DATAMINING SUPERVISE.

II Modèles paramétriques.

1) Modèles gaussiens: discriminations quadratique, linéaire

Nous allons généraliser l'exercice précédent au cas où les deux "nuages" de l'ensemble d'apprentissage sont des n_i-échantillons gaussiens dans R^p .

a-Rappels sur les vecteurs gaussiens.
Une variable $\mathbf{X} = (X_1, ..., X_p)$ dans R^p est un vecteur gaussien si $< \mathbf{l}, \mathbf{X} >= \Sigma_{j=1}^p l_j X_j$ est une v.a. normale pour toute forme linéaire l (la loi normale peut être dégénérée en une mesure de dirac δ_m).

En particulier, les X_i sont des gaussiennes $N(m_i, \sigma_i^2)$ (mais ce n'est pas suffisant).
Exemple 1: $X_1, ..., X_p$ gaussiennes $N(0, 1)$ mutuellement indépendantes;
Exemple 2: $\mathbf{Y} = \mathbf{XA} + \mathbf{m}$, $\mathbf{X}$ gaussien de l'exemple 1, $\mathbf{A}$ $p \times q$ application linéaire de R^p dans R^q ("à droite") et $\mathbf{m} =\in R^q$.

La fonction caractéristique de $< \mathbf{l}, \mathbf{X} >$ vaut
$\varphi_{<\mathbf{l},\mathbf{X}>}(t) = exp[itE(< \mathbf{l}, \mathbf{X} >) - \frac{t^2}{2} Var(< \mathbf{l}, \mathbf{X} >)]$
$= exp[it < \mathbf{l}, E(\mathbf{X}) > -\frac{t^2}{2} \mathbf{l}\mathbf{Var}\mathbf{X}^t\mathbf{l}]$
où $E(\mathbf{X}) = \mathbf{m}$ et $\mathbf{VarX} = (Cov(X_i, X_j))$.

D'où, avec $t = 1$, la fonction caractéristique de $\mathbf{X}$:
$\varphi_{\mathbf{X}}(\mathbf{l}) = exp[i < \mathbf{l}, E(\mathbf{X}) > -\frac{1}{2} \mathbf{l}\mathbf{Var}\mathbf{X}^t\mathbf{l}]$.

Théorème: soient $\mathbf{X} = (\mathbf{X}_{(1)}, \mathbf{X}_{(2)})$ et $\mathbf{Var X} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{pmatrix}$;
alors $\mathbf{V}_{12} = 0 <=> \mathbf{X}_{(1)}$ et $\mathbf{X}_{(2)}$ indépendantes.

$< \mathbf{l}, E(\mathbf{X}) > = < \mathbf{l}_{(1)}, E(\mathbf{X}_{(1)}) + < \mathbf{l}_{(2)}, E(\mathbf{X}_{(2)} >$,
$V_{12} = 0 => \mathbf{l Var X}\, {}^t\mathbf{l} = \mathbf{l}_{(1)}\mathbf{V}_{11}\mathbf{l}_{(1)} + \mathbf{l}_{(2)}\mathbf{V}_{22}\, {}^t\mathbf{l}_{(2)}$,
la fonction caractéristique se factorise: $\varphi_{\mathbf{X}}(\mathbf{l}) = \varphi_{\mathbf{X}_{(1)}}(\mathbf{l}_{(1)}).\varphi_{\mathbf{X}_{(2)}}(\mathbf{l}_{(2)})$
d'où l'indépendance (la réciproque est triviale).

Corollaire 1: tout vecteur gaussien $\mathbf{Y}$ est un exemple 2.
soit $\mathbf{O}$ orthogonale qui diagonalise $\mathbf{V} = \mathbf{Var Y}$, $\mathbf{V}$ de rang p:
${}^t\mathbf{OVO} = \Lambda = \begin{pmatrix} \Lambda_{(1)} & 0_{q,q-p} \\ 0_{q-p,q} & 0_{q-p,q-p} \end{pmatrix}$, $\Lambda_{(1)}$ $p \times p$ diagonale **positive** inversible, de racine carrée inversible $\Lambda_{(1)}^{1/2}$ d'inverse noté $\Lambda_{(1)}^{-1/2}$.
Choisir $\mathbf{m} = E(\mathbf{Y})$, $\mathbf{X} = (\mathbf{Y} - \mathbf{m})\mathbf{O} \begin{pmatrix} \Lambda_{(1)}^{-1/2} \\ 0 \end{pmatrix}$ et $\mathbf{A} = \begin{pmatrix} \Lambda_{(1)}^{1/2} & 0 \end{pmatrix} {}^t\mathbf{O}$.

D'après le corollaire 1, le support de $\mathbf{Y}$ appartient à $Im\mathbf{A} + \mathbf{m}$; $\mathbf{Y}$ ne peut admettre une densité (pour la mesure de Lebesgue de R^q) que si p=q, soit $\mathbf{V}$ inversible et donc $\mathbf{A}$ est une racine carrée de $\mathbf{V}$ inversible.

Corollaire 2: soit $\mathbf{V}$ inversible. $\mathbf{Y}$ admet la densité:
$(2\pi)^{-p/2}(det\mathbf{V})^{-1/2}\, exp - \frac{1}{2}(\mathbf{y} - \mathbf{m})\mathbf{V}^{-1}\, {}^t(\mathbf{y} - \mathbf{m})$.

La densité de $\mathbf{X}$ vaut $\Pi_{j=1}^{p} \frac{1}{\sqrt{2\pi}} exp - \frac{1}{2}x_j^2 = (2\pi)^{-p/2} exp - \frac{1}{2}\Sigma_{j=1}^{p} x_j^2$;
la transformation linéaire $\mathbf{x} = (\mathbf{y} - \mathbf{m})\mathbf{A}^{-1}$admet le jacobien $det(\mathbf{A}^{-1}) = (det\mathbf{A})^{-1} = (det\mathbf{V})^{-1/2}$.

b- Le n-échantillon gaussien dans R^p.

Soit $\{\mathbf{X_i} = (X_{i1}, ..., X_{ip}), i = 1...n\}$ un n-échantillon de loi $\mathcal{N}_p(\mathbf{m}, \mathbf{V})$.
$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ ... \\ \mathbf{X}_n \end{pmatrix} = \left(\mathbf{X}_{ij}\right)$, matrice nxp, est un vecteur gaussien dans R^{np} (appliquer la définition) d'espérance $1_n.\mathbf{m}$ et de variance npxnp définie par $Cov(X_{ij}, X_{lk}) = \mathrm{v_{jk}}$ si $i = l$ et 0 sinon.

Soit $\mathbf{O}$ nxn orthogonale; $\mathbf{Z} = \mathbf{OX}$ est gaussien d'espérance $\mathbf{O}1_n\mathbf{m}$.
Soit $\mathbf{U}_j$ (resp. $\mathbf{T}_j = \mathbf{OU}_j$) la j^{eme} colonne de $\mathbf{X}$ (resp. $\mathbf{Z}$).
$\mathbf{Cov}(\mathbf{T}_j, \mathbf{T}_k) = \mathbf{Cov}(\mathbf{OU}_j, \mathbf{OU}_k) = \mathbf{O}v_{jk}\mathbf{I}_n{}^t\mathbf{O} = v_{jk}\mathbf{I}_n$ montre que $\mathbf{Z}$ a même structure de variance que $\mathbf{X}$: lignes indépendantes de variance $\mathbf{V}$.

Choisissons $\mathbf{O}$ de première ligne $\frac{1}{\sqrt{n}}(1, ...1)$; alors $E(\mathbf{Z}) = \begin{pmatrix} \sqrt{n}\mathbf{m} \\ 0 \\ 0 \end{pmatrix}$;
$\mathbf{Z}_1 \sim \mathcal{N}_p(\sqrt{n}\mathbf{m}, \mathbf{V})$, $\mathbf{Z}_i \sim \mathcal{N}_p(0, \mathbf{V})$, $i = 2...n$, et sont mutuellement indépendantes. Notons:
$\overline{\mathbf{X}} = \frac{1}{n}\Sigma_{i=1}^n \mathbf{X}_i = \frac{1}{\sqrt{n}}\mathbf{Z}_1$,
$\mathbf{S} = \Sigma_{i=2}^n {}^t\mathbf{Z}_i\mathbf{Z}_i = \Sigma_{i=1}^n {}^t\mathbf{X}_i\mathbf{X}_i - \mathbf{Z}_1\mathbf{Z}_1 = \Sigma_{i=1}^n {}^t(\mathbf{X}_i - \overline{\mathbf{X}})(\mathbf{X}_i - \overline{\mathbf{X}})$.
Alors:

Théorème (Fisher):
$\overline{\mathbf{X}} =\sim \mathcal{N}_p(\mathbf{m}, \frac{1}{n}\mathbf{V})$ est un estimateur sans biais convergent de $\mathbf{m}$;
$\frac{1}{n-1}\mathbf{S}$ est un estimateur sans biais convergent de $\mathbf{V}$;
$\overline{\mathbf{X}}$ et $\mathbf{S}$ sont indépendantes.

$\mathbf{S}$ est un vecteur dans $R^{p(p+1)/2}$ de loi de **Wishart** $\mathcal{W}_p(n-1, \mathbf{V})$ (généralisation de la loi du χ^2) à $n-1$ degrés de liberté.

c-Prédiction dans le cas $\mu_0 \sim \mathcal{N}_p(\mathbf{m}_0, \mathbf{V}_0)$, $\mu_1 \sim \mathcal{N}_p(\mathbf{m}_1, \mathbf{V}_1)$.

On peut supposer que μ_0 et μ_1 sont à densité pour la mesure de Lebesgue sur R^p. Sinon, les supports des lois sont connus p.s. (aussitôt que l'ensemble d'apprentissage contient plus de p éléments pour chaque classe); si ces supports sont égaux, on est ramené au cas précédent avec $q < p$; s'ils sont distincts, ils sont disjoints p.s. et le prédicteur est déterministe.

Le prédicteur de Bayes choisit la classe 1 si $\frac{\mu_1(\mathbf{x})}{\mu_0(\mathbf{x})} > \frac{P(Y=0)}{P(Y=1)}$, soit $exp - \frac{1}{2}[(\mathbf{x}-\mathbf{m}_1)\mathbf{V}_1^{-1}\,{}^t(\mathbf{x}-\mathbf{m}_1) - (\mathbf{x}-\mathbf{m}_0)\mathbf{V}_0^{-1}\,{}^t(\mathbf{x}-\mathbf{m}_0)] > K$, $(\mathbf{x}-\mathbf{m}_1)\mathbf{V}_1^{-1}\,{}^t(\mathbf{x}-\mathbf{m}_1) - (\mathbf{x}-\mathbf{m}_0)\mathbf{V}_0^{-1}\,{}^t(\mathbf{x}-\mathbf{m}_0) < K_1$, ou:

$\mathbf{x}(\mathbf{V}_1^{-1} - \mathbf{V}_0^{-1})\,{}^t\mathbf{x} - 2\mathbf{x}(\mathbf{V}_1^{-1}\,{}^t\mathbf{m}_1 - \mathbf{V}_0^{-1}\,{}^t\mathbf{m}_0) < K_2$
avec $K_2 = -2.ln(\frac{P(Y=0).(det\mathbf{V}_1)^{1/2}}{P(Y=1).(det\mathbf{V}_0)^{1/2}}) - \mathbf{m}_1\mathbf{V}_1^{-1}\,{}^t\mathbf{m}_1 + \mathbf{m}_0\mathbf{V}_0^{-1}\,{}^t\mathbf{m}_0$
(fonction de discrimination **quadratique**).

Pour $\mathbf{V}_0 = \mathbf{V}_1 = \mathbf{V}$, on obtient la fonction **linéaire discriminante**
(choix de 1 si >0): $(\mathbf{x} - \frac{\mathbf{m}_1+\mathbf{m}_0}{2})\mathbf{V}^{-1}\,{}^t(\mathbf{m}_1 - \mathbf{m}_0) - ln(\frac{P(Y=0)}{P(Y=1)})$.

En pratique, les paramètres $P(Y=1)$, $\mathbf{m}_0$, $\mathbf{m}_1$, $\mathbf{V}_0$, $\mathbf{V}_1$ sont inconnus. On les estimera sur l'ensemble d'apprentissage $\mathcal{A}_n$ par $\frac{n_1}{n_0+n_1}$, $\overline{\mathbf{X}}_0$, $\overline{\mathbf{X}}_1$, $\widetilde{\mathbf{V}}_0 = \frac{1}{n_0-1}\mathbf{S}_0$, $\widetilde{\mathbf{V}}_1 = \frac{1}{n_1-1}\mathbf{S}_1$.

Dans le cas $\mathbf{V}_0 = \mathbf{V}_1 = \mathbf{V}$ ("MANOVA à un facteur à 2 niveaux"), on estimera $\mathbf{V}$ par $\widetilde{V} = \frac{1}{n_1+n_2-2}(\mathbf{S}_0+\mathbf{S}_1)$; on obtiendra la fonction de discrimination linéaire: $(\mathbf{x} - \frac{\overline{\mathbf{X}}_1+\overline{\mathbf{X}}_0}{2})\widetilde{\mathbf{V}}^{-1}\,{}^t(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_0) - ln(\frac{n_0}{n_1})$.

Exercice: dans le cas linéaire et $P(Y=1) = \frac{1}{2}$, calculer l'erreur du prédicteur de Bayes en fonction de $\Theta = \mathbf{V}^{-1/2}\,{}^t(\mathbf{m}_1 - \mathbf{m}_0)$.

2) Modèles à variables qualitatives.

On suppose maintenant que $\mathbf{X} = (X_1, ..., X_p)$ prend $K = \Pi_{j=1}^{p} k_j$ valeurs $\mathbf{m} = (m_1, ..., m_p) \in \Pi_{j=1}^{p}\{1, ..., k_j\}$; X_j est le j^{eme} **facteur** (aléatoire) à k_j niveaux.
$(Y, \mathbf{X})$ admet une loi multinômiale $\mathcal{M}_{2K}(1, \mathbf{P})$ de paramètre $\mathbf{P} = (p_{0,(1,...,1)}, ..., p_{1,(k_1,...k_p)}) \in \mathcal{S}_{2K-1}$ simplexe de R^{2K} .
$\mu_0(\mathbf{m}) = \frac{p_{0,\mathbf{m}}}{\Sigma_{\mathbf{m}} p_{0,\mathbf{m}}}$, $\mu_1(\mathbf{m}) = \frac{p_{1,\mathbf{m}}}{\Sigma_{\mathbf{m}} p_{1,\mathbf{m}}}$, $P(Y = 1|\mathbf{X} = \mathbf{m}) = \frac{p_{1,\mathbf{m}}}{p_{0,\mathbf{m}} + p_{1,\mathbf{m}}}$, le prédicteur de Bayes choisit 1 si $p_{1,\mathbf{m}} > p_{0,\mathbf{m}}$.

L'information de l'ensemble d'apprentissage $\mathcal{A}_n$ est résumée par la table de contingence $(1 + p)$-dimensionnelle composée de 2 strates $\mathbf{N}_0$, $\mathbf{N}_1$ (classes 0, 1)comportant chacune K cellules: $N_{0,\mathbf{m}}$ et $N_{1,\mathbf{m}}$ sont les nombres d'évènements $\{\mathbf{X}_i = \mathbf{m}\}$ observés dans les classes 0 et 1, $i = 1, ..., n$.

$(\mathbf{N}_0, \mathbf{N}_1)$ suit une loi multinômiale $\mathcal{M}_{2K}(n, \mathbf{P})$. On estimera $p_{0,\mathbf{m}}$ et $p_{1,m}$ par $\widehat{p}_{0,\mathbf{m}} = \frac{N_{0,\mathbf{m}}}{\Sigma_{\mathbf{m}} N_{0,\mathbf{m}}}$ et $\widehat{p}_{1,\mathbf{m}} = \frac{N_{1,\mathbf{m}}}{\Sigma_{\mathbf{m}} N_{1,\mathbf{m}}}$.

En pratique, ceci est impossible à réaliser aussitôt que le nombre de facteurs est grand: avec $p = 10$ et $k_j = 2$ pour tous j , $K = 1024$ cellules pour chaque classe; $N_{i,\mathbf{m}} \neq 0$ $\forall \mathbf{m}$ exige n très grand; sinon, il y a surparamétrisation; on est amené à diminuer considérablement le nombre de paramètres.

Par exemple, le **prédicteur bayésien naïf** suppose que tous les facteurs X_j sont indépendants (conditionnellement à Y), de lois respectives μ_{0j} et $\mu_{1,j}$. Alors $\mu_0(\mathbf{X}) = \Pi_{j=1}^{p} \mu_{0,j}(X_j)$ et $\mu_1(\mathbf{X}) = \Pi_{j=1}^{p} \mu_{1,j}(X_j)$, et la dimension du modèle devient $2k - 2p$, avec $k = \Sigma_{j=1}^{p} k_j$.
On estimera $\mu_{0,j}(m_j)$ et $\mu_{1,j}(m_j)$ par $\frac{\Sigma_{\mathbf{m}, m_j\, fixe} N_{0,\mathbf{m}}}{\Sigma_{\mathbf{m}} N_{0,\mathbf{m}}}$ et $\frac{\Sigma_{\mathbf{m}, m_j\, fixe} N_{1,\mathbf{m}}}{\Sigma_{\mathbf{m}} N_{1,\mathbf{m}}}$.

3) Modèles logistiques.

$p(\mathbf{X}) = P(Y=1|\mathbf{X}) = \frac{\mu_1(\mathbf{X}).P(Y=1)}{\mu_1(\mathbf{X}).P(Y=1)+\mu_0(\mathbf{X}).P(Y=0)}$ d'où $\frac{p(\mathbf{X})}{1-p(\mathbf{X})} = \frac{\mu_1(\mathbf{X}).P(Y=1)}{\mu_0(\mathbf{X}).P(Y=0)}$, soit $\ln(\frac{p(\mathbf{X})}{1-p(\mathbf{X})}) = ln(\frac{\mu_1(\mathbf{X})}{\mu_0(\mathbf{X})}) + ln(\frac{P(Y=1)}{P(Y=0)})$.

Dans le cas gaussien, nous avons obtenu une fonction linéaire $ln(\frac{p(\mathbf{X})}{1-p(\mathbf{X})}) = a + \mathbf{X}\mathbf{V}^{-1}({}^t\mathbf{m}_1 - {}^t\mathbf{m}_0)$ dans le cas $\mathbf{V}_0 = \mathbf{V}_1 = \mathbf{V}$, et quadratique sinon.
Dans le cas des modèles "exponentiels" (Binomiale, Poisson,...), la fonction $ln(\frac{p(\mathbf{X})}{1-p(\mathbf{X})})$ est également linéaire en $\mathbf{X}$.

De façon générale, on notera $\varphi_\theta(\mathbf{X}) = ln(\frac{p(\mathbf{X})}{1-p(\mathbf{X})})$, où θ décrit les paramètres du modèle; $u \to ln(\frac{u}{1-u})$ est la fonction **logit,** d'inverse la fonction **logistique** $t \to \frac{exp\,t}{1+exp\,t}$.

Le prédicteur de Bayes s'écrit $1_{\varphi_\theta(\mathbf{X})>0}$.

Conditionnellement en $\mathbf{X}$, la loi de Y est une Bernoulli $\mathcal{B}(p(\mathbf{X}))$ de densité $exp[y.ln(p(\mathbf{X})) + (1-y).ln(1-p(\mathbf{X}))] =$

$exp[y.\varphi_\theta(\mathbf{X}) - ln(1 + exp(\varphi_\theta(\mathbf{X}))]$ (par rapport à $\delta_0 + \delta_1$) .

On estime le paramètre θ par maximum de vraisemblance du **modèle conditionnel en les** $\mathbf{X}_i$ sur l'ensemble d'apprentissage $\mathcal{A}_n$, c'est-à-dire en maximisant, par une méthode numérique (comme le gradient conjugué...) la log-vraisemblance

$\Sigma_{i=1}^n[Y_i.\varphi_\theta(\mathbf{X}_i) - ln(1 + exp(\varphi_\theta(\mathbf{X}_i))]$.

La maximisation de la vraisemblance conditionnelle permet d'ignorer la loi marginale de $\mathbf{X}$: c'est une méthode **semi-paramétrique**.
Un tel modèle est appelé **modèle logistique**.

Dans le cas gaussien, les fonctions de discrimination linéaire ou quadratique ont été précédemment estimées par maximisation de la vraisemblance complète; les estimateurs obtenus ont une précision meilleure (sans biais et de variance plus petite), mais exigent l'hypothèse de normalité. Sans hypothèse sur la loi marginale de $\mathbf{X}$, le modèle logistique présente une meilleure **robustesse**.

D'autre part, cette absence d'hypothèse sur la loi de $\mathbf{X}$ permet de traiter conjointement variables **quantitatives** et **qualitatives**. Un modèle logistique pourra être décrit par une formule formellement identique à celle d'un modèle GLM ou d'un modèle non-linéaire; on pourra définir des familles paramétriques emboîtées...

Exemple 1:
le modèle bayésien naïf est de dimension $2k - 2p$ (paramétrage $a_{0,jh} = P(X_j = h|Y = 0)$, $a_{1,jh} = P(X_j = h|Y = 1)$,$h = 1...k_j$, $j = 1...p$ avec les contraintes $\Sigma_{h=1}^{k_j} a_{0,jh} = \Sigma_{h=1}^{k_j} a_{1,jh} = 1$, $\forall j$) .

Dans le modèle conditionnel , $\varphi_\theta(\mathbf{X}) = \Sigma_{j=1}^{p} \Sigma_{h=1}^{k_j} \theta_{jh} 1_{X_{j=h}}$, avec $\theta = (\theta_{jh} = ln(\frac{a_{1jh}}{a_{0,jh}})$, $h = 1...k_j$, $j = 1...p)$ de dimension $k - p + 1$; c'est un modèle **additif**.

Exemple 2: soit $\varphi_\theta(\mathbf{X}) = \Sigma_{j=1}^{p} \alpha_j 1_{X_j > u_j}$, de paramètre $\theta = (\alpha_1, ..., \alpha_p, u_1, ..., u_p)$: on recherche le meilleur prédicteur (par choix des α_j) parmi les sommes pondérée des meilleurs prédicteurs "élémentaires" $1_{X_j > u_j}$(par choix de u_j) (X_j quantitatifs ou qualitatifs). C'est une procédure de **boosting**.

Exercice:
l'ensemble de **Highleyman** comporte deux classes équiprobables :
conditionnellement à $Y = 0$, $\mathbf{X}$ est de loi $\mathcal{N}_2((1,1), \begin{pmatrix} 1 & 0 \\ 0 & .25 \end{pmatrix})$;
conditionnellement à $Y = 1$, $\mathbf{X}$ est de loi $\mathcal{N}_2((2,0), \begin{pmatrix} .01 & 0 \\ 0 & 4 \end{pmatrix})$.

1°) a) Calculer le prédicteur de Bayes; exprimer son erreur de prédiction; proposer des méthodes de calcul numérique.
b) Calculer la fonction discriminante linéaire; calculer son erreur de prédiction à l'aide de la fonction de répartition de la loi normale $\mathcal{N}$(0,1).

2°) a) Simuler (sous R) un ensemble d'apprentissage de Highleyman et un ensemble test.

b) Estimer les fonctions discriminantes linéaire et quadratique, et les erreurs de prédiction associées; calculer les erreurs apparentes. Comparer au 1°).

3°) Estimer les prédicteurs logistiques linéaire, quadratique et cubique et les erreurs de prédiction associées; calculer les erreurs apparentes.

4°) Estimer le prédicteur de boosting
$\alpha_1 1_{X_1<u_1} + \alpha_2 1_{X_1>u_2} + \alpha_3 1_{X_2<u_3} + \alpha_4 1_{X>u_4}$
et son erreur de prédiction; calculer l'erreur apparente.