

Méthodes en classification automatique

Qualité d'une classification

Yves Lechevallier

INRIA-Rocquencourt

E_mail : Yves.Lechevallier@inria.fr

Master ISI

1

Qualité d'une partition

- **Validation interne**

- À partir du critère optimisé par la méthode de classification

- **Validation externe**

- À partir d'informations externes, par exemple une partition.

Master ISI

2

Validation interne

- **Gain** entre la **partition en une classe** et la **partition en K classes** obtenue par la méthode de classification.
- La partition en une classe représente l'hypothèse que l'échantillon est **homogène**. La décomposition de cet échantillon en K classes est inutile

Master ISI

3

Critère d'adéquation

Critère d'adéquation entre $P=(C_1, \dots, C_K)$ et $L=(g_1, \dots, g_K)$.

$$\Delta(P, L) = \sum_{k=1}^K \sum_{s \in C_k} d^2(\mathbf{x}_s, g_k) = \sum_{k=1}^K \sum_{j=1}^p \sum_{s \in C_k} (\mathbf{x}_s^j - g_k^j)^2 \quad C_k \in P, g_k \in \mathbb{R}^p$$

Adéquation de la classe C_k avec le prototype U (Variabilité)

$$S(C_k, U) = \sum_{s \in C_k} d^2(\mathbf{x}_s, U)$$

Critère d'homogénéité de la classe C_k

$$I(C_k) = w(C_k, g_{C_k}) = \min_{U \in \mathbb{R}^p} \sum_{s \in C_k} d^2(\mathbf{x}_s, U)$$

Critère d'homogénéité de la partition $I(P) = \sum_{k=1}^K I(C_k)$

Master ISI

4

Décomposition du critère d'inertie

Quand les prototypes sont les **barycentres** des classes et d est la **distance euclidienne**, alors :

$$w(C, g_C) = \sum_{s \in C} d^2(\mathbf{x}_s, g_C) \text{ est l' } \mathbf{inertie} \text{ de la classe } C.$$

$$T = W + B \text{ avec } T = w(E, g_E) \text{ et } W = \sum_{k=1}^K w(C_k, g_k)$$

$$T_k^j = W_k^j + B_k^j \text{ avec } T_k^j = w_j(C_k, g_k^j) \quad W_k^j = w_j(C_k, g_k^j)$$

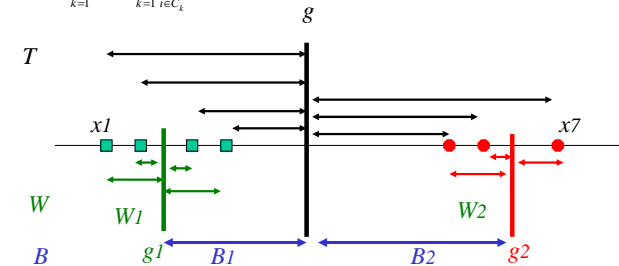
Somme sur les classes $T^j = \sum_{k=1}^K T_k^j$ Somme sur les variables $T_k = \sum_{j=1}^p T_k^j$

Master ISI

5

Décomposition du critère d'inertie

$$T^j = \sum_{k=1}^K T_k^j = \sum_{k=1}^K \sum_{i \in C_k} (x_i^j - g_E^j)^2$$



$$W^j = \sum_{k=1}^K W_k^j = \sum_{k=1}^K \sum_{i \in C_k} (x_i^j - g_k^j)^2 \quad B^j = \sum_{k=1}^K B_k^j = \sum_{k=1}^K n_k (g_k^j - g_E^j)^2$$

Master ISI

6

Décomposition du critère d'inertie

Inertie totale $T^j = \sum_{k=1}^K T_k^j = \sum_{k=1}^K \sum_{i \in C_k} (x_i^j - g_E^j)^2$

Inertie intra classes $W^j = \sum_{k=1}^K W_k^j = \sum_{k=1}^K \sum_{i \in C_k} (x_i^j - g_k^j)^2$

Inertie inter classes $B^j = \sum_{k=1}^K B_k^j = \sum_{k=1}^K n_k (g_k^j - g_E^j)^2$

Relations $T_k^j = W_k^j + B_k^j$

Master ISI

7

Critères de qualité

- Partition
- Variable
- Classe

Un **indice de qualité** est le ratio entre la valeur d'homogénéité d'une classe ou d'une variable et le critère homogénéité associé à la partition grossière $P_0 = E$ pour cette classe ou cette variable.

Un **indice de qualité** peut être interprété comme un gain entre l'hypothèse nulle « Absence de structure = Partition en une classe » et la partition en K classes.

Master ISI

8

Qualité d'une partition

Pour la partition $P_0=E$ l'inertie est définie par :

$$I(E) = w(E, g_E) = \sum_{s \in E} d^2(\mathbf{x}_s, g_E)$$

la **qualité** d'une **partition** P est définie par:

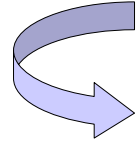
$$Q(P) = \frac{I(E) - I(P)}{I(E)} = \frac{T - W}{T} = \frac{B}{T} \quad \text{homogénéité d'une partition}$$

Cela mesure la part de l'information conservée en assimilant les objets de E aux prototypes des classes obtenues.

Perte d'information

Le critère $Q(P)$ représente la perte d'information en remplaçant le tableau des données Z par le tableau des centres de gravité.

Réduction des lignes d'un tableau



$$\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_i, \dots, \mathbf{z}_N] = \begin{bmatrix} z_1^1 & z_1^j & z_1^p \\ \vdots & \vdots & \vdots \\ z_i^1 & z_i^j & z_i^p \\ \vdots & \vdots & \vdots \\ z_N^1 & z_N^j & z_N^p \end{bmatrix}$$

$$\mathbf{G} = [\mathbf{g}_1, \dots, \mathbf{g}_K] = \begin{bmatrix} g_1^1 & g_1^j & g_1^p \\ \vdots & \vdots & \vdots \\ g_k^1 & g_k^j & g_k^p \\ \vdots & \vdots & \vdots \\ g_K^1 & g_K^j & g_K^p \end{bmatrix}$$

Qualité de la variable j

- une partition $P=(C_1, \dots, C_K)$

Pouvoir discriminant d'une variable

- un système $L=(g_1, \dots, g_K)$ de prototypes

Cet indice représente la part de l'homogénéité de la variable j prise en compte par la partition P :

$$Q_j(P) = \frac{T^j - W^j}{T^j} = \frac{B^j}{T^j}$$

Règle :

$Q_j(P) > Q(P)$ alors la variable j est discriminante par rapport à la partition P

Qualité de la classe k

- une partition $P=(C_1, \dots, C_K)$
- un système $L=(g_1, \dots, g_K)$ de prototypes

Pour chaque classe k la qualité est définie par :

$$Q(C_k) = 1 - \frac{W_k}{T_k} = \frac{B_k}{T_k}$$

Cette valeur mesure le gain de remplacer le prototype associé à E par le prototype de la classe C_k .

Une valeur proche de 1 caractérise une classe homogène et un prototype très différent du prototype global.

Contribution de la variable j

La contribution de la variable j est définie par le ratio entre le critère d'homogénéité calculé sur cette variable et le critère d'homogénéité défini sur l'ensemble des variables.

$$K_j(P) = \frac{W^j}{W}$$

Cette valeur peut être comparé à la contribution, de cette variable sur la partition grossière P_0 :

$$\frac{T^j}{T}$$

Master ISI

13

Contribution de la classe k

Cet indice mesure la contribution de la classe k au critère d'homogénéité de la partition P

$$K(C_k / P) = \frac{W_k}{W}$$

Cette valeur peut être comparé à la contribution, de cette classe sur la partition grossière P_0 :

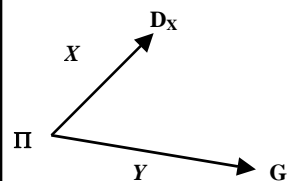
$$\frac{T_k}{T}$$

Master ISI

14

Validation externe

Nous avons une partition a priori qui représente la classification cible. Elle est représentée par la fonction Y



Un exemple est représenté par un couple (\mathbf{x}, y) où \mathbf{x} est sa description et y est l'indice de sa classe a priori.

Master ISI

15

Règle de Bayes d'erreur minimale

$\forall x \quad Y^*(x) = k$ où k est tel que $\Pr(k/x) = \max \Pr(h/x)$

Cette définition est peu opérationnelle, en effet, on connaît rarement la probabilité d'un classement sachant une description.

Théorème de Bayes $\Pr(k/x) = \frac{\pi_k L_k(x)}{L(x)}$

$\pi_k = \Pr[Y = k]$

$L_k(x) = \Pr[X = x / Y = k]$ est la densité de la classe k

$\forall x \quad Y^*(x) = k$ où k est tel que $\Pr(k/x) = \max \pi_k L_k(x)$

Master ISI

16

Les descriptions suivent une loi normale

Le descripteur X des exemples est constitué de p descripteurs numériques et que sa distribution, conditionnellement aux classes, suit une **loi normale multidimensionnelle** centrée sur le vecteur μ_k et de **matrice de variance-covariance** Σ_k .

La **vraisemblance conditionnelle** de X pour la classe k s'écrit alors

$$L_k(x) = \left((2\pi)^p \det \Sigma_k \right)^{-\frac{1}{2}} \exp \left(-\frac{1}{2} (x - \mu_k)' \Sigma_k^{-1} (x - \mu_k) \right)$$

Master ISI

17

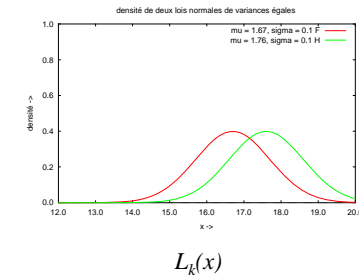
Exemple

Les variances et les probabilités a priori sont égales

La taille moyenne des femmes est égale à 1,67

La taille moyenne des hommes est égale à 1,76

$\mu_1=1,67$ et $\mu_2=1,76$



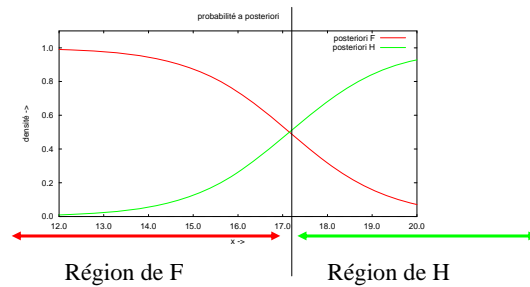
Master ISI

18

Règle de Bayes

$$\Pr(k / x) = \frac{\pi_k L_k(x)}{L(x)}$$

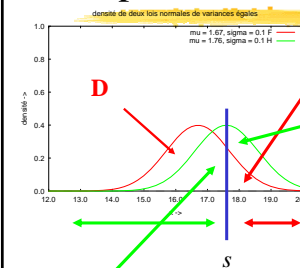
Cette règle minimise le pourcentage de mauvais classement



Master ISI

19

Construction d'un tableau de confusion à partir d'une fonction de décision



Qualité de la décision :
(A+D)/(A+B+C+D)

Classes a priori

	H	F
R_H	A	B
R_F	C	D

Classes d'affectation

Master ISI

20

Qualité d'un score

- Chaque sortie du réseau est associée à une classe a priori.
- L'objectif est d'analyser les scores de cette sortie
- Les exemples sont les observations de la classe a priori associée à cette sortie
- Les contre-exemples sont les observations des autres classes

Courbe ROC (1/3) Receiver Operating Characteristic curve

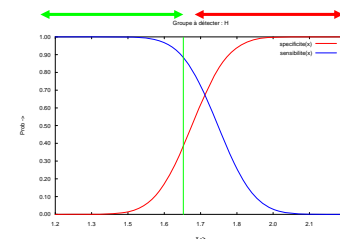
Pour un score s nous avons quatre comptages

- (A) Les Vrais Positifs sont les exemples ayant une valeur supérieure à s .
- (D) Les Vrais Négatifs sont les contre-exemples ayant une valeur inférieure à s .
- (C) Les Faux Négatifs sont les exemples ayant une valeur inférieure à s .
- (B) Les Faux Positifs sont les contre-exemples ayant une valeur supérieure à s .

Courbe ROC (2/3)

- On se fixe la classe a priori G (classe des exemples) et F est l'ensemble des autres classes a priori (classe des contre-exemples)
- La **sensibilité** du score s est égale à $P[S > s / G]$, la sensibilité est le pourcentage de Vrais Positifs
- La **spécificité** du score s est égale à $P[S < s / F]$, la spécificité est le pourcentage de Vrais Négatifs

Courbe ROC



Quand le score augmente

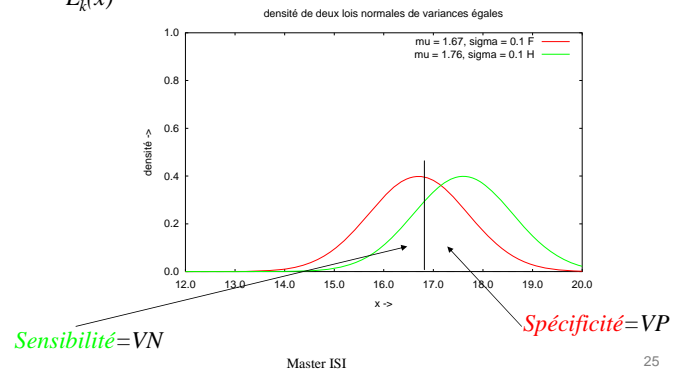
la **sensibilité** diminue cela signifie que le % d'exemples dépassant cette valeur diminue

La **spécificité** augmente cela signifie que le % de contre-exemples en dessous de cette valeur augmente

Si $s=1,6$ on a 90% des exemples qui dépassent cette valeur et 40% des contre-exemples qui sont en dessous de cette valeur

Courbe ROC

$L_k(x)$

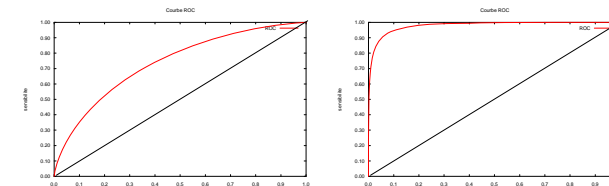


Courbe ROC : interprétation

La diagonale représente la courbe ROC d'un échantillon d'exemples et contre-exemples complètement mélangés

La courbe ROC de gauche est celle de notre exemple ($\mu_1=1,67$ et $\mu_2=1,76$)

La courbe ROC de droite est celle obtenue avec $\mu_1=1,57$ et $\mu_2=1,86$



La surface entre la diagonale et la courbe ROC est une mesure de séparabilité des exemples avec les contre-exemples.

Autre critère d'évaluation

L'évaluation de la qualité des classes C_i générées par la méthode de classification est basée sur sa comparaison avec les classes a priori U_k

n_{ki} est le nombre d'exemples classés dans la classe a priori U_k et ayant été affectés à la classe C_i obtenu par la méthode de classification.

n_k est le nombre d'exemples mises dans la classe a priori U_k
 n_i est le nombre d'exemples de la classe C_i
 n est le nombre d'exemples.

Master ISI

27

F mesure

La **F-measure** combine les mesures de **précision** et de **rappel** entre deux classes U_k et C_i de deux partitions.

La mesure de **rappel** est définie par $R(i,k) = n_{ki} / n_k$.
 C'est le pourcentage d'exemples de la classe a priori k que l'on retrouve dans la classe i obtenue par classification.

La mesure de **précision** est définie par $P(i,k) = n_{ki} / n_i$.
 C'est le pourcentage d'exemples de la classe i que l'on retrouve dans la classe a priori k .

Master ISI

28

F-mesure

La **F-mesure** proposée par (Van Rijsbergen, 1979) combine les mesures de **précision** et de **rappel** entre U_k et C_r .

La mesure de rappel est définie par $R(i,k) = n_{ik} / n_k$.

La mesure de précision est définie par $P(i,k) = n_{ik} / n_i$.

La F-mesure entre la partition a priori U en K groupes et la partition P par la méthode de classification est :

$$F = \sum_{k=1}^K (n_k / n) \max_j (2.R(k, j).P(k, j) / (R(k, j) + P(k, j)))$$

F mesure pour la classe a priori k :

$$F(k) = \max_j (2.R(k, j).P(k, j) / (R(k, j) + P(k, j)))$$