x}_0$ to π_1 if

$$-\frac{1}{2}\boldsymbol{x}_0{}^T\left(\boldsymbol{\Sigma}_1{}^{-1} - \boldsymbol{\Sigma}_2{}^{-1}\right)\boldsymbol{x}_0 + (\boldsymbol{\mu}_1{}^T\boldsymbol{\Sigma}_1{}^{-1} - \boldsymbol{\mu}_2{}^T\boldsymbol{\Sigma}_2{}^{-1})\,\boldsymbol{x}_0$$

$$-\frac{1}{2}\ln\left(\frac{\boldsymbol{\Sigma}_1}{\boldsymbol{\Sigma}_2}\right) + \frac{1}{2}(\boldsymbol{\mu}_1{}^T\boldsymbol{\Sigma}_1{}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2{}^T\boldsymbol{\Sigma}_2{}^{-1}\boldsymbol{\mu}_2)$$

$$\geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

Allocate $\boldsymbol{x}_0$ to π_2 if otherwise

- In most practical situations, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ are replaced by **sample mean and covariance matrices**.
- Note that R_1 and R_2 are defined by **quadratic** function of $\boldsymbol{x}_0$ (i.e., $-\frac{1}{2}\boldsymbol{x}_0{}^T\left(\boldsymbol{\Sigma}_1{}^{-1} - \boldsymbol{\Sigma}_2{}^{-1}\right)\boldsymbol{x}_0$). We thus call this classification rule as the **quadratic discriminant analysis** (QDA).

LDA, QDA

- **LDA**: Assume that K populations (classes) are all from normal distribution with equal covariance matrices.
- **QDA**: Assume that K populations (classes) are all from normal distribution with unequal covariance matrices.

Notes

- The quadratic classification rule is sensitive to departures from normality. I.e., if π_1 or π_2 is not from the multivariate normal, the quadratic rule can lead to large error rates or ECM.
- If the data are not multivariate normal, one can either
  1) transform the data to more nearly normal, or
  2) use a linear rule, which is less sensitive to normality but more sensitive to equal covariance assumption: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$.

Classification with several classes

- For $i, k = 1, \cdots, K$ (the # of classes),
  $f_i(\boldsymbol{x})$: the density associated with class π_i
  p_i: the prior probability of π_i
  $c(k|i)$: the cost of allocating a π_i object to π_k
  $c(i|i) = 0$
  R_i: the set of $\boldsymbol{x}$'s classified as π_i
  $P(k|i) = P(\text{classifying object as } \pi_k | \pi_i)$
  $$= \int_{R_k} f_i(\boldsymbol{x}) d\boldsymbol{x}$$

ECM with several classes

- The conditional expected cost of misclassifying a $\boldsymbol{x}$ from π_i into wrong population is

$$\mathrm{ECM}(i) = \sum_{\substack{k=1 \\ k\neq i}}^{K} P(k|i)c(k|i) \quad i = 1, \cdots, K$$

- The overall ECM

$$\begin{aligned}\mathrm{ECM} &= p_1 \mathrm{ECM}(1) + \cdots + p_K \mathrm{ECM}(K) \\ &= \sum_{i=1}^{K} p_i \left(\sum_{\substack{k=1 \\ k\neq i}}^{K} P(k|i)c(k|i) \right)\end{aligned}$$

Minimum ECM classification with several classes

- The classification regions that minimize the ECM are defined by allocating $\boldsymbol{x}_0$ to that population $\pi_k, k = 1, \cdots, K,$ for which
$$\sum_{\substack{i=1 \\ i \neq k}}^{K} p_i f_i(\boldsymbol{x}_0) c(k|i)$$
is the smallest.

Classification with normal populations

- If $f_i(\boldsymbol{x}) \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, \cdots, K$,
  - **LDA**: Assume that K classes with equal covariance matrices ($\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$).
  - **QDA**: Assume that K classes with unequal covariance matrices ($\boldsymbol{\Sigma}_1 \neq \cdots \neq \boldsymbol{\Sigma}_k$).

The number of linear discriminants in LDA

- When $K \geq 3$, we need more than 1 linear discriminant for classification in LDA.
- In LDA, the number of linear discriminants

$$s \leq \min(K-1, p)$$

部落格分類

- $p = 4$
- Measured variable:

  $\boldsymbol{X}$ = [commentscount, articlescount, usagedays, subscribercount]
- Class label Y = cateid (0=美食情報，1=休閒旅遊，2=職場甘苦)
- # of linear discriminants = 2 ($\leq \min(3-1, 4)$)

部落格分類：LDA

```
> (fmla <- as.formula(paste("cateid ~ ", paste(xname, collapse= "+"))))
cateid ~ commentscount + articlescount + usagedays + subscribercount
> fitlda<-lda(fmla, prior=c(1/3,1/3,1/3), data=blogtrain, na.action="na.omit")
>
Call:
lda(fmla, data = blogtrain, prior = c(1/3, 1/3, 1/3), na.action = "na.omit")

Prior probabilities of groups:
        0         1         2
0.3333333 0.3333333 0.3333333

Group means:
  commentscount articlescount usagedays subscribercount
0     1942.8122      580.6543  1590.488        219.3115
1     1945.8557      549.5644  1773.506        225.4696
2      625.4833      552.8141  1706.329        102.9145

Coefficients of linear discriminants:
                         LD1           LD2
commentscount    0.0001772913  1.897259e-05
articlescount   -0.0001089656 -8.468020e-05
usagedays       -0.0001529691  1.188610e-03
subscribercount  0.0011830598  1.029814e-04

Proportion of trace:
   LD1    LD2
0.7958 0.2042
```

1st linear discriminant

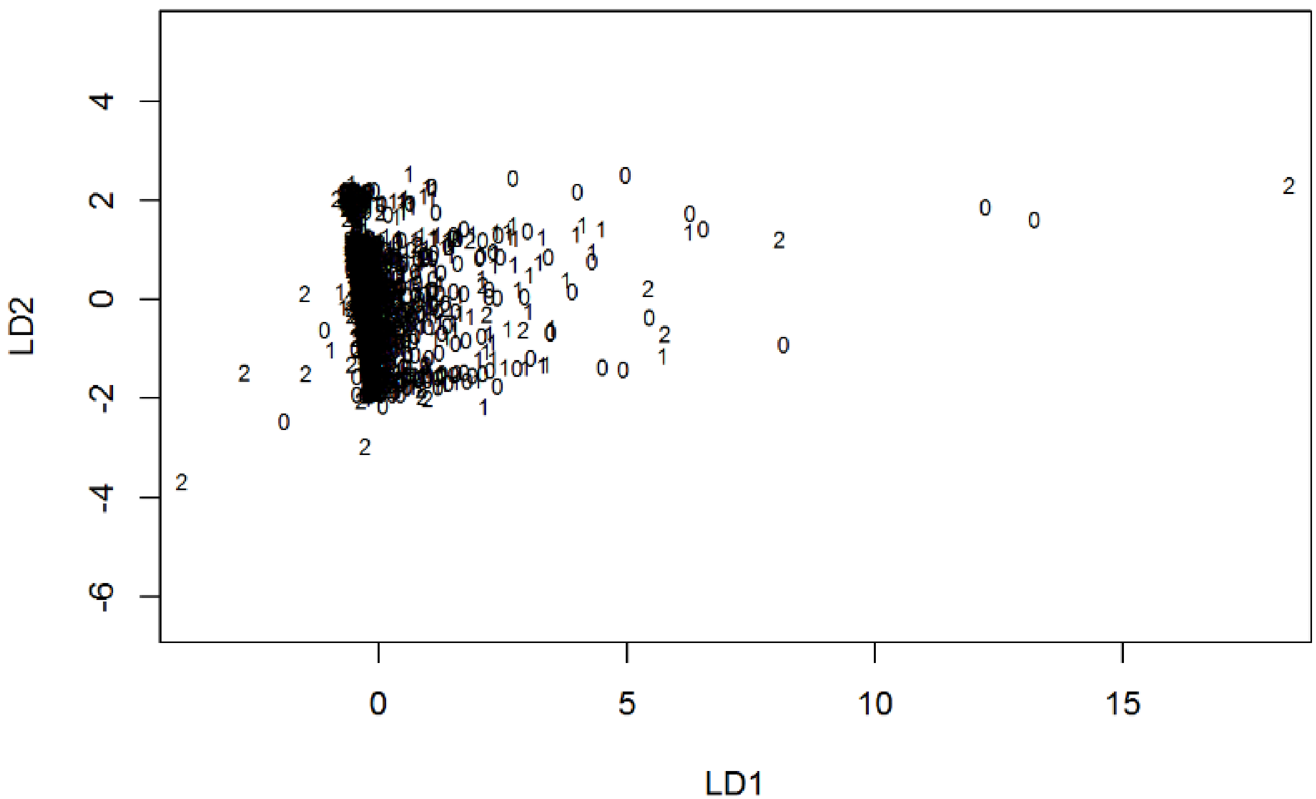
2nd linear discriminant

(RMD_example 15.2)

Use linear discriminants to classify objects

- Linear discriminants were derived for the purpose of obtaining a low-dimensional representation of the data.
- Although they were derived from considerations of separation, the discriminants also provide the basis for a classification rule.

部落格分類：LD plot



(RMD_example 15.2)

Evaluation of classification

- Leave-one-out cross-validation
  1. Omit one sample from the data, and develop a classification function based on the remaining $n - 1$ samples
  2. Classification the "holdout" sample, using the function constructed in 1
  3. Repeat 1 and 2 until all samples are classified
  4. Calculate the misclassification rate based on the classification results from 1, 2 and 3

Evaluation of classification

- *M*-fold cross-validation
  - The original sample is randomly partitioned into *M* equal size subsamples.
  - Proceed as the leave-one-out cross-validation except that now these *M* subsamples are cross-validated.
- Based on an additional dataset that is independent of the one used to build the classification

(RMD_examples 15.2, 15.3)