Chapitre 10
Classification supervisée

Lise BELLANGER
Lab-STICC UMR 6285, CNRS, Université Bretagne Sud, Vannes

Ce chapitre s'inspire comme le précédent de plusieurs chapitres de l'ouvrage de Bellanger & Tomassone (2014). Les lecteurs intéressés trouveront aussi de bons compléments dans les ouvrages suivants : Lebart, Morineau et Piron (2000), James *et al.* (2021), Nakache & Confais (2003), Saporta (2011). Les méthodes de classification supervisée sont utilisées dans de très nombreux domaines (santé, domaine bancaire, biologie, archéologie, ...). Nous aborderons ici trois d'entre elles : la méthode des k **plus proches voisins**, l'**analyse discriminante** et la **régression logistique**.

1 Idées générales

Nous avons présenté dans le chapitre 9 des méthodes de classification non supervisée qui ont pour objectif essentiel de créer ou de construire des classes. Contrairement à ces méthodes, celles de **classification supervisée** ou encore appelée classement consistent à placer chaque observation d'un échantillon dans une classe préalablement définie, en fonction de ses caractéristiques, c'est-à-dire en fonction des variables qui la décrivent. Le résultat de cette opération conduit donc à affecter chaque observation à la classe dont elle est la plus proche. Cette dénomination sous-entend que le tableau de données est structuré. Sur les n individus de l'échantillon, on a donc mesuré :

- une variable qualitative prenant q modalités à expliquer, notée $\boldsymbol{y} = (y_1, \ldots, y_n)^T \in \{1, \ldots, q\}^n$, où y_i est la valeur pour l'individu i
- p variables quantitatives formant la matrice de données $\mathbf{X}$ de dimension $n \times p$ où $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p$ sont les colonnes de $\mathbf{X}$.

On rappelle les notations :

- $\boldsymbol{x}^j = \left(x_1^j, \ldots, x_n^j\right)^T \in \mathbb{R}^n$ la colonne de $\mathbf{X}$ décrivant la j ème variable ;
- G_k le groupe des individus de l'échantillon possédant la modalité $k \in \{1, \ldots, q\}$;
- $n_k = \text{card}\,(G_k)$ le nombre d'individus possèdant la modalité k.

2 Méthode des k plus proches voisins

L'algorithme des k **plus proches voisins**, kNN (pour k Nearest Neighbors) en abrégé, est une méthode non paramétrique qui peut être utilisée aussi bien pour la régression que pour la classification supervisée. Son fonctionnement peut être assimilé à l'analogie suivante *"dis moi qui sont tes voisins, je te dirais qui tu es..."*.

Pour effectuer une prévision, l'algorithme kNN ne va pas estimer les paramètres d'un modèle comme c'est le cas pour la régression logistique par exemple. Il se base sur le jeu de données en entier pour produire un résultat. Pour une observation, qui ne fait pas partie du jeu de données, que l'on souhaite prédire, l'algorithme va chercher les k individus du jeu de données les plus proches de notre observation.

2.1 Illustration de la méthodes des k plus proches voisins

Prenons $k = 3$ voisins et $q = 2$ groupes, on cherche à classer une nouvelle observation (en jaune sur la figure 1). On calcule les distances entre l'observation jaune et ses trois plus proches voisins. On observe que ces voisins sont plus souvent verts que rouge, l'observation jaune est donc affectée au groupe vert.

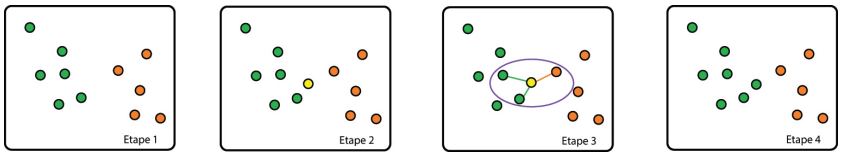


FIG. 1 : Illustration de la méthode kNN.

2.2 Le principe de l'algorithme KNN vu comme une méthode de classification supervisée

Sous forme mathématique, soient d une distance et $k \in \{1, \ldots, n\}$. Les k plus proches voisins de l'individu $\boldsymbol{e} \in \mathbb{R}^p$ à classer sont les k individus de $\Gamma = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$ qui ressemblent le plus à $\boldsymbol{e}$. Notant U_k l'ensemble des k voisins de $\boldsymbol{e}$, on a alors :

$$d\,(\boldsymbol{e}, \boldsymbol{e}_i) < d\,(\boldsymbol{e}, \boldsymbol{e}_j)\ \forall \boldsymbol{e}_i \in U_k \text{ et } \boldsymbol{e}_j \in \Gamma \backslash U_k$$

La méthode des k plus proches voisins propose d'affecter $\boldsymbol{e}$ au groupe (parmi les q groupes) auquel la majorité de ses k voisins appartiennent. Ce groupe peut s'écrire comme :

$$g^* = \underset{g\in\{1,\dots,q\}}{\arg\max} \sum_{i \text{ tq } \boldsymbol{e}_i\in U_k} 1_{\{y_i=g\}}$$

où $1_{\{y_i=g\}} = 1$ si $y_i = g$ (donc $\boldsymbol{e}_i$ appartient au groupe G_g) et 0 sinon.

2.3 Nombre de voisins à choisir

Le choix de la valeur k à utiliser pour effectuer une prévision avec kNN, varie en fonction du jeu de données. En règle générale, moins on utilise de voisins (k petit), plus on risque les problèmes de sous-apprentissage. Par ailleurs, plus k est choisi grand, plus la prévision est fiable, avec toutefois des risques de sur ajustement si k est choisi trop proche de n.

Le choix du nombre de voisins k ainsi que la méthode de calcul de la distance d peut ne pas être évident. Il faut essayer plusieurs combinaisons et évaluer la qualité de l'algorithme des k plus proches voisins à l'aide d'une validation croisée.

3 Analyse discriminante

Le classement est toujours associé à une discrimination préalable, même si ce n'est pas précisé de façon explicite : on ne peut classer des observations que dans des classes préalablement définies. La terminologie la plus ancienne et la plus classique est celle d'**analyse discriminante**, terme qui regroupe les deux opérations celle de discrimination et celle classement. Dans un contexte particulier d'application, l'un ou l'autre de ces deux aspects peut être privilégié. Cette technique statistique permet de décrire et de prévoir l'appartenance à des populations prédéfinies (classes, groupes, modalités* de la variable à prédire, ...) d'un ensemble d'observations (individus*, objets, ...) à partir d'un ensemble de variables*.

3.1 Objectifs et particularités de l'analyse discriminante

L'analyse discriminante recouvre deux objectifs différents ; mais néanmoins souvent complémentaires :

- le premier est descriptif. Il consiste, par une approche géométrique, à déterminer les caractères discriminants d'une population répartie en groupes ; il se traduit souvent par une représentation graphique (de type nuages de points) incluant la définition des frontières séparant les régions de plus forte concentration des individus de chacune des classes (ressemblant souvent à des cartes[1]) comme dans les analyses factorielles. On parlera d'**analyse factorielle discriminante** (en abrégé AFD).

[1] Ce type de représentation qui s'apparente à une carte est également mentionné dans le chapitre 9. Il s'agit d'une représentation graphique dans un plan factoriel, qui peut être exploité, par exemple, pour visualiser les effets de répartition entre des classes d'échantillons aux signatures de compositions élémentaires différentes dans le cas des études de provenance (*cf.* 2.1.4 et 2.1.5).

- le second est prédictif, c'est un outil d'aide pour affecter une nouvelle observation à une population* (on dit aussi un groupe) en associant cette affectation à un degré d'appartenance dont la mesure est généralement fournie par une probabilité* ou un « score ». Il repose sur un cadre probabiliste (*cf.* 7).

La distinction entre ces deux approches n'est pas aussi tranchée. Il est possible de dériver des règles géométriques d'affectation à partir de l'analyse factorielle discriminante.

3.2 Approche géométrique

Cette approche repose sur des notions de distances et ne fait pas intervenir d'hypothèses probabilistes. L'objectif est de proposer un nouveau système de représentation, des variables latentes formées à partir de combinaisons linéaires des variables explicatives initiales, qui permettent de discerner le plus possible les q groupes d'observations. C'est en ce sens que l'analyse discriminante est alors proche des méthodes d'analyse factorielle puisqu'elle permet une représentation graphique dans un espace de dimension réduite ; c'est alors une Analyse en Composantes Principales (ACP) calculée sur les centres de gravité conditionnels des nuages de points avec une métrique particulière. Dans ce cadre, l'analyse discriminante est une méthode essentiellement descriptive

[2] ; on parle dans ce cas d'analyse factorielle discriminante (en abrégé AFD).

Sous forme mathématique, l'objectif de l'AFD est donc de trouver un système de nouvelles variables, combinaisons linéaires des variables explicatives, qui discriminent au mieux les groupes définis par les q modalités de la variable à expliquer. Ces variables notées $\boldsymbol{F}$ sont définies comme un système de vecteurs de $\mathbb{R}^n$, combinaisons linéaires des vecteurs $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p$ (colonnes de $\mathbf{X}$) :

$$\boldsymbol{F} = \mathbf{X}\boldsymbol{u} = u_1\boldsymbol{x}^1 + \cdots + u_p\boldsymbol{x}^p$$

où $\mathbf{u} = (u_1, \ldots, u_p)^T \in \mathbb{R}^p$ est le vecteur des coefficients de cette combinaison linéaire.

Décomposition de la variabilité

Dans le cas le plus général, les n observations ont chacune un poids $p_i \geq 0$ $(i = 1, \cdots, n)$ regroupés dans la matrice diagonale $\mathbf{D}_{n\times n} = \text{diag}\,(p_i)$; la somme des poids p_i valant 1. En général, on prend $p_i = \frac{1}{n}$. Les n observations forment un nuage de n points dans $\mathbb{R}^p$, composé des q sous-nuages $G_k (k = 1, \cdots, q)$ qui ont chacun un poids $q_k = \sum_{i\in G_k} \quad p_i$. On définit deux niveaux de variabilité :

- la **variabilité intra-classe** (« within ») $\mathbf{W}$ qui est la moyenne des **matrices** $p \times p$ **de variance-covariance** [3] $\mathbf{W}_k$ des classes G_k. Si on appelle $\boldsymbol{x}_i \in \mathbb{R}^p$ $(i \in G_k)$ une

[2] L'approche géométrique est celle qui a été mise en oeuvre dans certains cas d'études de provenance des artéfacts archéologiques (*cf.* 2.1.5 pour plus de détails).

[3] La *matrice de variance-covariance* est une matrice carrée qui regroupe les variances et covariances de plusieurs variables. Les éléments sur la diagonale représentent les variances de chaque variable (mesure

observation appartenant à G_k et si $\boldsymbol{g}_k \in \mathbb{R}^p$ est le **barycentre** de la classe G_k, la matrice $\mathbf{W}_k$ s'écrit :

$$\mathbf{W}_k = \frac{1}{q_k} \sum_{i \in G_k} p_i \left(\boldsymbol{x}_i - \boldsymbol{g}_k\right) \left(\boldsymbol{x}_i - \boldsymbol{g}_k\right)^T$$

d'où l'on déduit la matrice $p \times p$ de variance-covariance intra-classe $\mathbf{W}$:

$$\mathbf{W} = \sum_{k=1}^{q} q_k \mathbf{W}_k$$

- la **variabilité inter-classe** (« between ») fournie par la matrice $p \times p$ de variance-covariance $\mathbf{B}$ entre les classes (où $\overline{\boldsymbol{g}} \in \mathbb{R}^p$ est le centre de gravité des n points affectés des poids p_i). Cette variabilité représente la variance des barycentres $\boldsymbol{g}_k$:

$$\mathbf{B} = \sum_{k=1}^{q} q_k \left(g_k - \overline{\boldsymbol{g}}\right) \left(g_k - \overline{\boldsymbol{g}}\right)^T$$

Le théorème de Huygens (*cf.* 9) nous dit que si $\mathbf{T}$ est la matrice $p \times p$ de variance-covariance totale, alors :

$$\mathbf{T} = \mathbf{B} + \mathbf{W}$$

Cette dernière relation, assez usuelle en statistique, est illustrée sur la figure 2.

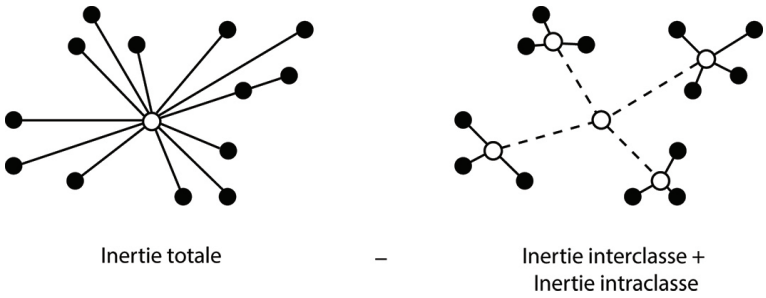


FIG. 2 : Illustration du théorème de Huygens.

Dans la suite nous supposerons que les variables explicatives, contenues dans la matrice $\mathbf{X}$, sont centrées, c'est-à-dire que $\overline{\boldsymbol{g}} = \mathbf{0}$. Dans le cas classique où le poids p_i de

de leur dispersion), tandis que les éléments hors diagonale donnent les covariances entre chaque paire de variables (indiquant si elles varient ensemble). La matrice est *symétrique*, ce qui signifie que la covariance entre la variable A et la variable B est la même que celle entre B et A.

chaque observation $\boldsymbol{i}$ vaut $\frac{1}{n}$, en notant les effectifs des q sous-nuages $n_1, n_2, \cdots, n_q$, on a :

$$\mathbf{B} = \frac{1}{n} \sum_{k=1}^{q} n_k \boldsymbol{g}_k \left(\boldsymbol{g}_k\right)^{\mathrm{T}}$$

$$\mathbf{W} = \frac{1}{n} \sum_{k=1}^{q} n_k \mathbf{W}_k$$

Recherche d'un nouveau repère, sens du critère à maximiser

Dans l'approche géométrique, comme pour une analyse factorielle classique, on essaie de représenter les observations de façon à ce que les sous-nuages soient aussi distincts que possible et que dans ces sous-nuages les observations soient bien regroupées. On va donc rechercher parmi toutes les combinaisons linéaires des variables, celles qui ont une variance intra-classe minimale et une variance inter-classe maximale.

En AFD, comme en ACP (*cf.* 6, §1), on munit $\mathbb{R}^p$ d'une métrique $\mathbf{Q}_{p\times p}$, bien souvent égale à $\mathbf{I}_p$, on note :

- $\boldsymbol{a} \in \mathbb{R}^p$ l'axe discriminant $\left(\boldsymbol{a}^T\mathbf{Q}\boldsymbol{a} = 1\right)$,
- $\boldsymbol{u} \in \mathbb{R}^p$ le facteur discriminant associé $\boldsymbol{u} = \mathbf{Q}\boldsymbol{a}$ tel que $\boldsymbol{u}^T\mathbf{Q}^{-1}\boldsymbol{u} = 1$ et
- $\boldsymbol{F} = \mathbf{X}\boldsymbol{u} \in \mathbb{R}^n$ la variable discriminante.

On définit maintenant le critère qui mesure la capacité d'un axe à discriminer les groupes. Si on projette les observations sur l'axe discriminant $\boldsymbol{a}$, les q barycentres doivent être les plus séparés possible et les observations de chaque sous-nuage doivent être les plus regroupées possible autour de leur barycentre. Pour cela, on utilise la variance intra-classe et la variance inter-classe de la variable discriminante $\boldsymbol{F}$ constituée des coordonnées des projections des individus sur l'axe. On peut montrer que :

- la variance totale de $\boldsymbol{F}$ ou inertie totale projetée sur $\boldsymbol{a}$ vaut $\boldsymbol{u}^T\mathbf{T}u$,
- la variance intra-classe de $\boldsymbol{F}$ ou inertie intra-classe projetée sur $\boldsymbol{a}$ vaut $\boldsymbol{u}^T\mathbf{W}\boldsymbol{u}$,
- la variance inter-classe de $\boldsymbol{F}$ ou inertie inter-classe projetée sur $\boldsymbol{a}$ vaut $\boldsymbol{u}^T\mathbf{B}\boldsymbol{u}$.

Le théorème de Huygens évoqué précédemment entraîne que :

$$\boldsymbol{u}^T\mathbf{T}\boldsymbol{u} = \boldsymbol{u}^T\mathbf{B}\boldsymbol{u} + \boldsymbol{u}^T\mathbf{W}\boldsymbol{u}$$

La quantité $\boldsymbol{u}^T\mathbf{T}\boldsymbol{u}$ étant indépendante des groupes, minimiser $\boldsymbol{u}^T\mathbf{W}\boldsymbol{u}$ est équivalent à maximiser $\boldsymbol{u}^T\mathbf{B}\boldsymbol{u}$. On aura une bonne discrimination des groupes si :

- les barycentres $\boldsymbol{g}_k$ une fois projetés sont bien éloignés i.e. $\boldsymbol{u}^T\mathbf{W}\boldsymbol{u}$ est maximum,
- les groupes projetés ne sont pas trop dispersés i.e. $\boldsymbol{u}^T\mathbf{B}\boldsymbol{u}$ est minimum.

On peut alors prendre comme critère à maximiser soit le rapport ***« inertie inter-classe/inertie intra-classe »*** soit le rapport ***« inertie inter-classe/inertie totale »***, ce qui s'écrit :

$$\frac{\text{inter}}{\text{totale}} \Rightarrow \max_{\boldsymbol{u}\in\mathbb{R}^p} \frac{\boldsymbol{u}^{\mathrm{T}}\mathbf{B}\boldsymbol{u}}{\boldsymbol{u}^{\mathrm{T}}\mathbf{T}\boldsymbol{u}}$$

ou

$$\frac{\text{inter}}{\text{intra}} \Rightarrow \max_{\boldsymbol{v}\in\mathbb{R}^p} \frac{\boldsymbol{v}^{\mathrm{T}}\mathbf{B}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\mathbf{W}\boldsymbol{v}}$$

Les pratiques anglaise et française diffèrent un peu : les français prennent souvent le premier rapport, les anglais le second. Les approches fournissent bien sûr des résultats fonctionnellement liés. La pratique anglaise en mettant l'accent sur le rapport $\frac{\text{inter}}{\text{intra}}$ privilégie « l'aspect analyse de variance » de la discrimination, la pratique française en mettant l'accent sur le rapport $\frac{\text{inter}}{\text{totale}}$ en privilégie l'aspect régression (c'est la proportion de la variance de $\boldsymbol{F}$ expliquée par $\boldsymbol{y}$ encore appelé le rapport de corrélation entre $\boldsymbol{F}$ et $\boldsymbol{y}$. Dans le cadre de la discrimination parfaite, ce rapport sera égal à 1. Plus généralement, la discrimination est d'autant meilleure que le ratio est proche de 1.

Dans la première expression $\frac{\boldsymbol{u}^T\mathbf{B}\boldsymbol{u}}{\boldsymbol{u}^T\mathbf{T}\boldsymbol{u}}$ (pratique française), il est équivalent de maximiser la fonction et de chercher le maximum de la forme quadratique $\boldsymbol{u}^T\mathbf{B}\boldsymbol{u}$ sous la contrainte quadratique $\boldsymbol{u}^T\mathbf{T}\boldsymbol{u} = 1$. Ce qui conduit à :

$$\mathbf{B}\boldsymbol{u} = \lambda\mathbf{T}\boldsymbol{u}$$

Si la matrice $\mathbf{T}$ est inversible, la première variable discriminante $\boldsymbol{F} = \mathbf{X}\boldsymbol{u}$ est telle que le facteur discriminant $\boldsymbol{u}$, vérifie $\mathbf{T}^{-1}\mathbf{B}\boldsymbol{u} = \lambda\boldsymbol{u}$. Ce qui se traduit par : le rapport $\frac{\boldsymbol{u}^T\mathbf{B}\boldsymbol{u}}{\boldsymbol{u}^T\mathbf{T}\boldsymbol{u}}$ est maximal pour u_1 vecteur propre de $\mathbf{T}^{-1}\mathbf{B}$ associé à la plus grande valeur propre notée λ_1 ; la valeur du maximum étant λ_1. La plus grande valeur propre λ_1 qui est par définition le quotient de la variance inter de la fonction discriminante sur la variance totale est toujours comprise entre 0 et 1 ; on l'appelle parfois **pouvoir discriminant**. On a toujours $0 \leq \lambda \leq 1$ et le nombre de valeurs propres non nulles est égal au plus à $(p, q-1)$.

Les facteurs discriminants $\boldsymbol{u}$, donc les variables discriminantes $\boldsymbol{F} = \mathbf{X}\boldsymbol{u}$, sont indépendants de la métrique $\mathbf{Q}$. Le choix de la métrique $\mathbf{Q}$ est donc indifférent pour la construction de variables et des facteurs discriminants, seule la norme change. Ainsi, en pratique si on choisit $\mathbf{Q} = \mathbf{T}^{-1}$ (resp. $\mathbf{W}^{-1}$), on peut choisir parmi tous les vecteurs propres colinéaires, celui qui est $\mathbf{T}$ (resp. $\mathbf{W}$)-normé à 1.

En résumé, on a les définitions suivantes :

- le premier facteur discriminant est $\boldsymbol{u}_1$, c'est le premier vecteur propre de $\mathbf{T}^{-1}\mathbf{B}$,
- la première variable discriminante est $\boldsymbol{F}_1 = \mathbf{X}\boldsymbol{u}_1$,

- le pouvoir discriminant est $\lambda_1 = \frac{\boldsymbol{u}_1^T \mathbf{B} \boldsymbol{u}_1}{\boldsymbol{u}_1^T \mathbf{T} \boldsymbol{u}_1}$.

La décomposition spectrale de $\mathbf{T}^{-1}\mathbf{B}$ donne aussi les autres axes discriminants de l'AFD. En effet, on cherche ensuite $\boldsymbol{F}_2 = \mathbf{X}\boldsymbol{u}_2$, non corrélée à $\boldsymbol{F}_1$, telle $\frac{\boldsymbol{u}_2^T \mathbf{B} \boldsymbol{u}_2}{\boldsymbol{u}_2^T \mathbf{T} \boldsymbol{u}_2}$ que soit maximum et ainsi de suite. On peut montrer que les vecteurs propres de $\mathbf{T}^{-1}\mathbf{B}$ sont les facteurs discriminants et les valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_r$ avec $r \leq \min(p, q-1)$ sont les pouvoirs discriminants. Si on suppose que $q-1$, on pourra construire au plus $q-1$ axes discriminants car le rang r de $\boldsymbol{T}^{-1}\mathbf{B}$ est au plus $\min(p, q-1)$.

La pratique anglo-saxonne, qui nécessite la maximisation de *inter/intra*, revient à chercher les solutions de $\mathbf{W}^{-1}\mathbf{B}\boldsymbol{v} = \mu\boldsymbol{v}$ ou de $\mathbf{B}v = \mu\mathbf{W}\boldsymbol{v}$. On peut voir que :

$$\mu = \frac{\lambda}{1-\lambda} \text{ ou } \lambda = \frac{\mu}{1+\mu}$$

avec $0 \leq \lambda \leq 1$ et $0 \leq \mu \leq \infty$. On retrouve ce que nous avons dit plus haut : λ a les caractéristiques d'un coefficient de détermination R^2 en régression, alors que μ a celles d'une statistique d'analyse de variance comme un F de Fisher. Souvent on choisit la contrainte $v^T W v = 1$; alors $v^T B v = \frac{\lambda}{1-\lambda} = \mu$ et $v^T T v = \frac{1}{1-\lambda}$.

Du point de vue des vecteurs propres, ceci signifie qu'à :

$\boldsymbol{u}$ vecteur propre de $\mathbf{T}^{-1}\mathbf{B}$, de valeur propre λ avec $\boldsymbol{u}^T\mathbf{B}\boldsymbol{u} = 1$

correspond

$\boldsymbol{v}$ vecteur propre de $\mathbf{W}^{-1}\mathbf{B}$, de valeur propre $\mu = \frac{\lambda}{1-\lambda}$ avec $\boldsymbol{v}^T\mathbf{W}\boldsymbol{v} = 1$.

Les vecteurs $\boldsymbol{u}$ et $\boldsymbol{v}$ sont liés par la relation : $\boldsymbol{u} = \sqrt{1-\lambda}\boldsymbol{v}$. μ est évidemment toujours supérieur à λ.

Remarque : L'AFD peut être vue comme une ACP du nuage des q centres de gravité avec la métrique $\mathbf{T}^{-1}$ ou avec la métrique $\mathbf{W}^{-1}$. Oon peut aussi en tirer diverses conclusions utiles pour les interprétations : les variables discriminantes sont non corrélées deux à deux et on peut les interpréter à l'aide des cercles des corrélations vus précédemment (voir chapitre Analyse factorielle).

Règles géométriques d'affectation ou règles d'affectation de Mahalanobis-Fisher

Après avoir estimé la meilleure représentation de la séparation en q classes préexistantes, on peut rechercher la classe d'affectation d'un individu $\boldsymbol{e} \in \mathbb{R}^p$ pour lequel on connaît les valeurs des p variables $\boldsymbol{x} = \left\lfloor x^1, \cdots, x^p \right\rfloor$, on entre alors dans l'étape de classement.

La façon la plus simple consiste à calculer sa **distance de Mahalanobis** aux barycentres $\boldsymbol{g}_k$ de toutes les classes G_k, $(k = 1, , q)$. Cette distance vaut :

$$D^2(\boldsymbol{e}, \boldsymbol{g}_k) = (\boldsymbol{x} - \boldsymbol{g}_k)^{\mathrm{T}} \mathbf{W}^{-1} (\boldsymbol{x} - \boldsymbol{g}_k) = \boldsymbol{x}^T\mathbf{W}^{-1}\boldsymbol{x} - 2\boldsymbol{g}_k^T\mathbf{W}^{-1}\boldsymbol{x} + \boldsymbol{g}_k^T\mathbf{W}^{-1}\boldsymbol{g}_k$$

On classe alors $\boldsymbol{e}$ dans la classe dont il est le plus proche. La règle va donc consister à chercher à minimiser $D^2(\boldsymbol{e}, \boldsymbol{g}_k)$. Puisque $\boldsymbol{x}^T\mathbf{W}^{-1}\boldsymbol{x}$ ne dépend pas de la classe k, ceci est équivalent à choisir G_{k^*} le groupe auquel $\boldsymbol{e}$ prenant les valeurs $\boldsymbol{x} = \left\lfloor x^1, \cdots, x^p \right\rfloor$, a le plus de chance d'appartenir tel que :

$$k^* = \underset{k\in\{1,\ldots q\}}{\operatorname{argmax}} 2\boldsymbol{g}_k^{\mathrm{T}}\mathbf{W}^{-1}\boldsymbol{x} - \boldsymbol{g}_k^T\mathbf{W}^{-1}\boldsymbol{g}_k$$

Cette règle est linéaire par rapport aux coordonnées de $\boldsymbol{e}$. On dispose donc pour chaque classe k d'une **fonction de classement** ou **fonction discriminante** : on calcule pour chaque observation les q valeurs de ces fonctions linéaires et on affecte l'observation à la classe k pour laquelle la fonction de classement possède la valeur maximale puisqu'elle correspond à la distance minimale $D^2(\boldsymbol{e}, \boldsymbol{g}_k)$.

On peut noter que la distance de Mahalanobis, ainsi utilisée, fournit une affectation (en choisissant la plus petite distance) mais pas un *degré de crédibilité* de cette affectation : pour l'obtenir il va falloir imposer une propriété supplémentaire à la distribution des observations. Une autre insuffisance des règles géométriques provient du fait qu'elles conduisent à des affectations erronées dès que les dispersions des classes sont très différentes entre elles : l'usage d'une même métrique quelle que soit la classe n'est alors pas justifiée. Il faut donc envisager des règles d'affectation probabilistes : c'est ce que nous allons aborder maintenant.

3.3 Approche probabiliste

Nous allons définir successivement les probabilités d'appartenance à une population et leur usage pour le classement d'observations dans la suite naturelle de l'AFD présentée dans la section précédente.

Recherche d'une probabilité a posteriori

On modélise les variables Y, $\boldsymbol{X} = \left(X^1, \cdots, X^p\right)$ par des variables aléatoires réelles, en gardant les mêmes notations par convention. L'objectif est de produire une **règle d'affectation** qui permette de prédire, pour un individu $\boldsymbol{e}$ donné prenant les valeurs $\boldsymbol{x} = \left\lfloor x^1, \cdots, x^p \right\rfloor$, la valeur appartenant à $\{y_1, \cdots, y_K\}$ associée de Y à partir des valeurs prises par $\boldsymbol{x}$. On suppose que les q classes sont en proportion $p_1, p_2, \cdots, p_K$ (qui fournissent donc des **probabilités a priori** d'appartenance à l'une des classes) dans la population totale et que la distribution de $\boldsymbol{X}$ dans chaque classe G_k, $(k = 1, \cdots, q)$ est définie par une densité (ou une loi discrète) $f_k(\boldsymbol{x}) = P[\boldsymbol{X} = \boldsymbol{x} \mid G_k]$.

Pour un point $\boldsymbol{x} = \left\lfloor x^1, \cdots, x^p \right\rfloor$ la probabilité qu'il provienne de la classe G_k est donnée par la formule de Bayes*

$$P(G_k \mid \boldsymbol{x}) = \frac{p_k f_k(\boldsymbol{x})}{\sum_{i=1}^{q} p_i f_i(\boldsymbol{x})}; k = 1, \ldots, q$$

La règle bayesienne conduit à affecter l'observation à la classe k dont la **probabilité a posteriori** $P(G_k \mid \boldsymbol{x})$ est maximale. Comme les dénominateurs de cette expression sont les mêmes quelle que soit la classe k, on doit chercher celle pour laquelle $p_k f_k(\boldsymbol{x})$

est maximale, et alors $Y = y_k$. Toute la problématique de l'analyse discriminante revient alors à proposer une estimation de la quantité $f_k(\boldsymbol{x})$. On peut alors se donner une famille de lois de probabilité pour $f_k(\boldsymbol{x})$; par exemple la loi Normale, comme nous le verrons ci-après.

Loi Normale multidimensionnelle

Si $\boldsymbol{X}$ suit une loi $N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ dans chaque classe k :

$$f_k(\boldsymbol{x}) = P[\boldsymbol{X} = \boldsymbol{x} \mid G_k] = \frac{1}{(2\pi)^{p/2}\left(|\boldsymbol{\Sigma}_k|\right)^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right]$$

On estime généralement les matrices $\boldsymbol{\Sigma}_k$ par $\mathbf{W}_k/(n_k-1)$ et $\boldsymbol{\mu}_k$ par $\boldsymbol{g}_k$, $k = 1, \cdots, q$.

La règle bayesienne consistant à rechercher $\max_k(p_k f_k(\boldsymbol{x}))$, on peut montrer que le groupe auquel $\boldsymbol{e}$ prenant les valeurs $\boldsymbol{x} = \left\lfloor x^1, \cdots, x^p \right\rfloor$, a le plus de chance d'appartenir est G_{k^*} avec :

$$k^* = \underset{k\in\{1,\ldots q\}}{\operatorname{argmin}} (\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) - 2\ln p_k + \ln\left(|\boldsymbol{\Sigma}_k|\right)$$

On voit que :

- si les matrices de variance-covariance $\boldsymbol{\Sigma}_k$ ne sont pas égales, la règle est alors *quadratique* et qu'il faut comparer q fonctions quadratiques de $\boldsymbol{x}$.
- si on suppose l'hypothèse d'**homoscédasticité*** [4] satisfaite i.e. "$\boldsymbol{\Sigma}_1 = \cdots = \boldsymbol{\Sigma}_K = \boldsymbol{\Sigma}$" alors la règle devient *linéaire* :

$$\begin{aligned} k^* &= \underset{k\in\{1,\ldots q\}}{\operatorname{argmin}} (\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) - 2\ln p_k + \ln\left(|\boldsymbol{\Sigma}_k|\right) \\ &= \underset{k\in\{1,\ldots q\}}{\operatorname{argmin}} \Delta_p^2(\boldsymbol{x}, \boldsymbol{\mu}_k) - 2\ln p_k + \ln\left(|\boldsymbol{\Sigma}_k|\right) \end{aligned}$$

  où $\Delta_p^2(\boldsymbol{x}, \boldsymbol{\mu}_k)$ est la distance de Mahalanobis théorique entre $\boldsymbol{x}$ et $\boldsymbol{\mu}_k$, $\ln\left(|\boldsymbol{\Sigma}_k|\right)$ est une constante et $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$ que l'on obtient dans le calcul ne dépend que de l'observation ; si bien que l'on peut montrer que :

$$k^* = \underset{k\in\{1,\ldots q\}}{\operatorname{argmax}} \left\{ \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p_k \right\}$$

  Dans la formule ci-dessus, on voit que, si les probabilités a priori p_k sont égales, cette règle est strictement identique à la règle géométrique $-2\boldsymbol{g}_k^T\mathbf{W}^{-1}\boldsymbol{x} - \boldsymbol{g}_k^T\mathbf{W}^{-1}\boldsymbol{g}_k$ vue dans la partie 10.3.2. Par conséquent la règle géométrique (où $\boldsymbol{\mu}_k$ et $\boldsymbol{\Sigma}$ sont remplacés par leurs estimations $\boldsymbol{g}_k$ et $\mathbf{W}$)) est aussi optimale dans la situation gaussienne sous hypothèse d'homoscédasticité.

[4] L'hypothèse d'*homoscédasticité* signifie que la matrice de variance-covariance est la même pour toutes les classes, c'est-à-dire que $\Sigma_1 = \cdots = \Sigma_K = \Sigma$. Autrement dit, la dispersion des variables et leurs relations (covariances) ne dépendent pas de la classe. Sous cette hypothèse, la règle de classification devient linéaire, ce qui simplifie le calcul des frontières entre les classes.

Une autre approche que celle présentée ici, consiste à chercher un modèle pour la *probabilité a posteriori* $P(G_k \mid \boldsymbol{x})$. C'est ce que nous allons voir avec la régression logistique dans la section 4.

4 Régression logistique

Nous n'aborderons dans cette partie que la **régression logistique** binaire. Elle fait partie des méthodes de classification supervisée ; mais elle peut aussi s'inscrire dans le cadre du **modèle linéaire généralisé** (GLM, en abrégé pour le terme anglais). Pour plus de détails techniques, nous renvoyons aux ouvrages de Dobson (2002), Hosmer & Lemeshow (2000) et d'Agresti (1990).

4.1 Principes de la régression logistique binaire

On considère que la variable Y à expliquer ou prédire prend deux valeurs 0 ou 1 et qu'il existe p variables explicatives $X^1, \cdots, X^p$. L'ensemble des individus vérifiant $\{Y = 1\}$ correspondra donc à la classe G_1 tandis que l'ensemble des individus vérifiant $\{Y = 0\}$ correspondra à la classe G_2.

Partant des données, on souhaite estimer la probabilité inconnue qu'un individu $\boldsymbol{e}$ vérifiant $\boldsymbol{X} = \boldsymbol{x} = \left\lfloor x^1, \cdots, x^p \right\rfloor$ appartienne à la classe G_1 :

$$\pi(\boldsymbol{x}) = P(G_1 \mid \boldsymbol{x}) = P[Y = 1 \mid X = \boldsymbol{x}]$$

Si l'estimation de $\pi(\boldsymbol{x})$ est supérieure à 0.5, alors $\boldsymbol{e}$ a plus de chances d'appartenir à G_1.

Définitions : On appelle

- **transformation logit**, la fonction :

$$\text{logit}(y) = \ln\left(\frac{y}{1-y}\right) \in \mathbb{R},\ y \in]0, 1]$$

  Son inverse est la fonction : $\text{logit}^{-1}(x) = \frac{\exp(x)}{1+\exp(x)} \in]0, 1], x \in \mathbb{R}$

- **modèle de régression logistique**, la modélisation :

$$\text{logit}(\pi(\boldsymbol{x})) = \beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p$$

  Où $\beta_0, \cdots, \beta_p$ désignent $p + 1$ réels inconnus appelés *paramètres du modèle*.

Dans le cadre du modèle linéaire généralisé, $\text{logit}(\pi)$ est une des fonctions de lien largement utilisée[5], ne serait-ce qu'en raison de son lien avec l'analyse discriminante. Elle est particulièrement bien adaptée à la modélisation d'une courbe en S dite courbe sigmoïdale. Cette modélisation simple n'est pas unique d'autres fonctions de lien existent,

[5] Sur l'utilisation du modèle Logit, nous renvoyons à l'article de Sanders 1997 (https ://-doi.org/10.4000/cybergeo.1547) en lien avec le chapitre 2.2.8 sur la classification hiérarchique des établissements. Aussi, à l'article Disser et al. 2014 (https ://doi.org/10.1016/j.jas.2013.10.034) qui utilise un modèle logit pour classer des artefacts métalliques selon leur procédé de production d'origine (direct ou indirect), en lien avec les notions de provenance des matériaux archéologiques (*cf.* 2.1.2).

comme : la fonction *probit* ou *inverse de la loi Normale* $N(0,1)$; la fonction *Log-log* qui a la forme $\ln(-\ln(1-\pi))$. En utilisant la fonction de lien logit, on en déduit l'expression de $\pi(\boldsymbol{x})$ suivante :

$$\pi(\boldsymbol{x}) = \text{logit}^{-1}(\beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p) = \frac{\exp(\beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p)}{1 + \exp(\beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p)}$$

4.2 Estimation des paramètres : cas de la fonction de lien logit

En procédant par substitution dans l'expression de $\pi(\boldsymbol{x})$ ci-dessus, des estimations ponctuelles de $\beta_0, \cdots, \beta_p$ amèneront une estimation ponctuelle de $\pi(\boldsymbol{x})$. Les paramètres de la régression logistique sont estimés par maximum de vraisemblance*. On note $\hat{\beta}_0, \ldots, \hat{\beta}_p$ ces estimateurs. En pratique, ils sont approchés avec l'algorithme de Newton-Raphson (voir par exemple Hosmer & Lemeshow (2000)).

Une estimation ponctuelle de $\pi(\boldsymbol{x})$ est donnée par :

$$\hat{\pi}(\boldsymbol{x}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x^1 + \cdots + \hat{\beta}_p x^p)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x^1 + \cdots + \hat{\beta}_p x^p)}$$

Le groupe auquel $\boldsymbol{e}$ a le plus de chances d'appartenir est :

$$G^* = \begin{array}{l} G_1 \text{ si } \hat{\pi}(\boldsymbol{x}) \geq 0.5 \\ G_2 \text{ sinon} \end{array}$$

4.3 Taux d'erreur de classement

Après avoir obtenu les prévisions grâce au modèle de régression logistique, il faut en évaluer l'efficacité à l'aide de critères adaptés aux méthodes de classification supervisée. Le taux d'erreur de classement en est un. Il est donné par le nombre d'individus mal affectés sur le nombre total d'individus. Plus il est proche de 0, meilleur est la qualité prédictive du modèle. On convient que la qualité de la classification est mauvaise lorsque ce taux est supérieur à 0.5.

Il existe de nombreuses autres critères d'évaluation externe : taux de succès, sensibilité (encore appelé taux de vrais positifs), spécificité (proportion de vrais négatifs détectés), aire sous la courbe ROC, …(*cf.* 2.2.5) Enfin, comme dans toute modélisation, les étapes de validation interne et externe ne doivent pas être négligées. Sur ces deux points fondamentaux, non abordés ici, nous renvoyons vers les ouvrages déjà cités.

5 Bilan

Les méthodes de classification supervisée permettent d'affecter un nouvelle individu à un groupe préalablement défini.

- kNN est une méthode non paramétrique qui calcule la similarité entre l'observation à classer et les différentes observations du jeu de donnée, aucune estimation

de paramètres de modèle. C'est un algorithme simple avec l'inconvénient qu'il doit garder en mémoire l'ensemble des observations pour pouvoir effectuer sa prévision. Ainsi il faut faire attention à la taille du jeu d'entrainement.

- L'analyse discriminante et la régression logistique sont deux méthodes de classement paramétriques concurrentes. Elles ont chacune des avantages et des inconvénients inhérents aux suppositions sous-jacentes de chacune ; il sera intéressant, voire important de confronter leurs résultats si les deux méthodes peuvent être envisagées.
- En présence de variables explicatives quantitatives quasi-colinéaires, l'analyse discriminante comme la régression logistique conduiront à des modèles instables. Des remèdes pourront être :
  - la régression logistique de type PLS ou Ridge,
  - l'analyse discriminante de type PLS ou Ridge,
  - l'utilisation de méthodes non paramétriques.

Parmi les méthodes de classification supervisée non abordées ici, citons :

- les arbres de décision,
- les forêts aléatoires,
- les machines à vecteurs de support ou séparateurs à vaste marge support,
- les réseaux de neurones.