Problématique de l'apprentissage statistique principes généraux de l'analyse discriminante

Gilles Celeux
Inria Rhône-Alpes

Apprentissage statistique

La problématique est décisionnelle. Il s'agit de prédire une valeur, de diagnostiquer un état ou d'évaluer un risque
à partir de données multidimensionnelles constituées de descripteurs sur un ensemble d'apprentissage.
Lorsque la variable à prédire est quantitative, on est en situation de régression
Lorsqu'elle la variable à prédire est discrète, on est en situation d'analyse discriminante ou de classification supervisée.

Analyse discriminante

On veut prédire l'appartenance d'un individu, décrit par d variables explicatives à un groupe parmi g groupes $G_1, \ldots, G_g$ disjoints, définis a priori.

Pour ce faire on dispose d'un ensemble d'apprentissage

$$A = \{(\mathbf{x}_i, z_i), \ldots, (\mathbf{x}_n, z_n), \mathbf{x}_i \in \mathbf{R}^d \text{ et } z_i \in \{1, \ldots, g\}\},$$

$\mathbf{x}_i$ contient les valeurs du $i$ème individu sur les d descripteurs,
z_i indique le numéro de groupe auquel il appartient.

L'échantillon étiqueté A va être utilisé pour construire une fonction de décision $\delta(\mathbf{x})$ qui à tout vecteur de $\mathbf{R}^d$ va associer un des g groupes a priori de sorte à minimiser le risque d'erreur.

Le modèle probabiliste

On définit

- les probabilités a priori des groupes $p(G_1), \ldots, p(G_g)$,
- les densités de probabilité par groupe $p(\mathbf{x}|G_1), \ldots, p(\mathbf{x}|G_g)$
- les coûts de mauvais classement d'une observation du groupe G_k dans le groupe G_ℓ $c(\ell|k)$ pour $\ell, k = 1, \ldots, g$, avec $c(k|k) = 0$.

Toute fonction de décision δ définit une partition de $\mathbf{R}^d$, $(D_1, \ldots, D_g)$ avec $D_k = \{\mathbf{x} \in \mathbf{R}^d / \delta(\mathbf{x}) = k\}$. Le risque moyen de δ s'écrit

$$R(\delta) = \sum_{k=1}^{g} p(G_k) \sum_{\ell=1}^{g} c(\ell|k) \int_{D_\ell} p(\mathbf{x}|G_k) d\mathbf{x}.$$

La règle de Bayes

La règle de décision optimale δ^*, dite règle de Bayes, est celle qui minimise $R(\delta)$. Par exemple, dans le cas où $g = 2$, elle s'écrit

$$\begin{aligned}\delta^*(\mathbf{x}) &= 1 \Longleftrightarrow c(2|1)p((G_1)p(\mathbf{x}|G_1) \geq c(1|2)p(G_2)p(\mathbf{x}|G_2)\\ \delta^*(\mathbf{x}) &= 2 \Longleftrightarrow c(2|1)p((G_1)p(\mathbf{x}|G_1) < c(1|2)p(G_2)p(\mathbf{x}|G_2).\end{aligned}$$

Dans le cas d'égalité des coûts de mauvais classement, elle revient à affecter un point $\mathbf{x}$ au groupe de plus grande probabilité conditionnelle $p(G_k|\mathbf{x}) \propto p(G_k)p(\mathbf{x}|G_k)$.

Les méthodes de l'analyse discriminante diffèrent par les hypothèses faites sur les densités par groupe ou sur les probabilités conditionnelles des groupes et par les méthodes d'estimation de leurs caractéristiques.

Les schémas d'échantillonnage

Comment l'ensemble d'apprentissage A a-t-il été constitué ?

Les deux schémas les plus répandus sont

- le schéma de mélange où A est tiré au hasard dans la population étudiée. Les $\mathbf{x}_i$ sont des réalisations d'une loi de mélange
$$p(\mathbf{x}) = \sum_{k=1}^{g} p(G_k)p(\mathbf{x}|G_k).$$
Les probabilités a priori sont estimées par les proportions n_k/n, n_k désignant le nombre de points issus du groupe G_k dans A.
- le schéma rétrospectif ou cas-témoin où l'échantillon A est la concaténation de g échantillons indépendants de taille fixée n_k et de loi $p(\mathbf{x}|G_k)$ pour $k = 1, \ldots, g$. Le schéma cas-témoin est bien adapté à la prise en compte de groupes a priori rares, mais il demande que les probabilités a priori des groupes soient connus.

Des stratégies différentes pour approcher la règle de Bayes

- Estimation des densités par groupe $p(\mathbf{x}|G_k)$
  - hypothèses paramétriques (densités gaussiennes par exemple)
  - approche non paramétrique (méthode des k plus proches voisins, par exemple)
- Estimation directe des probabilités conditionnelles d'appartenance aux groupes $p(G_k|\mathbf{x})$.
  - hypothèses semi-paramétriques (régression logistique)
  - approche non paramétrique (discrimination par arbres)
- Un autre point de vue est de chercher directement une fonction de décision d'une forme donnée. C'est en particulier le cas des réseaux de neurones ou de la méthode SVM.

Une méthode de référence : l'ADL

L'analyse discriminante linéaire (ADL) suppose que les densités par groupe sont gaussiennes $\mathcal{N}(\mu_k, \Sigma)$ de même matrice variance Σ.
Elle crée des séparations linéaires entre groupes. Si $g = 2$, il vient

$$\delta(\mathbf{x}) = 1 \Longleftrightarrow (\mathbf{x} - \frac{\mu_1 + \mu_2}{2})^t \Sigma^{-1} (\mu_1 - \mu_2) \geq 0.$$

Les paramètres de l'ADL sont estimés par leurs valeurs empiriques

$$\hat{\mu}_k = \bar{\mathbf{x}}_k, \quad k = 1, \ldots, g$$

$$\hat{\Sigma} = \frac{\Sigma_{k=1}^{g} \Sigma_{i/z_i=k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^t}{n - g}.$$

L'ADL est très utilisée pour les raisons suivantes :

- Bon compromis pertinence/complexité (dilemme biais-variance).
- La sélection de variables se fait de manière quasi optimale.
- L'ADL fournit des résultats stables et robustes.

Sélection de variables

Chacune des d variables explicatives apporte de l'information discriminante et du bruit d'échantillonnage. En pratique on distingue trois types de variables

- les variables utiles,
- les variables redondantes dont l'information discriminante est essentiellement contenue dans d'autres variables,
- les variables nuisibles.

Ainsi la sélection de variables est souvent une étape indispensable pour obtenir une fonction de décision fiable.
Certaines méthodes comme les arbres de décision ou la régression logistique réalisent cette sélection intrinséquement, mais en général il faut sélectionner les variables préalablement à la construction de la fonction de décision.

Critères de sélection

Dans le cas où $g = 2$, on dispose de la distance de Mahalanobis D, critère à maximiser qui mesure la séparation des deux groupes,

$$D^2 = ||\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2||^2_{W^{-1}} = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)W^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^t.$$

Dans le cas général, l'équivalent de la distance de Mahalanobis est le lambda de Wilks

$$\Lambda = \frac{|W|}{|T|},$$

$$T = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^t$$

$$W = \frac{1}{n}\sum_{k=1}^{g}\sum_{i/z_i=k}(\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^t.$$

Ce critère à minimiser mesure l'homogénité moyenne des groupes. Ces critères sont essentiellement adaptés à l'ADL.

Procédures de sélection

- Sélection exhaustive par un algorithme branch and bounds possible pour 2 groupes et quelques dizaine de variables candidates.
- Sélection ascendante : on sélectionne la meilleure, puis le meilleur couple contenant la première variable, etc. Rapide mais sous-optimale.
- sélection descendante : Cette procédure part de toutes les variables, élimine la moins bonne variable, puis la deuxième moins bonne variable sachant que la première a été éliminée, etc.

Pour l'ADL, on peut se placer dans un contexte d'analyse de variance et définir une procédure de sélection ascendante minimisant le lambda de Wilks. Cela permet la remise en cause de variables qui perdraient leur pouvoir discriminant au fur à mesure de la sélection.

Procédure d'arrêt de la sélection

- Une technique universelle consiste à minimiser le taux d'erreur de classement estimée par validation croisée. L'inconvénient majeur de cette méthode est sa lenteur pour les techniques lourdes.

- Pour les techniques paramétriques, on peut considérer des critères de vraisemblance pénalisée comme le critère

$$AIC(J) = -2\ln(L(m(J)) + 2M(J)$$

ou le critère BIC (Bayesian Information Criterion)

$$BIC(J) = -2\ln(L(m(J)) + M(J)\ln(n).$$

$L(m(J))$ étant la vraisemblance du modèle utilisé avec un ensemble de J variables, $M(J)$ étant le nombre de paramètres indépendants de ce modèle et n la taille de l'échantillon d'apprentissage A.

Évaluation des performances

La règle de décision construite à partir de A est destinée à être appliquée sur une population de taille potentiellement infinie. Il est important d'estimer les taux d'erreur de cette règle.
Quels taux d'erreur ? Les taux d'erreur conditionnels.

Deux mauvaises façons d'évaluer un taux d'erreur.

- Estimation par resubstitution : on évalue le taux d'erreur de la règle de décision δ sur le même ensemble A qui a servi à la construire : biais d'optimisme.
- Estimation par un calcul théorique sous les hypothèses du modèle de la règle de décision δ. Là encore, biais d'optimisme car non prise en compte des fluctuations d'échantillonnage, ni surtout remise en compte du modèle.

Techniques de rééchantillonnage

L'échantillon-test : le taux d'erreur est mesuré sur un échantillon T disjoint de A et tiré au tirage au hasard dans l'échantillon initial.

La validation croisée : on divise A aléatoirement en L parties égales. Pour $\ell = 1, \ldots, L$, on construit la règle de décision sur la base de A privé de sa ℓème partie et on affecte cette dernière aux groupes a priori suivant cette règle de décision.
Si $L = n$ on retombe sur la procédure standard de leave-one-out.
Si, à l'opposé, $L = 2$ on définit ainsi une procédure d'half sampling.

Plus L est choisi grand plus le biais d'estimation est petit, mais plus la variance est importante. Un choix de L entre 5 et 10 est classique.

Le bootstrap substitue à la distribution de probabilité inconnue des données, la distribution empirique de A. Il construit B (typiquement $B = 100$) échantillons bootstrap de taille n par tirage aléatoire avec remise dans A.

Pour chaque échantillon bootstrap b, on obtient deux taux d'erreur, le taux apparent sur l'échantillon b et une estimation du taux d'erreur réels sur l'échantillon A.

La différence entre ces deux taux d'erreur est une estimation du biais d'optimisme du taux apparent associé au pseudo-échantillon b.

La moyenne O_B de ces optimismes sur les B échantillons bootstrap est l'estimateur de l'optimisme du taux apparent d'erreur calculé par resubstitution E_A. L'estimateur bootstrap du taux d'erreur est donc

$$E_B = E_A + O_B.$$

Évaluation des performances

La règle de décision construite à partir de A est destinée à être appliquée sur une population de taille potentiellement infinie. Il est important d'estimer les taux d'erreur de cette règle.
Quels taux d'erreur ? Les taux d'erreur conditionnels.

Deux mauvaises façons d'évaluer un taux d'erreur.

- Estimation par resubstitution : on évalue le taux d'erreur de la règle de décision δ sur le même ensemble A qui a servi à la construire : biais d'optimisme.
- Estimation par un calcul théorique sous les hypothèses du modèle de la règle de décision δ. Là encore, biais d'optimisme car non prise en compte des fluctuations d'échantillonnage, ni surtout remise en compte du modèle.

Sélection de modèles et rééchantillonnage

- De nombreuses méthodes utilisent un paramètre externe de réglage.
- Pour fixer la valeur de ce type de paramètre, le plus simple est de disposer de trois échantillons : un échantillon A d'apprentissage A, un échantillon de réglage R du paramètre externe et d'un échantillon test T.
- En réalité, on est souvent obligé d'avoir recours à la validation croisée pour réaliser ces tâches.
- Il faut donc faire une double validation croisée : la première pour calculer le paramètre externe optimal, l'autre pour évaluer les taux d'erreur de la règle finale.

La régression logistique

Posant $\Pi(\mathbf{x}) = p(G_1|\mathbf{x})$, l'équation de la régression logistique linéaire est

$$\text{logit}(\Pi(\mathbf{x})) = \ln\left(\frac{\Pi(\mathbf{x})}{1-\Pi(\mathbf{x})}\right) = \beta_0 + \sum_{j=1}^{d} \beta_j x^j.$$

La régression logistique linéaire est une méthode semi-paramétrique qui rentre dans la famille des modèles linéaires généralisés avec une fonction de lien logit.

Pour estimer les paramètres $\beta = (\beta_0, \beta_1, \ldots, \beta_d)$, les équations non linéaires du maximum de vraisemblance sont résolues par un algorithme de type Newton-Raphson indépendamment du schéma d'échantillonnage.

La régression logistique bénéficie des outils des modèles linéaires généralisés qui permettent notamment la sélection de variables par des tests de nullité des coefficients β ainsi que des tests d'adéquation du modèle (test du rapport de vraisemblance, test de Wald, . . .).

La régression logistique (2)

- La généralisation de la régression logistique au cas $(g > 2)$ se fait en général par rapport à un groupe de référence.

- Du fait qu'elle modélise directement les probabilités conditionnelles d'appartenance aux groupes, elle permet le traitement conjoint de variables continues et dicrètes.

- Elle est très utilisée en médecine à cause des risques relatifs estimés (*odds-ratio*) pour les variables explicatives binaires. Ce risque relatif estimé s'écrit pour une variable j

$$RRE(j) = \exp(\beta_j).$$

  Les individus pour lesquels cette variable j prend la valeur 1 ont une probabilité d'appartenir au groupe à risque $RRE(j)$ fois supérieure à ceux pour lesquels elle vaut 0.

La régression logistique (3)

- Les performances et les qualités (robustesse, stabilité...), de la régression logistique sont analogues à celles de l'ADL.
- La régression logistique n'est pas restreinte au cadre linéaire.
  - Il est possible de considérer un développement quadratique du logit de $\Pi(\mathbf{x})$
  - ou un développement d'ordre supérieur. Cette possibilité est notamment intéressante lorsque les descripteurs sont qualitatifs pour tenir compte des interactions non linéaires.
  - Voire un développement non paramétrique utilisant une base de fonctions splines.
- Ainsi la régression logistique peut être vue dans une démarche analogue à celle des Support Vector Machines.

La méthode des K plus proches voisins (1)

- Parmi les méthodes non paramétriques les méthodes de lissage (méthode des noyaux) sont peu répandues en apprentissage statistique multidimensionnel.

- La méthode de référence est la méthode des K plus proches voisins.

- Dans sa version de base, elle se résume ainsi: Pour chaque vecteur $\mathbf{x}$ à classer, on examine ses K plus proches voisins dans l'échantillon d'apprentissage A et on l'affecte au groupe majoritaire.

- Le choix du nombre K de voisins est primordial. Une bonne stratégie, consiste à le choisir de sorte à minimiser le taux d'erreur évalué par la validation croisée standard.

La méthode des K plus proches voisins (2)

- Cette stratégie revient à faire une estimation locale de la densité autour d'un petit voisinage de volume V du point $\mathbf{x}$ (schéma d'échantillonnage de mélange)

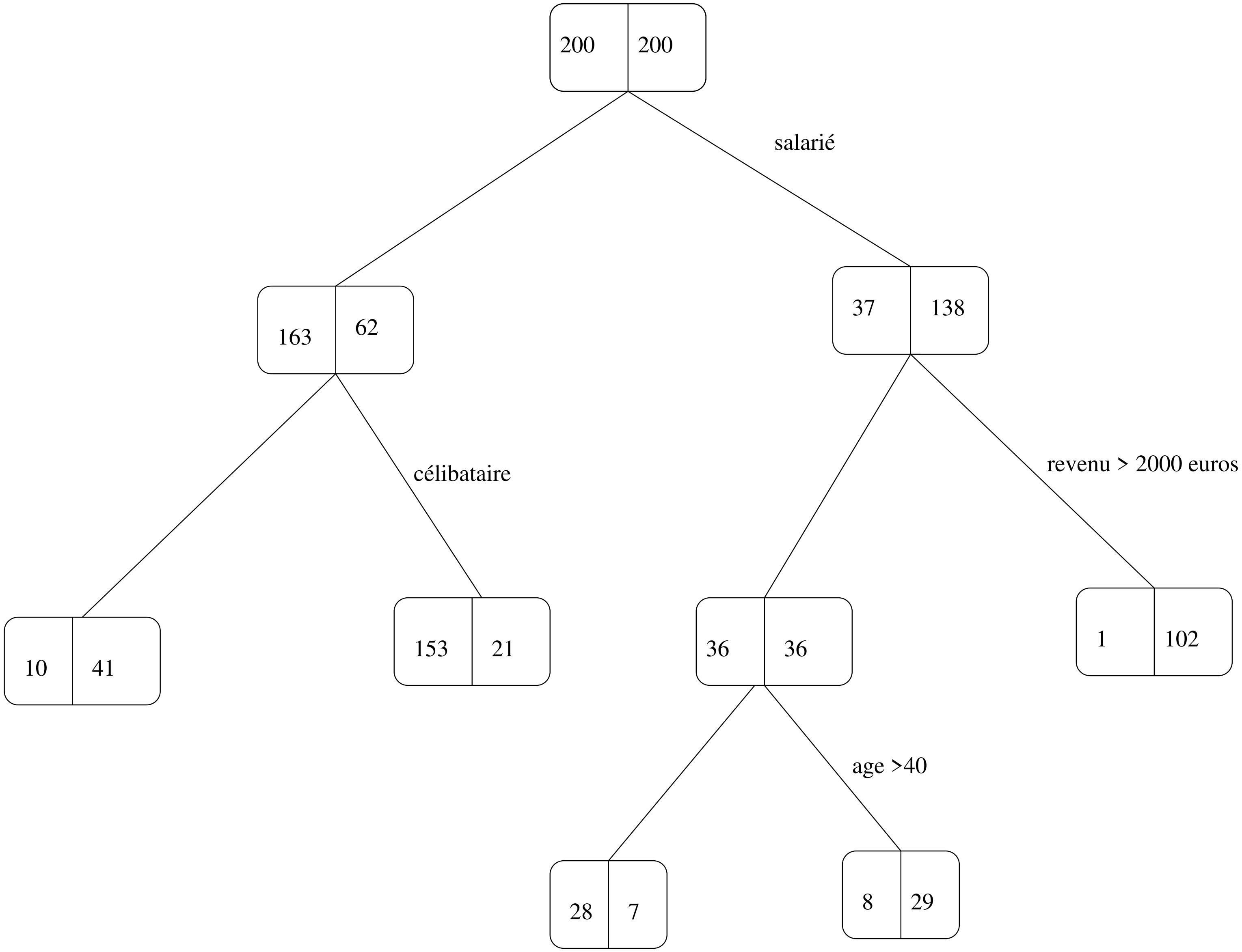
$$p(G_k|\mathbf{x}) = \frac{p(\mathbf{x}|G_k)p(G_k)}{p(\mathbf{x})} \approx \frac{(a_k/(n_k V))(n_k/n)}{a/(nV)} = \frac{a_k}{a}$$

  où a est le nombre de points de A dans le petit voisinage et où a_k est le nombre de points de A émanant du groupe G_k dans ce petit voisinage.

- Pour être efficace la méthode des K plus proches voisins réclame une taille d'échantillon importante.

- Cette méthode présente surtout de l'intérêt dans les situations hautement non linéaires.

Arbres de décision (1)



Arbres de décision (2)

- Critère d'évaluation d'une coupure se fonde sur la notion d'impureté d'un nœud :

$$i(t) = -\sum_{k=1}^{g} p(k|t) \log p(k|t),$$

ou

$$i(t) = -\sum_{k \neq \ell} p(k|t) p(\ell|t).$$

Le critère d'évaluation d'une coupure est alors la réduction d'impureté du nœud

$$\Delta i(t) = i(t) - p_l i(t_l) - p_r i(t_r),$$

p_l (resp. p_r) représentant la probabilité empirique (poids) de tomber dans le nœud gauche (resp. droit).

Arbres de décision (3)

- Le point crucial est l'élagage des branches inutiles. Cette question relève de la résolution du dilemme biais-variance d'une règle de décision.

- La bonne stratégie consiste à construire l'arbre le plus profond possible et à l'élaguer à l'aide d'un échantillon-test ou par une procédure de validation croisée.

- La procédure la plus efficace utilise la notion de coût-complexité d'un arbre T définie par

$$C_\alpha(T) = R(T) + \alpha|T|$$

  $R(T)$ étant le coût de classement de l'arbre T estimé par resubstitution, $|T|$ étant son nombre de nœuds terminaux et α étant un scalaire positif. Pour tout α, il existe un arbre de coût-complexité minimum.

Arbres de décision (4)

- Les arbres de décision réalisent une règle de décision simple, compacte, non paramétrique et qui réalisent de manière intrinsèque une sélection des variables.
- Ils exigent des tailles d'échantillons importantes pour de bonnes performances.
- Leurs taux d'erreur sont rarement très bons.
- Leur gros défaut réside dans leur instabilité due au fait que le choix des premiers nœuds oriente fortement le direction de développement de l'arbre.
- On verra que le boosting réduit cette variabilité et améliore considérablement les performances.

Les réseaux de neurones

- Ces méthodes ne cherchent pas à modéliser la distribution de probabilité des groupes ou les probabilités conditionnelles d'appartenance aux groupes.
- Elles visent à construire directement des frontières entre les groupes par minimisation d'un critère des moindres carrés.
- Ainsi la méthode du perceptron recherche à partir des points de A un hyperplan séparateur entre les groupes.
- Même si les hypothèses de l'ADL ne sont pas vérifiés, pour peu que les groupes soient bien séparables par un hyperplan sur l'ensemble A, le perceptron le trouvera.
- À l'inverse, cette technique risque de passer assez loin de l'hyperplan séparateur optimal de l'ADL dans le cas où il n'est pas possible de séparer les groupes par un hyperplan sur A.

Le perceptron

Soit $\mathbf{x} = (1, x^1, \ldots, x^d)$, le perceptron cherche $\mathbf{w} = (w_0, \ldots, w_d)$ de sorte à construire une règle de décision de la forme

$$\delta(\mathbf{x}) = 1 \text{ si } h\left(\sum_{j=0}^{d} w_j f_j(x^j)\right) = 1$$
$$\delta(\mathbf{x}) = 2 \text{ si } h\left(\sum_{j=0}^{d} w_j f_j(x^j)\right) = 0$$

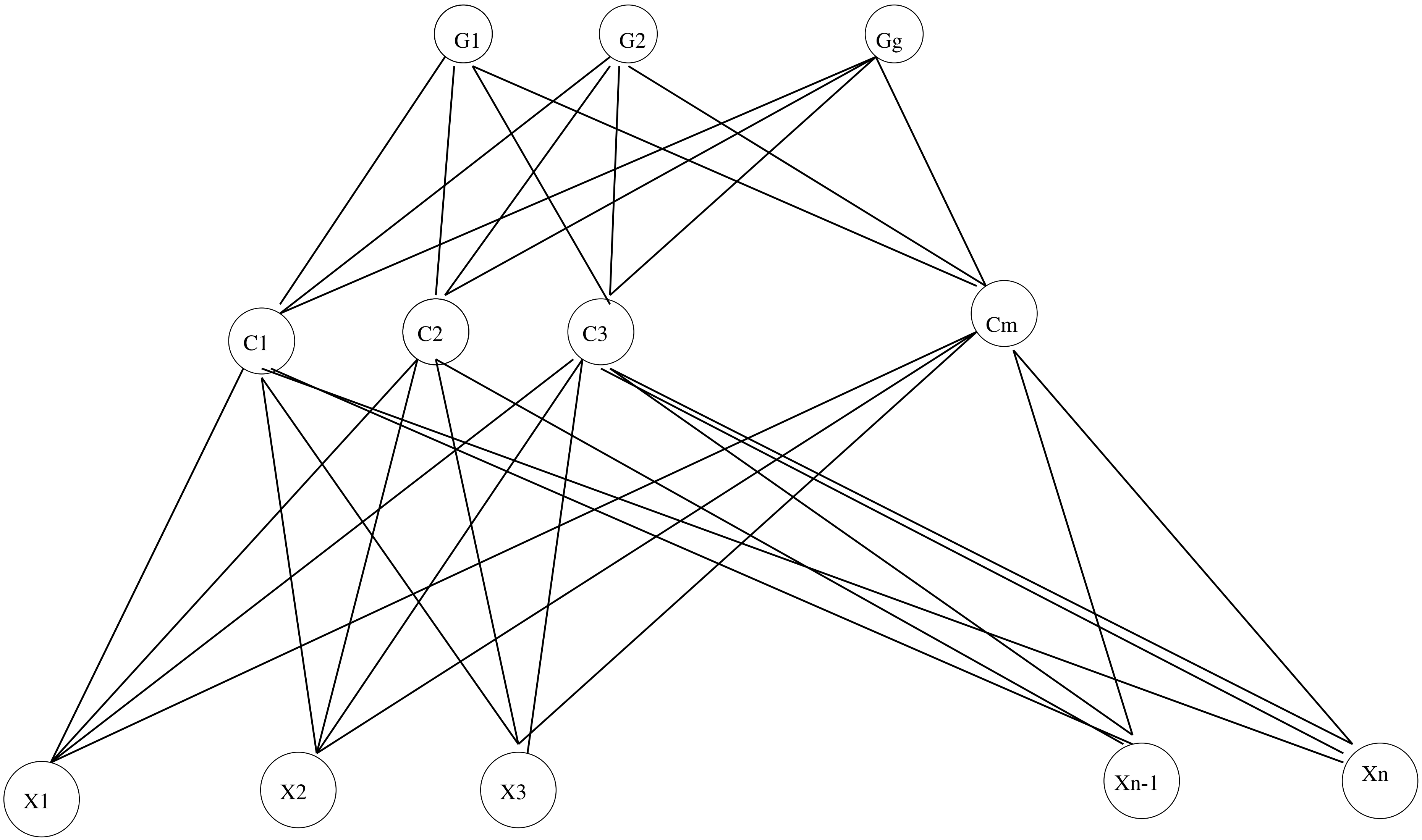
où la fonction h est la fonction d'Heaviside $h(a) = \begin{cases} 0 \text{ si } a < 0 \\ 1 \text{ si } a > 0, \end{cases}$ et où les f_j sont fixés à l'avance. La fonction à mimimiser est

$$E(\mathbf{w}) = \sum_{i \in C_A} \mathbf{w}^t \mathbf{f}_i y_i$$

où C_A ={points mal classés de A}, $\mathbf{f}_i = (1, f_1(x_i^1), \ldots, f_d(x_i^d))$; $y_i = 1$ si $i \in G_1$ et $y_i = -1$ sinon. L'algorithme est un algorithme séquentiel du gradient $\mathbf{w}^{r+1} = \mathbf{w}^r + \alpha(z_r - \delta_r)\mathbf{x}$, z_r étant le groupe de l'obs. présentée au réseau et δ_r son affectation avec le vecteur $\mathbf{w}^r$.

Le perceptron multicouche (1)

Le perceptron multicouche introduit une couche cachée et une fonction d'activation h moins abrupte que la fonction d'Heaviside.



Le perceptron multicouche (2)

Il se traduit par la construction de g fonctions de la forme

$$y_k(\mathbf{x}) = \tilde{h}\left[\sum_{\ell=0}^{m} \beta_{k\ell} h\left(\sum_{j=0}^{d} \alpha_{\ell j} x^j\right)\right], \quad k = 1, \ldots, g,$$

m étant le nombre de couches cachées. D'où la règle de décision

$$\delta(\mathbf{x}) = k \text{ ssi } k = \text{arg}\max(y_k(\mathbf{x})).$$

Ce réseau utilise 2 systèmes de poids adaptatifs et 2 fonctions d'activation. La fonction d'activation de sortie $\tilde{h}$ est la fonction softmax

$$\tilde{h}(a_k) = \frac{\exp(a_k)}{\Sigma_{\ell=1}^{g} \exp(a_\ell)}.$$

La fonction d'activation h de la couche cachée est

$$h(a) = \frac{1}{1+\exp(-a)}$$

ou

$$h(a) = \tanh(a)$$

.

Le perceptron multicouche (3)

- Un perceptron multicouche peut générer des séparations entre les groupes arbitrairement compliquées.
- Les poids β et α sont calculés pour minimiser l'erreur quadratique $R(\mathbf{w}) = \Sigma_{k=1}^{g} \Sigma_{i=1}^{n} \left(z_i^k - y_k(\mathbf{x}_i)\right)^2$, où ici $z_i^k = 1$ si $i \in G_k$ et 0 sinon.
- Cela se fait par l'algorithme de rétro-propragation du gradient.
- La mise en œuvre du perceptron multicouche demande tout un savoir-faire heuristique pour surmonter toutes ses difficultés.
- Le principal problème est le surapprentissage. On régularise les poids du réseau par un critère d'erreur pénalisée

$$\tilde{R}(\mathbf{w}) = R(\mathbf{w}) + \lambda \sum_l w_l^2$$

  où λ est un paramètre qui règle le lissage de la fonction de décision.

L'analyse discriminante prédictive

- Dans un cadre paramétrique, l'approche prédictive consiste à munir θ_k d'une loi a priori $p(\theta_k)$ non informative et à fonder la règle de décision sur la densité intégrée $p_k(\mathbf{x}) = \int p(\mathbf{x}|\theta_k, G_k)p(\theta_k)d\theta_k$.
- Pour l'ADL, on considère la loi impropre $p(\theta_k) \propto |\Sigma|^{-a/2}$ qui conduit à une densité prédictive t de Student multivarié à $n-g-a+2$ degrés de libertés. Cette distribution a des queues plus lourdes que la distribution gaussienne.
- En pratique, l'analyse discriminante linéaire prédictive prend son intérêt pour les petits échantillons et lorsque les tailles n_k sont déséquilibrées avec certaines valeurs très faibles.
- Dans le cas quadratique, la discrimination quadratique prédictive peut produire de bonnes performances avec des tailles d'échantillon faibles et irréalistes pour la discrimination quadratique classique.

L'analyse discriminante régularisée (ADR)

Dans le cadre gaussien, l'ADR fait dépendre l'estimation des matrices variances des groupes Σ_k de deux paramètres de régularisation λ et γ.

$$\hat{\Sigma}_k(\lambda, \gamma) = (1 - \gamma)\hat{\Sigma}_k(\lambda) + \gamma(\frac{tr(\hat{\Sigma}_k(\lambda))}{d})\mathbf{I},$$

o“u

$$\hat{\Sigma}_k(\lambda) = \frac{(1-\lambda)(n_k - 1)\hat{\Sigma}_k + \lambda(n-g)\hat{\Sigma}}{(1-\lambda)(n_k - 1) + \lambda(n-g)}.$$

- Le paramètre de complexité λ $(0 \leq \lambda \leq 1)$ contrle la contribution des estimateurs empiriques $\hat{\Sigma}_k$ et $\hat{\Sigma}$s,
- le paramètre γ $(0 \leq \gamma \leq 1)$ contrle le rétrécissement des valeurs propres vers l'égalité.
- Les paramètres λ et γ sont calculés de sorte à minimiser le taux d'erreur obtenu par validation croisée.

L'analyse discriminante par décomposition spectrale (ADDS)

ADDS se fonde sur la décomposition spectrale des matrices variances

$$\Sigma_k = \lambda_k D_k A_k D_k' \tag{1}$$

où $\lambda_k = |\Sigma_k|^{1/d}$, D_k est la matrice des vecteurs propres de Σ_k et A_k est une matrice diagonale dont la diagonale contient les valeurs propres normalisées de Σ_k en ordre décroissant. Le paramètre λ_k détermine le volume du groupe G_k, D_k son orientation et A_k sa forme.

En autorisant ou non ces quantités à varier entre les groupes, on obtient différents modèles plus ou moins parcimonieux et faciles à interpréter.

ADDS met en compétition 14 modèles et choisit celui qui fournit le plus petit taux d'erreur évalué par validation croisée, les paramètres des modèles étant estimés par l'approche du maximum de vraisemblance.

Modélisation par des mélanges

Si les groupes sont hétérogènes, on peut envisager de modéliser les densités par groupe par un mélange de lois gaussiennes

$$p(\mathbf{x}|G_k) = \sum_{\ell=1}^{r_k} \pi_{k\ell}\Phi(\mathbf{x}|m_{k\ell}, \Sigma_{k\ell})$$

Les proportions du mélange $\pi_{k\ell}$ vérifiant $\Sigma_{\ell=1}^{r_k} \pi_{k\ell} = 1$.

- Les g lois de mélange sont estimés par l'algorithme EM.
- Problème : nombre très important de paramètres. Parades :
  - Se restreindre à une matrice variance commune pour tous les composants du mélange et tous les groupes, $\Sigma_{k\ell} = \Sigma$, r_k le nombre de composants par groupe étant fixé.
  - Poser $\Sigma_{k\ell} = \sigma^2_{k\ell} I$. Les densités par groupe sont modélisées par un ensemble de "boules" gaussiennes à volumes libres. Estimation des r_k par minimisation du taux d'erreur évalué par validation croisée.

Fonctions radiales de base (FRD)

Les FDR proposent une architecture de réseau de la forme

$$y_k(\mathbf{x}) = \sum_{j=0}^{m} w_{kj} F_j(\mathbf{x}), \;\; k = 1, \ldots, g.$$

avec $F_0(\mathbf{x}) = 1$. Une fonction radiale type assurant que le réseau peut approcher toute fonction de manière optimale est

$$F_j(\mathbf{x}) = \exp\left(\frac{- \parallel \mathbf{x} - \mu_j \parallel^2}{2\sigma_j^2}\right).$$

L'apprentissage se fait en deux temps.

- Les paramètres des FRD sont estimés indépendamment des sorties,
- les poids de la couche cachée sont déterminés par une résolution linéaire pour minimiser l'erreur quadratique des sorties.

Fonctions radiales de base (2)

- L'art des FRD réside dans le choix des centres et des distances aux centres.
  - Cela peut se faire par un algorithme des centres mobiles
  - ou mieux par un mélange de lois normales estimé par l'algorithme EM. Les proportions du mélange étant écartées lors de la deuxième phase de l'pprentissage.
- La première phase exploratoire va fournir une représentation parcimonieuse des données d'entrée.
- Les FRD constituent un outil plus souple que le perceptron multicouche si les fonctions radiales F_j sont bien choisies.
- Mais une bonne utilisation des FDR demande un savoir faire heuristique encore plus important.