DATAMINING SUPERVISE.

I Le modèle probabiliste (à deux classes).

Soit $(X = \{X_1, ..., X_p\}, Y)$ une variable aléatoire à valeurs dans $R^p \times \{0,1\}$; X est la variable **prédictive** et Y la **réponse**.

Un **prédicteur** de Y sera une fonction $g(X)$ où $g : R^p \to \{0,1\}$.

Remarque: le problème est dit **séparable** si la donnée de X détermine Y de façon unique (les supports des lois de X conditionnellement à $Y = 0$ et $Y = 1$ sont dispoints); l'aléatoire intervient du fait que l'on observera un nombre fini de couples (X_i, Y_i) pour en déduire un prédicteur sur tout l'espace R^p .

Notons :

- $h(X)$ la probabilité de $\{Y = 1\}$ conditionnellement à X :
$h(X) = P(Y = 1|X) = E(1_{Y=1}|X)$;

$-\mu$ la loi de probabilité (marginale) de X.

Soit $C \subset R^p \times \{0,1\}$;

$C = (C \bigcap \{Y = 0\}) \bigcup (C \bigcap \{Y = 1\}) = C_0 \times \{Y = 0\} \bigcup C_1 \times \{Y = 1\}$

$P((X,Y) \in C) = \int_{C_0} (1 - h(x).d\mu(x) + \int_{C_1} h(x).d\mu(x)$.

Coûts des erreurs:
a_0 si $\{Y=0\}$ est prédit par $g(X)=1$,
a_1 si $\{Y=1\}$ est prédit par $g(X)=0$.

Le **coût moyen** $\mathcal{C}(g)$ est défini par:

$$\mathcal{C}(g) = a_0.P(\{Y=0\}\bigcap\{g(X)=1\}) + a_1.P(\{Y=1\}\bigcap\{g(X)=0\})$$
$$= a_0.E(1_{Y=0}.1_{g(X)=1}) + a_1.E(1_{Y=1}.1_{g(X)=0}) \ .$$

Le **coût moyen** $\mathcal{C}(g|X)$ **conditionnel** à X vaut:

$$= a_0.1_{g(X)=1}E(1_{Y=0}|X) + a_1.1_{g(X)=0}.E(1_{Y=1}|X)$$
$$= a_0.1_{g(X)=1}.(1-h(X)) + a_1.(1-1_{g(X)=1}).h(X)$$
$$= a_0.1_{g(X)=1} + (a_1-(a_0+a_1).1_{g(X)=1}).h(X) \ .$$

Soit $g^*(X)$ un autre prédicteur de Y .

$$C(g|X) - C(g^*|X)$$
$$= a_0.(1_{g(X)=1} - 1_{g^*(X)=1}) - (a_0+a_1).(1_{g(X)=1} - 1_{g^*(X)=1}).h(X)$$
$$= (1_{g(X)=1} - 1_{g^*(X)=1})(a_0 - (a_0+a_1).h(X)) \ .$$

Choisissons $g^*(X) = 1_{h(X)>\frac{a_0}{a_0+a_1}}$; alors les deux parenthèses sont toujours de même signe quel que soit g. On peut conclure:

Théorème:

Le prédicteur $g^*(X) = 1_{h(X)>\frac{a_0}{a_0+a_1}}$ minimise le coût moyen conditionnel et le coût moyen de prédiction; il est appelé **prédicteur de Bayes** ($h(X) - \frac{a_0}{a_0+a_1}$ est le **score** de Bayes) .

Lien avec la théorie des tests.

Le théorème "des probabilités composées" de Bayes dit que

$P(Y=1|X=x) = \frac{P(X=x|Y=1).P(Y=1)}{P(X=x|Y=0).P(Y=0)+P(X=x|Y=1).P(Y=1)}$.

$h(x) > \frac{a_0}{a_0+a_1}$ équivaut à $\frac{P(X=x|Y=1)}{P(X=x|Y=0)} > \frac{a_0}{a_1}.\frac{P(Y=0)}{P(Y=1)}$, ou

$a_1.P(X=x|Y=1).P(Y=1) > a_0.P(X=x|Y=0).P(Y=0)$

$a_1.\mu_1(x).P(Y=1) - a_0.\mu_0(x).P(Y=0) > 0$

en notant $\mu_0(x) = P(X=x|Y=0)$ et $\mu_1(x) = P(X=x|Y=1)$
les lois de X sous les hypothèses $H_0 : Y=0$ et $H_1 : Y=1$.

Le **lemme de Neyman-Pearson** dit que $\{x \in R^p \,|\, \mu_1(x) > K.\mu_0(x)\}$ est la région de rejet d'un test le plus puissant parmi les tests de même niveau:
en théorie des tests, on minimise l'erreur de 2ème espèce (à erreur de première espèce fixée) ;
en classification, on minimise la somme pondérée des erreurs.

Cas $a_0 = a_1 (= 1)$.
$\mathcal{C}(g) = P(g(X) \neq Y)$; $g^*(X) = 1_{h(X) > \frac{1}{2}}$.

$\varepsilon^*(X) = \mathcal{C}(g^*|X) = P(g^*(X) \neq Y|X) = 1_{h(X)>\frac{1}{2}} + (1 - 2.1_{h(X)>\frac{1}{2}}).h(X)$
$= \begin{cases} 1 - h(X) \, si \, h(X) > \frac{1}{2} \\ h(X) \, si \, h(X) \leq \frac{1}{2} \end{cases} = min(h(X), 1 - h(X))$.

Ainsi $\varepsilon^* = \mathcal{C}(g^*) = E(min(h(X), 1 - h(X))$.

Prédicteur (ou échantillonneur) de Gibbs.

C'est, conditionnellement à X, une variable de Bernoulli : étant donné X, on choisit au hasard $\widetilde{g}(X) = 1$ avec probabilité $h(X)$.

$E(1_{\widetilde{g}(X)=1}.1_{Y=0}|X) = P(1_{\widetilde{g}(X)=1}|X)P(1_{Y=0}|X) = h(X)(1-h(X))$
$E(1_{\widetilde{g}(X)=0}.1_{Y=1}|X) = P(1_{\widetilde{g}(X)=0}|X)P(1_{Y=1}|X) = (1-h(X))h(X)$

$\widetilde{\varepsilon}(X) = \mathcal{C}(\widetilde{g}|X) = 2h(X).(1-h(X))$
$= 2.min(h(X), 1-h(X)).(1-min(h(X), 1-h(X))$
$= 2\varepsilon^*(X)(1-\varepsilon^*(X))$
$\widetilde{\varepsilon} = E(\widetilde{\varepsilon}(X)) = 2.E(\varepsilon^*(X)(1-\varepsilon^*(X)) \leq 2.E(\varepsilon^*(X)) = 2.\varepsilon^*$, d'où:

Proposition: $\varepsilon^* \leq \widetilde{\varepsilon} \leq 2\varepsilon^*$
(pour ε^* petit, le prédicteur de Gibbs sera un **bon** prédicteur !!)

Et maintenant des statistiques.

En réalité, la loi de (X, Y) est inconnue. Il nous faut estimer la fonction g^*(ou la fonction h) à l'aide des réalisations (X_i, Y_i) d'un **ensemble d'apprentissage** $\mathcal{A}_n$.

Hypothèse fondamentale 1 (n-échantillon)

Les couples (X_i, Y_i) , $i = 1...n$, de l'ensemble d'apprentissage constituent un n-échantillon, c'est-à-dire sont indépendants de même loi.

Cette hypothèse permettra en particulier d'utiliser des théorèmes limites (loi des grands nombres, TLC...) pour obtenir des propriétés des estimateurs pour n grand...

Hypothèse fondamentale 2 ("généralisation")

Le prédicteur obtenu sur l'ensemble d'apprentissage est valide pour de nouvelles données d'étude $\mathcal{E}$ où il s'agit de prédire des Y_i inconnus.

Remarque: nous avons vu que h est déterminé par μ_0 , μ_1 et $P(Y=1)$. Si $P(Y=1)$ est différente sur $\mathcal{E}$, il faudra utiliser l'estimateur $\widehat{h_E}(X) = \frac{P_E(Y=1)}{P_A(Y=1)} h_A\widehat{(X)}$, et donc connaître (ou estimer) $P_E(Y=1)$...

Le comportement de l'estimateur obtenu par apprentissage sur un nouveau jeu de données appelé **ensemble test** $\mathcal{T}$ indépendant de l'apprentissage, permet de **valider** l'hypothèse de généralisation.

Pour chercher le prédicteur, plusieurs méthodes sont possibles:

- on estime directement g^*(ou h), par exemple en estimant μ_0 , μ_1 et $P(Y=1)$;

- on minimise un estimateur de $P(g(X) \neq Y)$ sur une famille de fonctions $\mathcal{G}$.

Un(e suite de) prédicteur(s) $\widehat{g}_n$ sera **universellement convergent** si $P(\widehat{g}_n(X) \neq Y) \to \varepsilon^*$ quand $n \to +\infty$ (n est la taille de l'ensemble d'apprentissage $\mathcal{A}_n$) pour toute loi sur (X, Y).

On sait construire des prédicteurs universellement convergents, mais malheureusement la vitesse de convergence peut être aussi lente que l'on veut (par un choix approprié de la loi de (X, Y)).

Erreur de prédiction.

Pour chercher un prédicteur, juger de son intérêt ou encore comparer deux prédicteurs, il est très important d'estimer l'**erreur de prédiction** $P(g(X) \neq Y)$**.**

Sous l'hypothèse 1, $\widehat{\varepsilon}_n(g) = \frac{1}{n}\Sigma_1^n 1_{g(X_i)\neq Y_i}$ est un estimateur convergent (loi des grands nombres).

Calculée sur l'ensemble d'apprentissage, $\widehat{\varepsilon}_{A_n}(g)$ est dite **erreur empirique** ou **apparente** de g ; elle peut servir pour la recherche d'un prédicteur. Mais l'erreur apparente $\widehat{\varepsilon}_{A_n}(\widehat{g})$ d'un prédicteur estimé $\widehat{g}$ sera en géneral un mauvais estimateur de l'erreur de prédiction de $\widehat{g}$ (car $\widehat{\varepsilon}_{A_n}(\widehat{g})$ et $\widehat{g}$ sont très corrélés); par exemple l'erreur apparente du "$1 - NN$ prédicteur" est nulle !

Au contraire, calculée sur l'ensemble test $\mathcal{T}_m$, l'**erreur de test** $\widehat{\varepsilon}_{\mathcal{T}_m}(\widehat{g})$ est un estimateur sans biais convergent de $P(\widehat{g}(X) \neq Y)$ (conditionnellement à $\widehat{g}$) vu l'indépendance de $\widehat{g}$ et de l'ensemble $\mathcal{T}_m$.

En pratique, on cherchera des familles de prédicteurs $\mathcal{G}$ telles que:

1) le prédicteur qui minimise l'erreur apparente sur $\mathcal{G}$ soit proche du prédicteur optimal de la famille $\mathcal{G}$:
$P(\widehat{\varepsilon}_{A_n}(\widehat{g}_n) - Inf_{g\in\mathcal{G}}\varepsilon(g) > \varepsilon) \to 0$ quand $n \to +\infty$;

2) la famille $\mathcal{G}$ soit suffisamment "riche", pour que le prédicteur optimal de la famille $\mathcal{G}$ soit proche du prédicteur de Bayes.

Le phénomène de surparamétrisation:

on sera attentif à ce que $\mathcal{G}$ ne soit pas "trop" riche :
l'erreur apparente sera d'autant plus petite que la famille $\mathcal{G}$ est riche; le prédicteur sera très bon sur $\mathcal{A}_n$, mais cela au détriment de l'erreur de prédiction sur R^p tout entier.

Pour l'éviter, on choisira une suite croissante $\{\mathcal{G}_h\}$ de familles de prédicteurs et on partagera l'ensemble d'apprentissage $\mathcal{A}_n$ en deux parties indépendantes $\mathcal{A}_{n,1}$ et $\mathcal{A}_{n,2}$ par tirage aléatoire:

-pour chaque h, le prédicteur $\widehat{g_h}$ minimisera $\widehat{\varepsilon}(g)$, $g \in \mathcal{G}_h$, sur $\mathcal{A}_{n,1}$;

-on choisira $\widehat{h}$ qui minimise l'erreur de prédiction de $\widehat{g_h}$ estimée sur $\mathcal{A}_{n,2}$ (choix de h par **validation croisée**).

Le choix de familles $\{\mathcal{G}_h\}$ satisfaisant aux conditions 1 et 2 pourra s'appuyer sur des théorèmes de majoration d'erreur (Vapnik-Chervonenkis, mesures de **complexité**, d'entropie...).

Extension à k classes, $k > 2$.

Une stratégie consiste à effectuer des classifications par paires puis en déduire la classe finale.

Cette stratégie est cohérente si l'on suppose tous les coûts d'erreurs a_i égaux et les classes équiprobables: étant donné X , on choisit $\{Y = i\}$ contre $\{Y = j\}$ si $P(X|Y = i) > P(X|Y = j)$ (prédicteur de Bayes car $P(Y = i) = P(Y = j)$).

Ceci revient à choisir la classe i qui maximise $P(X|Y = i\}$ parmi les $i = 1...k$.

Exercice: exemple élémentaire.

$p = 1$, $\mu_0 = \mathcal{N}(m_0, 1)$, $\mu_1 = \mathcal{N}(m_1, 1)$, $P(Y = 1) = \frac{1}{2}$.

1) Montrer que le prédicteur de Bayes $g^*(X)$ choisit la classe 1 si:
$m_0 < m_1$ et $X > \frac{m_0+m_1}{2}$ ou bien $m_0 > m_1$ et $X < \frac{m_0+m_1}{2}$
et la classe 0 sinon.
Calculer son erreur de prédiction ε^*.

2) On ne connait pas m_0 et m_1. L'ensemble d'apprentissage $\mathcal{A}_n$ est constitué de deux m-échantillons ($n = 2m$) indépendants de lois respectives μ_0 et μ_1 ; on estime m_0 et m_1 par $\overline{X_0}$ et $\overline{X_1}$.

Calculer $\widehat{g^*}(X)$ puis son erreur de prédiction $\varepsilon^*_{\widehat{g^*}}$.

3) Effectuer des simulations en faisant varier n; on pourra choisir $m_0 + m_1 = 0$ et $\theta = m_1 - m_0 = 0.1, 0.2, 0.5, 1, 2, 5$.
Tracer les graphes des erreurs apparentes et des estimations d'erreur de prédiction en fonction de n pour les divers θ .