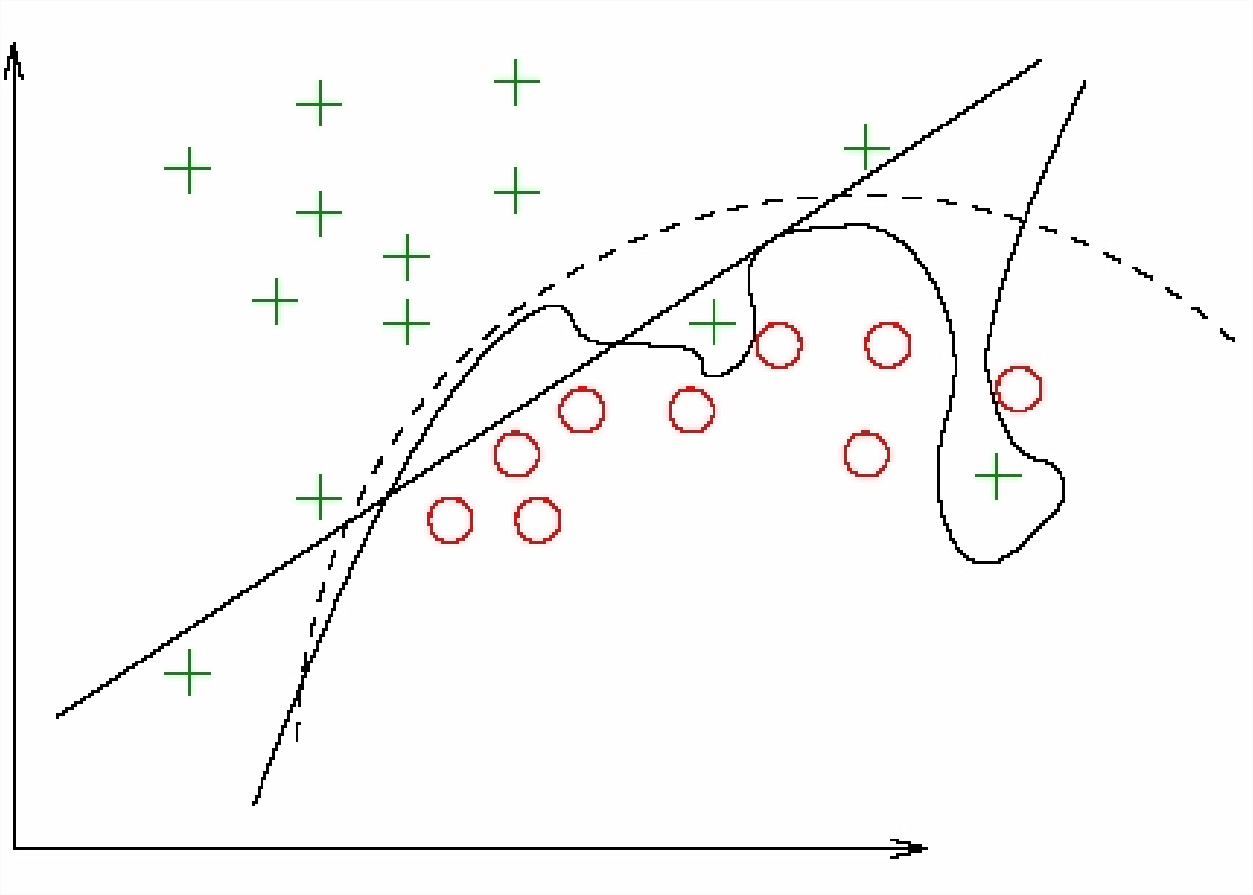
Apprentissage statistique

Apprentissage supervisé: suite

Cours pour non-spécialistes

Aurélien Garivier

Sur-apprentissage



Sur-apprentissage

- On dit qu'une règle de classification est en sur-apprentissage si elle "colle trop" aux données d'entraînement.
- Le sur-apprentissage apparaît facilement dans les modèles complexes, quand on cherche à apprendre trop de choses par rapport à la richesse des données disponibles.
- L'inverse du sur-apprentissage est le "sous-apprentissage" (pas une notion aussi claire) : il consiste à utiliser un modèle trop grossier.
- Le sur-apprentissage est plus trompeur, car l'ajustement aux données d'entraînement paraît très bon !
  Erreur d'apprentissage faible, ex : R^2 en ajustement linéaire.
- Beaucoup de méthodes d'apprentissage font intervenir un ou plusieurs paramètres. Suivant la valeur de celui-ci, on peut tomber dans le sur-apprentissage : il faut ajuster les paramètres pour l'éviter.

Equilibre biais-variance

On peut souvent décomposer le risque d'une méthode comme somme de deux termes :

Décomposition biais-variance

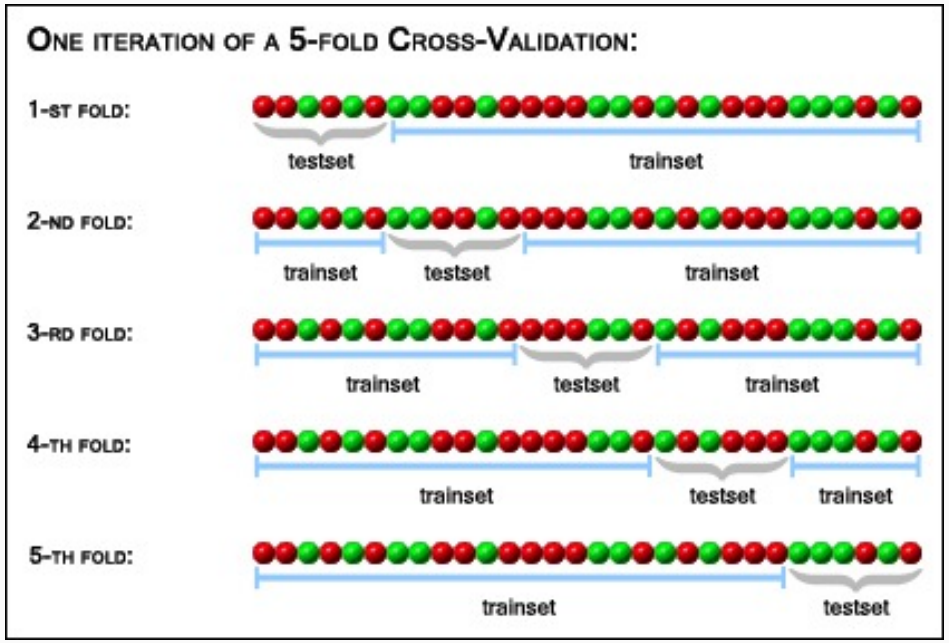
risque $=$ (variabilité $+$) approximation $+$ estimation

- Les deux types d'erreur varient en sens contraire avec les paramètres.
- L'objectif est de trouver un bon *équilibre* entre les deux types d'erreur afin de minimiser leur somme.
- De manière générale il faut préférer les modèles parcimonieux.

Validation croisée

- Pour ajuster les paramètres, il existe
  - des méthodes structurelles (ex : critères AIC, BIC, etc.) ;
  - des méthodes de validation sur les données (ex : échantillon de test, validation croisée).
- Méthode 0 : découpage train/test

  Inconvénient : on apprend sur peu de données
- Solution très souvent adoptée : *validation croisée*

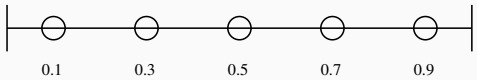


Choix de modèle et fléau de la dimension

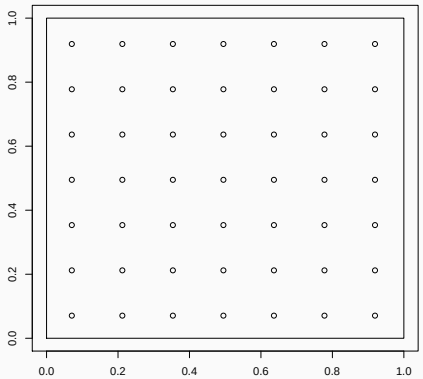
Curse of dimensionality (Bellman)

Dans $[0,1]^d$, combien de points doit contenir une grille régulière pour que tout point ait un voisin de la grille au plus à distance 0.1 ?

$d = 1$: 5 points



$d = 2$: 49 points



$d = 10$: 976562500000 points

$d = 20$: 976562500000000000000000000 points

Modèle plus riche = meilleur ?

- Exemple : régression polynomiale

$$Y_i = f\left(\frac{i}{n}\right) + \epsilon_i \ , \qquad \epsilon_i \sim N^{\circ}\left(0, \sigma^2\right)$$

- f fonction régulière (par exemple polynomiale)
- estimation par régression polynomiale

$\Longrightarrow$ le meilleur modèle n'est pas le plus gros

$\Longrightarrow$ si f est polynomiale, son degré ne donne pas non plus la solution

$\Longrightarrow$ plus le nombre d'observation est grand, plus on peut considérer des modèles riches

- des méthodes non basées sur des modèles donnent aussi d'excellents résultats (ex : k-NN)

Régression logistique

PLAN

- Introduction
- Odds
- Modélisation d'une variable qualitative à 2 modalités
- Modèle binomial
- Choix de modèle

Objectif : Explication d'observations constituées d'effectifs.

Les lois concernées sont discrètes et associées à des dénombrements :

- loi de Poisson
- loi binomiale
- loi multinomiale.

Tous ces modèles appartiennent à la famille du *modèle linéaire général* et partagent à ce titre beaucoup de concepts : famille exponentielle, estimation par maximum de vraisemblance, tests, diagnostics, résidus.

Soit Y une variable qualitative à m modalités. On désigne la chance ou l'odds de voir se réaliser la $j^{\text{ème}}$ modalité plutôt que la $k^{\text{ème}}$ par le rapport

$$\Omega_{jk} = \frac{\pi_j}{\pi_k},$$

où π_j est la probabilité d'apparition de la $j^{\text{ème}}$ modalité.

Cette quantité est estimée par $\frac{n_j}{n_k}$ des effectifs observés sur un échantillon.

Si Y est une variable de Bernoulli de paramètre π, alors l'odds est égal à $\frac{\pi}{1-\pi}$, la cote ou chance de gain.

Soit Y une variable qualitative à 2 modalités : 1 ou 0, succès ou échec, présence ou absence de verglas...

Les modèles de régression linéaire adaptés à l'explication d'une variable quantitative ne s'appliquent plus directement car le régresseur linéaire usuel $\mathbf{X}\beta$ ne prend pas des valeurs simplement binaires.

Objectif : Adapter la modélisation linéaire à cette situation en cherchant à expliquer les probabilités

$$\pi = \mathbb{P}(Y = 1) \text{ ou } 1 - \pi = \mathbb{P}(Y = 0)$$

ou plutôt une transformation de celles-ci, par l'observation conjointe des variables explicatives.

Idée : Faire intervenir une fonction réelle monotone g opérant de $[0; 1]$ dans $\mathbb{R}$ et chercher un modèle linéaire de la forme

$$g(\pi) = \mathbf{x}'\beta.$$

Plusieurs possibilités :

- **probit** : g est la fonction inverse de la fonction de répartition d'une loi normale, mais son expression n'est pas explicite.
- **log-log** : $g(\pi) = \ln[-\ln(1-\pi)]$, mais cette fonction est dissymétrique.
- **logit** : $g(\pi) = \mathit{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$ avec $g^{-1}(x) = \frac{e^x}{1+e^x}$.

La régression logistique s'interprète comme la recherche d'une modélisation linéaire du "log odds".

On va chercher β tel que

$$\frac{\pi_X}{1-\pi_X} \simeq e^{X'\beta}.$$

- La régression logistique s'exprime généralement sous la forme

$$\frac{\pi_{x_1,\cdots,x_p}}{1-\pi_{x_1,\cdots,x_p}} = e^{\alpha_0+\alpha_1 x_1+\cdots+\alpha_p x_p}.$$

- Si $\alpha_j \simeq 0$, alors l'odds ratio ne dépend pas de x_j.
- Un incrément de 1 dans x_j correspond à un incrément de e^{α_j} dans l'odds ratio.
- Si $\alpha_j > 0$, alors un incrément de 1 dans x_j entraîne une augmentation de e^{α_j} du risque
(ou de la chance) que Y prenne la valeur 1.

$$\frac{\frac{\pi_{x_1,\cdots,x_j+1,\cdots,x_p}}{1-\pi_{x_1,\cdots,x_j+1,\cdots,x_p}}}{\frac{\pi_{x_1,\cdots,x_p}}{1-\pi_{x_1,\cdots,x_p}}} = e^{\alpha_j}.$$

- On teste l'effet de chaque variable en testant l'hypothèse

$$H_0 : \beta_i = 0.$$

- On rejette l'hypothèse H_0 dès que la p-valeur de H_0 est trop faible.
  - En effet

$$P(\text{Observations}|H_0) = \text{p-valeur}.$$

    Si la p-valeur est trop faible, alors aux vues de ces observations, l'hypothèse H_0 est peu vraisemblable et par conséquent le coefficient β_i est significativement non nul.
- Les logiciels scientifiques donnent cette p-valeur.
- On peut également déterminer des régions de confiance pour la régression logistique sur π_X grâce à g^{-1}(région de confiance classique).

Plan d'expérience

- Pour $i \in \{1, \cdots, I\}$, on effectue n_i mesures de la variable Y.
- y_i désigne le nombre de réalisations $Y = 1$.
- On suppose que π_i est la probabilité de succès de Y sachant que les $x^1, \cdots, x^p$ appartiennent au goupe i

$$\pi_i = P(Y = 1 | X \in \text{Groupe } i).$$

- On suppose que les groupes sont homogènes dans leurs réalisations de Y : la probabilité pour Y d'être égal à 1 au sein d'un même groupe est indépendante de la valeur de X dans ce groupe.
- Proposition : En effectuant n_i mesures dans le groupe i et en supposant tous les échantillons indépendants, si Y_i désigne le nombre de valeurs $Y = 1$, alors

$$P(Y_i = y_i) = C_{n_i}^{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i}.$$

ESTIMATION DES COEFFICIENTS

- On suppose que le vecteur des fonctions *logit* des probabilités π_i appartient au sous-espace $Vect\{X^1, \ldots, X^p\}$

$$logit(\pi_i) = \mathbf{x}_i'\beta \quad \text{ou} \quad \pi_i = \frac{e^{\mathbf{x}_i'\beta}}{1 + e^{\mathbf{x}_i'\beta}},\ i = 1, \ldots, I.$$

- La log-vraisemblance s'écrit

$$L(\beta) = \Pi_{i=1}^{I} C_n^{y_i} g^{-1}(\mathbf{x}_i'\beta)^{y_i}(1 - g^{-1}(\mathbf{x}_i'\beta))^{n_i - y_i}.$$

- β est estimé par maximisation de la log-vraisemblance. Il n'y a pas de solution analytique, celle-ci est obtenue par des méthodes numériques itératives (algorithme de Newton Raphson).
- L'optimisation fournit une estimation $\mathbf{b}$ de β. On peut alors en déduire les estimations ou prévisions des probabilités π_i et celles des effectifs

$$\widehat{\pi}_i = \frac{e^{\mathbf{x}_i'\mathbf{b}}}{1 + e^{\mathbf{x}_i'\mathbf{b}}} \quad \text{et} \quad \widehat{y}_i = n_i\widehat{\pi}_i.$$

REMARQUES

- La matrice $\mathbf{X}$ issue de la planification expérimentale est construite avec les mêmes règles que celles utilisées dans le cadre de la covariance mixant variables explicatives quantitatives et qualitatives.

- La situation décrite précédemment correspond à l'observation de données groupées. Dans de nombreuses situations concrètes et souvent dès qu'il y a des variables explicatives quantitatives, les observations $\mathbf{x}_i$ sont toutes distinctes. Ceci revient donc à fixer $n_i = 1, i = 1, \ldots, I$ dans les expressions précédentes et la loi de Bernoulli remplace la loi binomiale.

Le test du rapport de vraisemblance et le test de Wald permettent de comparer un modèle avec un sous-modèle et d'évaleur l'intérêt de la présence des termes complémentaires. Stratégie descendante à partir du modèle complet.

Idée : Supprimer, un terme à la fois, la composante d'intéraction ou l'effet principal qui apparaît comme le moins significatif au sens du rapport de vraisemblance ou du test de Wald.

ATTENTION du fait de l'utilisation d'une transformation non linéaire (*logit*), même si des facteurs sont orthogonaux, aucune propriété d'orthogonalité ne peut être prise en compte pour l'étude des hypothèses.

- Élimination des termes un par un et ré-estimation du modèle.
- Un terme prinicipal ne peut être supprimé que s'il n'intervient plus dans les termes d'interaction.

Les logiciels calculent en plus l'AIC pour finaliser le choix pour une meilleure qualité prédictive.

Qualités du classifieur de régression logistique

Interprétabilité : NON

Critique : OUI ! (très utilisé pour le scoring)

Consistance : NON (sauf si le modèle est exact)

Minimax : NON

Parameter-free : OUI

Vitesse : OUI

Online : possible

Analyse discriminante décisionnelle

- Idée : chaque caractéristique est associé à la classe dont le barycentre des caractéristiques dans l'échantillon d'apprentissage est le plus proche

$\Longleftrightarrow$ recherche directe d'une frontière linéaire

- Problème : hétérogénéité des variables
- Nombreuses variantes :

LDA = AFD (analyse factorielle discriminante)

interprétation bayésienne

k-NN est parfois vu comme une variante non-paramétrique

noyau : idem pour la méthode du noyau

Arbres de décision

PLAN

- Introduction
- Construction d'un arbre binaire
  - Principe
  - Critère de division
  - Règle d'arrêt
  - Affectation
- Critères d'homogénéité
  - Y quantitative
  - Y qualitative
- Élagage
  - Construction de la séquence d'arbres
  - Recherche de l'arbre optimal

La segmentation par arbre est une approche non-paramétrique de l'analyse discriminante.

But : expliquer une variable réponse (qualitative ou quantitative) à l'aide d'autres variables.

Principe : construire un arbre à l'aide de divisions successives des individus d'un ensemble E en deux segments (appelés aussi noeuds) homogènes par rapport à une variable Y (binaire, nominale, ordinale ou quantitative) en utilisant l'information de p variables $X^1, \ldots, X^p$ (binaires, nominales, ordinales ou quantitatives).

L'arbre obtenu est sous forme d'un arbre inversé comportant à la racine l'échantillon total E à segmenter et les autres segments sont

- soit des segments intermédiaires (encore divisibles),
- soit des segments terminaux.

L'ensemble des segments terminaux constitue une partition de l'ensemble E en classes homogènes et distinctes, relativement à la variable Y.

Il s'agit d'arbre de classement si Y est qualitative et d'arbre de régression si Y est quantitative.

Avantages / Inconvénients

- La méthode CART (Classification And Regression Tree) fournit des solutions sous formes graphiques simples à interpréter.
- Elle est complémentaire des méthodes statistiques classiques, très calculatoire et efficace à condition d'avoir de grandes tailles d'échantillon.
- Elle est capable de gérer à la fois les variables quantitatives ET qualitatives simultanément.
- Peu d'hypothèses requises !
- Algorithme étant basé sur une stratégie pas à pas hiérarchisée, il peut passer à côté d'un optimum global.

Soient p variables quantitatives ou qualitatives explicatives X^j et une variable à expliquer Y qualitative à m modalités $\{\tau_l, l = 1, \ldots, m\}$ ou quantitative réelle, observée sur un échantillon de n individus.

La construction d'un arbre de discrimination binaire consiste à déterminer une séquence de noeuds.

- Un noeud est défini par le choix conjoint d'une variable parmi les explicatives et d'une division qui induit une partition en deux classes.
- Une division est elle-même définie par une valeur seuil de la variable quantitative sélectionnée ou un partage en deux groupes des modalités si la variable est qualitative.
- À la racine ou au noeud initial correspond l'ensemble de l'échantillon. La procédure est ensuite itérée sur chacun des sous-ensembles.

L'algorithme considéré nécessite

1. la définition d'un critère permettant de sélectionner la "meilleure" division parmi toutes celles admissibles pour les différentes variables ;
2. une règle permettant de décider qu'un noeud est terminal : il devient alors feuille ;
3. l'affectation de chaque feuille à l'une des classes ou à une valeur de la variable à expliquer.

Le point 2. correspond encore à la recherche d'un modèle parcimonieux. Un arbre trop détaillé, associé à une sur-paramétrisation, est instable et donc probablement plus défaillant pour la prévision d'autres observations.

Breiman et al. ont mise en place une stratégie de recherche de l'arbre optimal.

1. Construire l'arbre maximal A_{max}.
2. Ordonner les sous-arbres selon une séquence emboîtée suivant la décroissance d'un critère pénalisé de déviance ou de taux de mal-classés.
3. Sélectionner le sous-arbre optimal : c'est la procédure d'élagage.

CRITÈRE DE DIVISION

Une division est dite admissible si aucun des segments descendants n'est vide.

- Si la variable explicative est qualitative ordinale à m modalités, elle conduit à $m - 1$ divisions binaires admissibles.
- Si elle est nominale, le nombre de divisions devient égal à $2^{m-1} - 1$.
- Pour une variable quantitative à m valeurs disctinctes, on se ramène au cas ordinal.

Le critère de division repose sur la définition d'une fonction d'hétérogénéité ou de désordre.

Objectif : Partager les individus en deux groupes les plus homogènes au sens de la variable à expliquer.

CRITÈRE DE DIVISION

L'hétérogénéité d'un noeud se mesure par une fonction non négative qui doit être

1. nulle si et seulement si le segment est homogène : tous les individus appartiennent à la même modalité ou prennent la même valeur de Y,
2. maximale lorsque les valeurs de Y sont équiprobables ou très dispersées.

La division du noeud k crèe deux fils notés $(k+1)$ et $(k+2)$, mais une renumérotation sera nécessaire pour respecter la séquence de sous-arbres.

Parmi toutes les divisions admissibles du noeud k, l'algorithme retient celle qui rend la somme $D_{(k+1)} + D_{(k+2)}$ des désordres des noeuds fils minimale, c-à-d

$$\max_{\{\text{divisions de } X^j\,;\, j=1,\ldots,p\}} D_k - D_{(k+1)} - D_{(k+2)}.$$

RÈGLE D'ARRÊT

La croissance de l'arbre s'arrête à un noeud qui devient donc feuille

- lorsqu'il est homogène, c-à-d lorsqu'il n'existe plus de division admissible,
- si le nombre d'observations qu'il contient est inférieur à un seuil fixé par l'utilisateur d_{min}. En général, $1 \leq d_{min} \leq 5$.
- si le nombre de noeuds est supérieur à n_{max}, nombre fixé par l'utilisateur.

AFFECTATION

Une fois les critères d'arrêt atteints, il faut affecter une valeur à chaque feuille.

- Si Y est quantitative, attribution de la valeur moyenne aux observations de cette feuille.
- Si Y est qualitative, chaque feuille est affectée à une modalité τ_l de Y en considérant le mode conditionnel
  - celle la plus représentée dans la feuille, c-à-d celle ayant la proportion la plus élevée à l'intérieur de cette feuille. Il est alors facile de comparer le nombre de données mal classées.
  - la modalité la moins coûteuse si des coûts de mauvais classements sont donnés.
  - la classe *a posteriori* la plus probable au sens bayésien si des probabilités *a priori* sont connues.

Soit une partition de n individus en deux sous-populations E_1 et E_2 de tailles respectives n_1 et n_2. Soit μ_{ij} la valeur "théorique" de Y pour l'individu i du sous-ensemble E_j.

L'hétérogénéité du sous-ensemble E_j est mesurée par

$$D_j = \sum_{i=1}^{n_j} (\mu_{ij} - \mu_{.j})^2 \text{ avec } \mu_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} \mu_{ij}.$$

Alors l'hétérogénéité de la partition est définie par

$$D = D_1 + D_2 = \sum_{j=1}^{2} \sum_{i=1}^{n_j} (\mu_{ij} - \mu_{.j})^2.$$

C'est l'inertie intragroupe qui vaut 0 si et seulement si $\forall i, j, \mu_{ij} = \mu_{.j}$.

La différence entre l'hétérogénéité de l'ensemble non partagé et celle de la partition est

$$\begin{aligned}\Delta &= \sum_{j=1}^{2}\sum_{i=1}^{n_j}(\mu_{ij}-\mu_{..})^2 - \sum_{j=1}^{2}\sum_{i=1}^{n_j}(\mu_{ij}-\mu_{.j})^2 \text{ où } \mu_{..} = \frac{1}{n}\sum_{i,j}\mu_{ij}.\\ &= \sum_{j=1}^{2} n_j(\mu_{..}-\mu_{.j})^2\\ &= \frac{n_1 n_2}{n}(\mu_{.1}-\mu_{.2})^2.\end{aligned}$$

Δ correspond au "désordre" des barycentres et est homogène à la variance intergroupe.
Objectif : À chaque étape, maximiser Δ, c-à-d trouver la variable explicative induisant une partition en deux sous-ensembles associée à une inertie intragroupe minimale ou encore qui rende l'inertie intergroupe maximale.

Les quantités sont estimées

$$D_j \text{ par } \widehat{D}_j = \sum_{i=1}^{n_j} (y_{ij} - y_{.j})^2 \quad \text{et} \quad D \text{ par } \widehat{D} = \sum_{j=1}^{2} \sum_{i=1}^{n_j} (y_{ij} - y_{.j})^2.$$

Soit Y une variable à expliquer à m modalités τ_l. L'arbre induit une partition pour laquelle n_{+k} désigne l'effectif du $k^{\text{ème}}$ noeud.

Soit $p_{lk} = \mathbb{P}[\tau_l|k]$ avec $\sum_{l=1}^{m} p_{lk} = 1$, la probabilité qu'un élément du $k^{\text{ème}}$ noeud appartienne à la $l^{\text{ème}}$ classe.

Le désordre du $k^{\text{ème}}$ noeud, défini à partir de l'entropie, s'écrit

$$D_k = -2 \sum_{l=1}^{m} n_{+k} p_{lk} \log(p_{lk}).$$

L'hétérogénéité de la partition est encore

$$D = \sum_{k=1}^{2} D_k = -2 \sum_{k=1}^{2} \sum_{l=1}^{m} n_{+k} p_{lk} \log(p_{lk}).$$

Cette quantité est positive et nulle si et seulement si les probabilités p_{lk} ne prennent que des valeurs 0 sauf une égale à 1 correspondant à l'absence de mélange.

ENTROPIE

Soit n_{lk} l'effectif observé de la $l^{\text{ème}}$ classe dans le $k^{\text{ème}}$ noeud :

$$n_{+k} = \sum_{l=1}^{m} n_{lk}.$$

Les quantités sont estimées

$$D_k \text{ par } \widehat{D}_k = -2 \sum_{l=1}^{m} n_{+k} \frac{n_{lk}}{n_{+k}} \log \left(\frac{n_{lk}}{n_{+k}} \right)$$

et

$$D \text{ par } \widehat{D} = -2 \sum_{k=1}^{2} \sum_{l=1}^{m} n_{+k} \frac{n_{lk}}{n_{+k}} \log \left(\frac{n_{lk}}{n_{+k}} \right).$$

Critère de Gini

Le critère de Gini du noeud k est défini par $D_k = \sum_{l \neq h} p_{lk} p_{hk}$ avec $l, h = 1, \ldots, m$ et est estimé par $\widehat{D}_k = \sum_{l \neq h} \frac{n_{lk}}{n_{+k}} \frac{n_{hk}}{n_{+k}}$.

- Le désordre D_k est maximal si $p_{lk} = \frac{1}{m}$; l'échantillon présente autant d'éléments de chaque modalité.
- D_k est nul si l'échantillon est pur : $p_{lk} = 1$ et $p_{hk} = 0$ si $h \neq l$.
- D_k représente la probabilité de mauvais classement pour un individu tiré au hasard parmi les individus du noeud k.

Critère de Gini

Le désordre de l'échantillon initial de taille n est estimé par $\widehat{D} = \sum_{l \neq h} \frac{n_l}{n} \frac{n_h}{n}$, où n_l représente l'effectif observé de la $l^{\text{ème}}$ modalité dans l'échantillon initial.

La réduction d'impureté correspond à une division binaire est alors estimée par

$$\widehat{\Delta} = \widehat{D} - \frac{n_{+1}}{n} \widehat{D}_1 - \frac{n_{+2}}{n} \widehat{D}_2.$$

Objectif : Rechercher le meilleur compromis entre

- un arbre très détaillé, fortement dépendant des observations qui ont permis son estimation, qui fournira un modèle de prévision très instable
- un arbre trop robuste mais grossier qui donne des prédictions trop approximatives.

Principe

- Construire une suite emboîtée de sous-arbres de l'arbre maximum par élagage successif.
- Choisir, parmi cette suite, l'arbre optimal au sens d'un critère.

La solution obtenue par algorithme pas à pas n'est pas nécessairement, globalement optimale mais l'efficacité et la fiabilité sont préférées à l'optimalité.

CONSTRUCTION DE LA SÉQUENCE D'ARBRES

Pour un arbre A donné, on note K le nombre de feuilles ou noeuds terminaux de A ; la valeur de K exprime la complexité de A.

La qualité de discrimination d'un arbre A se mesure par le critère $D(A) = \sum_{k=1}^{K} D_k(A)$ où $D_k(A)$ est le nombre de mal classés ou la déviance ou le coût de mauvais classement de la $k^{\text{ème}}$ feuille de l'arbre A.

CONSTRUCTION DE LA SÉQUENCE D'ARBRES

La construction de la séquence d'arbres emboîtés repose sur une pénalisation de la complexité de l'arbre

$$C(A) = D(A) + \gamma K.$$

- Pour $\gamma = 0$, $A_{max} = A_K$ minimise $C(A)$.
- En faisant croître γ, l'une des divisions de A_K, celle pour laquelle l'amélioration de D est la plus faible (inférieure à γ) apparaît comme superflue et les deux feuilles sont regroupées (élaguées) dans le noeud père qui devient terminal ; $A_K \supset A_{K-1}$.

Construction de la séquence d'arbres

Soit $\mathcal{N}$ un noeud. On appelle $A_{\mathcal{N}}$ le sous-arbre (ou la branche) de A extrait(e) à partir de $\mathcal{N}$, donc constitué des descendants de $\mathcal{N}$ et de la racine $\mathcal{N}$. On appelle A' le sous-arbre de A auquel on a enlevé la branche $A_{\mathcal{N}}$. On a alors

$$C(A') = C(A) + C(\mathcal{N}) - C(A_{\mathcal{N}}).$$

Par conséquent,

$$C(A') \geq C(A) \Longleftrightarrow \gamma \leq \frac{D(\mathcal{N}) - D(A_{\mathcal{N}})}{|A_{\mathcal{N}}| - 1} = \alpha.$$

Ceci signifie que si la valeur de γ fixée est inférieure à α, le coût du sous-arbre élagué A' est supérieur à celui de A : on gardera donc l'arbre complet A.

CONSTRUCTION DE LA SÉQUENCE D'ARBRES

Le procédé est itéré pour la construction de la séquence emboîtée

$$A_{max} = A_K \supset A_{K-1} \supset \ldots \supset A_1$$

où A_1, le noeud racine, regroupe l'ensemble de l'échantillon.

Un graphe représente la décroissance ou l'éboulis de la déviance (ou du taux de mal classé) en fonction du nombre croissant de feuilles dans l'arbre ou en fonction de la valeur décroissante du coefficient de pénalisation γ (graphes équivalents).

Élagage lorsque l'augmentation de la complexité de l'arbre n'est plus compensée par la diminution de la déviance.

RECHERCHE DE L'ARBRE OPTIMAL

Les procédures d'élagage diffèrent par la façon d'estimer l'erreur de prédiction. Quand l'amélioration du critère est jugée trop petite ou négligeable, on élague l'arbre au nombre de feuilles obtenues.

- L'évaluation de la déviance ou du taux de mauvais classement estimée par resubsitution sur l'échantillon d'apprentissage est biaisée (trop optimiste).

RECHERCHE DE L'ARBRE OPTIMAL

- Une estimation sans biais est obtenue par l'utilisation d'un autre échantillon (validation) ou encore par validation croisée.

  La procédure de validation croisée a une particularité : la séquence d'arbres obtenue est différente pour chaque estimation sur l'un des sous-échantillons.

  - L'erreur moyenne n'est pas calculée pour chaque sous-arbre avec un nombre de feuilles donné mais pour chaque sous-arbre correspondant à une valeur fixée du coefficient de pénalisation γ.
  - À la valeur de γ minimisant l'estimation de l'erreur de prévision, correspond ensuite l'arbre jugé optimal dans la séquence estimée sur tout l'échantillon d'apprentissage.

Recherche de l'arbre optimal

Le principe de sélection d'un arbre optimal est donc décrit par l'algorithme suivant

- Construction de l'arbre maximal A_{max}.
- Construction de la séquence $A_K, \ldots, A_1$ d'arbres emboîtés.
- Estimation sans biais (échantillon de validation ou validation croisée) des déviances $D(A_K), \ldots, D(A_1)$.
- Représentation de $D(A_k)$ en fonction de k ou de γ.
- Choix de k rendant $D(A_k)$ minimum.

Qualités du classifieur CART

Interprétabilité : OUI !

Critique : OUI mais pas très précis

Consistance : OUI (sous certaines réserves) MAIS instable !

Minimax : NON !

Parameter-free : NON

Vitesse : OUI

Online : NON