DATAMINING SUPERVISE.

IV Prédicteurs linéaires "adaptés".

Reprenons l'étude des prédicteurs linéaires, "adaptés" au cas non paramétrique.

1°) Le perceptron de Rosenblatt.

Hypothèse: les supports des lois de $\mathbf{X}$, conditionnellement à Y, ont des enveloppes convexes fermées disjointes de distance positive:
séparation exacte possible par un hyperplan (cas séparable).

Problème: soit $\mathcal{A}_n = \{(\mathbf{X}_i, Y_i) \in R^p \times \{-1,1\}, i = 1...n\}$ un ensemble d'apprentissage; chercher un hyperplan $< \mathbf{w}, \mathbf{x} > +b = 0$ de séparation, soit $< \overline{\mathbf{w}}, \overline{\mathbf{x}} >= 0$, avec $\overline{\mathbf{w}} = (\mathbf{w}, b)$ et $\overline{\mathbf{x}} = (\mathbf{x}, 1)$.

Un point $\mathbf{X}_i$ sera mal classé si $Y_i < \overline{\mathbf{w}}, \overline{\mathbf{X}}_i > \leq 0$ (explique le choix $Y_i \in \{-1,1\}$). On note $R = Max_{i=1...n}\{\left\|\overline{\mathbf{X}}_i\right\|\}$. L'hypothèse de séparabilité entraîne l'existence de $\mathcal{H}$: $< \overline{\mathbf{w}}_{opt}, \overline{\mathbf{x}} >= 0$ (avec $\left\|\overline{\mathbf{w}}_{opt}\right\| = 1$) tel que $Min_{i=1...n}\{Y_i < \overline{\mathbf{w}}_{opt}, \overline{\mathbf{X}}_i > > 0\} = \rho > 0$.

Algorithme du perceptron.
$\overline{\mathbf{w}}_0 = 0$.
Pour $i = 1...n$,
$\overline{\mathbf{w}}_k = \overline{\mathbf{w}}_{k-1} + Y_i \overline{\mathbf{X}}_i$ si $\mathbf{X}_i$ est mal classé: $Y_i < \overline{\mathbf{w}}_{k-1}, \overline{\mathbf{X}}_i > \leq 0$,
et on recommence jusqu'à ce que tous les $\mathbf{X}_i$ soient bien classés.

Théorème de finitude (Novikoff).
L'algorithme du perceptron converge en un nombre fini d'étapes.

$\|\overline{\mathbf{w}}_k\|^2 = \|\overline{\mathbf{w}}_{k-1}\|^2 + \left\|\overline{\mathbf{X}}_i\right\|^2 + 2Y_i < \overline{\mathbf{w}}_{k-1}, \overline{\mathbf{X}}_i > \leq$
$\|\overline{\mathbf{w}}_{k-1}\|^2 + R^2 \leq kR^2$ par récurrence sur k.
$< \overline{\mathbf{w}}_k, \overline{\mathbf{w}}_{opt} > = < \overline{\mathbf{w}}_{k-1}, \overline{\mathbf{w}}_{opt} > + Y_i < \overline{\mathbf{X}}_i, \overline{\mathbf{w}}_{opt} > \geq$
$< \overline{\mathbf{w}}_{k-1}, \overline{\mathbf{w}}_{opt} > + \rho \geq k\rho$.
L'inégalité de Cauchy-Schwarz entraîne
$k\rho \leq < \overline{\mathbf{w}}_k, \overline{\mathbf{w}}_{opt} > \leq \|\overline{\mathbf{w}}_k\| \left\|\overline{\mathbf{w}}_{opt}\right\| \leq R\sqrt{k}$, donc $k \leq \frac{R^2}{\rho^2}$.

Remarques:
-dans le cas non séparable, l'algorithme ne peut converger aussitôt que les enveloppes convexes de $\{\mathbf{X}_i | Y_i = 1\}$ et $\{\mathbf{X}_i | Y_i = -1\}$ se coupent;
-on peut choisir une initialisation arbitraire; la solution dépendra de l'initialisation; problème du choix d'un "bon" hyperplan?
-pour $\overline{\mathbf{w}}_0 = 0$, $\overline{\mathbf{w}}_{k-1} = \Sigma_{l=1}^{k-1} \alpha_l Y_l \overline{\mathbf{X}}_l$ par récurrence; l'algorithme se
réécrit: $\overline{\mathbf{w}}_k = \overline{\mathbf{w}}_{k-1} + Y_i.\overline{\mathbf{X}}_i$ si $Y_i \Sigma_{l=1}^{k-1} \alpha_l Y_l < \overline{\mathbf{X}}_l, \overline{\mathbf{X}}_i > \leq 0$;
il ne dépend des $\mathbf{X}_i$ que par l'intermédiaire des$< \overline{\mathbf{X}}_l, \overline{\mathbf{X}}_i >$, donc de la matrice de Gram$(< \mathbf{X}_l, \mathbf{X}_i >)$;
-la solution vaut $\overline{\mathbf{w}}^* = \Sigma_{i=1}^{n} d_i Y_i \overline{\mathbf{X}}_i$; les $\overline{\mathbf{X}}_i$ avec $d_i \neq 0$ sont appelés "**vecteurs du support**"(support vectors) de $\overline{\mathbf{w}}^*$.

2°) Les réseaux de neurones.

A) On généralise le perceptron par des **réseaux de neurones multicouches**:
$f(<\mathbf{w}_{li},\mathbf{x}> +b_{li})$ est la réponse du neurone i de la couche l à des excitations $\mathbf{x}=(x_1,...,x_p)$, et constitue une partie des excitations des neurones de la couche $l+1$.

On considère en général des réseaux à deux couches, constitués d'**une couche cachée** et d'**un neurone final**:
l'entrée $\mathbf{x}=(x_1,...,x_p)$ excite les k neurones de la couche cachée 1; les réponses de ces neurones excitent le neurone final 2.
La réponse $f(\Sigma_{i=1}^k w_{2,i}\, f(<\mathbf{w}_{1,i},\mathbf{x}> +b_{1,i})+b_2)$ permet de définir le prédicteur; il y a $k(p+1)+k+1$ paramètres.

Choix de la fonction f :
f linéaire ne permet rien de plus qu'un neurone unique;
f fonction de seuil (threshold) permet des extensions non-linéaires;
f fonction sigmoïde(logistique...) présente des avantages par rapport à une fonction seuil:
-avantages théoriques: traitement des problèmes de régression; liens avec les modèles logistiques (M.V. conditionnelle);
-avantages pratiques: lors de l'estimation des poids $\mathbf{w}_i$, par minimisation d'une fonction de coût..

Propriétés théoriques: théorème d'approximation uniforme des fonctions continues (par le théorème de Stone-Weierstrass) par la réponse lorsque le nombre de neurones de la couche cachée tend vers l'infini. On peut montrer qu'un modèle multi-couches peut se ramener à un modèle à 2 couches (couche cachée + neurone final).

B) Estimation des paramètres: la fonction de coût est la somme de carrés des écarts $\Sigma_{i=1}^{n}[Y_i - reponse(\mathbf{X}_i, \mathbf{w}, \mathbf{b})]^2$ des réponses du réseau aux Y_i de l'ensemble d'apprentissage $\mathcal{A}_n$, que l'on minimise en $(\mathbf{w}, b)$ par la méthode de rétro-propagation ou des méthodes numériques (méthode du gradient...).

Les problèmes:
-problème 1: la fonction de coût est-elle bien adaptée au problème de classification ?
-problème 2: **nombreux minima** locaux (tenir compte des k! permutations possibles sur les k neurones de la couche cachée...), d'où l'instabilité des configurations associées (c'est-à-dire les $\mathbf{w}_{il}$ et les b_{il}). En général, on réalisera plusieurs essais (avec des initialisations différentes), et on choisira le meilleur par validation croisée.

-problème 3: **lenteur** de la vitesse de convergence; la poursuite de l'algorithme se traduit souvent par un "pseudo-accroissement" de la complexité du modèle.
-problème 4: choix du nombre de couches cachées, et du nombre de neurones par couche (k dans le cas d'une seule couche cachée), qui traduisent la **complexité** du réseau de neurones. Pour éviter toute surparamétrisation, il est nécessaire d'estimer l'erreur de prédiction par validation croisée.

Conclusion.
Les difficultés rencontrées lors de l'utilisation des réseaux de neurones conduisent à s'interroger sur leurs qualités effectives.

3°) La méthode SVM.

A) Choix de Vapnik.

Revenons au perceptron dans le cas linéairement séparable, et au problème du choix "optimal" de l'hyperplan $< \mathbf{w}, \mathbf{x} > +b = 0$, lorqu'on l'on choisit la classe $\{Y = sgn(< \mathbf{w}, \mathbf{X} > +b - 1)\}$ (par homogénéité sur $(\mathbf{w}, b)$).

On choisit l'hyperplan médiateur du plus petit segment qui joint les convexes engendrés par les $\{\mathbf{X}_i | Y_i = 1\}$ et les $\{\mathbf{X}_i | Y_i = -1\}$:
- cet hyperplan minimise le majorant de Vapnik-Chervonenkis (très pessimiste) de $P(Erreur\, pred - Err\, app)$ (permet d'estimer l'erreur de prédiction par l'erreur apparente);

- il maximise les distances (les **marges**) de l'hyperplan aux points $\mathbf{X}_i$, égales à $\frac{|< \mathbf{w}, \mathbf{X}_i > +b|}{\|\mathbf{w}\|} \geq \frac{1}{\|\mathbf{w}\|}$.

Il nous faut minimiser $\|\mathbf{w}\|^2$, avec les contraintes affines $Y_i(< \mathbf{w}, \mathbf{X}_i > +b) - 1 \geq 0$, $i = 1...n$. La convexité de $\|\mathbf{w}\|^2$ et sa croissance à l'infini entraînent l'existence d'une solution **unique**.

La méthode des multiplicateurs de Lagrange (Appendice) conduit à:

1)minimiser $L(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \Sigma_{i=1}^n \alpha_i [Y_i(< \mathbf{w}, \mathbf{X}_i > +b) - 1]$ en $(\mathbf{w}, b)$ avec $\alpha_l \geq 0 \, \forall l$.
Dérivons en $(\mathbf{w}, b)$: $\widehat{\mathbf{w}}(\alpha) = \Sigma_{l=1}^n \alpha_l Y_l \mathbf{X}_l$, $\Sigma_{l=1}^n \alpha_l Y_l = 0$.
$L(\widehat{\mathbf{w}}, \widehat{b}, \alpha) = -\frac{1}{2} \Sigma_{i=1}^n \Sigma_{l=1}^n \alpha_i \alpha_l Y_i Y_l < \mathbf{X}_i, \mathbf{X}_l > + \Sigma_{i=1}^n \alpha_i$.

2)minimiser $< \alpha, G(\hat{\mathbf{w}}, \hat{b}, \alpha) >= \Sigma_{i=1}^{n}\Sigma_{l=1}^{n}\alpha_i\alpha_l Y_i Y_l < \mathbf{X}_i, \mathbf{X}_l > -\Sigma_{i=1}^{n}\alpha_i$ en α avec $\Sigma_{l=1}^{n}\alpha_l Y_l = 0$ et $\alpha_l \geq 0\,\forall l$ ou maximiser $-2L(\hat{\mathbf{w}}, \hat{b}, \alpha) + \Sigma_{i=1}^{n}\alpha_i$: problème **dual**.

Algorithme SMO: on fixe $n-2$ des α_i, sauf i_1 et i_2; $\Sigma_{l=1}^{n}\alpha_l Y_l = 0$ s'écrit $\Sigma_{i\neq i_1,i_2}\alpha_l Y_l + \alpha_{i_1}Y_{i_1} + \alpha_{i_2}Y_{i_2} = 0$, et on maximise par rapport au seul paramètre "libre" α_{i_1}...

Extension au cas non séparable.

On introduit des variables "ressort" ξ_i pour assouplir les contraintes:
$Y_i(< \mathbf{w}, \mathbf{X}_i > +b) \geq 1 - \xi_i$, $\xi_i \geq 0$. Le problème dual a la même forme que dans le cas séparable, à part le fait que les α_i sont

bornés: $0 \leq \alpha_i \leq C\,\forall i$.

B) Espace intermédiaire, extension par noyau.

Lorsqu'il est impossible de séparer les classes par un hyperplan, on peut plonger R^p dans un espace (plus gros) $\mathcal{F}$, dit **espace intermédiaire** ("feature space") par une application $\phi : R^p \to \mathcal{F}$, dans lequel les classes deviennent séparables.
On substitue $< \phi(\mathbf{X}_i), \phi(\mathbf{X}_l) >= K(\mathbf{X}_i, \mathbf{X}_l)$ à $< \mathbf{X}_i, \mathbf{X}_l >$ dans l'algorithme précédent. Le **noyau** $K(\mathbf{x}, \mathbf{y})$ est de type positif; en effet
$\Sigma_{i=1}^{k}\Sigma_{l=1}^{k}a_i a_l K(\mathbf{x}_i, \mathbf{x}_l) = \Sigma_{i=1}^{k}\Sigma_{l=1}^{k}a_i a_l < \phi(\mathbf{x}_i), \phi(\mathbf{x}_l) >=$
$< \Sigma_{i=1}^{k}a_i\phi(\mathbf{x}_i), \Sigma_{l=1}^{k}a_l\phi(\mathbf{x}_l) >= \left\|\Sigma_{i=1}^{k}a_i\phi(\mathbf{x}_i)\right\|^2 \geq 0$.

Réciproquement, si le noyau $K(\mathbf{x},\mathbf{y})$ est de type positif, il existe $\mathcal{F}$ et $\phi : R^p \to \mathcal{F}$ tels que $K(\mathbf{x},\mathbf{y}) =< \phi(\mathbf{x}), \phi(\mathbf{y}) >$ (**Théorème de Mercer**). Il nous suffit donc de connaître $K(\mathbf{x},\mathbf{y})$.

[Idée de démonstration: à tout $\mathbf{x} \in R^p$, on associe $Z_\mathbf{x}$ v.a.r. gaussienne telle que $Cov(Z_\mathbf{x}, Z_\mathbf{y}) = K(\mathbf{x},\mathbf{y})$; l'application $\mathbf{x} \to Z_\mathbf{x}$ de R^p dans $\mathcal{F} = L^2(\Omega, \mathcal{A}, P)$ est l'application ϕ cherchée].

Exemples de noyaux:

a) noyau **polynômial** de degré k: $K(\mathbf{x},\mathbf{y}) = (< \mathbf{x}, \mathbf{y} > +a)^k$
(on exhibe directement ϕ).

b) noyau **gaussien** $K(\mathbf{x},\mathbf{y}) = exp - \frac{\sigma^2}{2} \|\mathbf{x} - \mathbf{y}\|^2$.
On a $exp - \frac{1}{2} \|\mathbf{u}\|^2 = (\frac{1}{2\pi})^{\frac{p}{2}} \int_{R^p} e^{i<\mathbf{u},\mathbf{t}>} e^{-\frac{1}{2}\|\mathbf{t}\|^2} d\mathbf{t}$ (f.c. d'une $\mathcal{N}_p(0, \mathbf{I}_p)$).

$\Sigma_{h=1}^n \Sigma_{l=1}^n a_h a_l exp - \frac{1}{2} \|\mathbf{x}_h - \mathbf{x}_l\|^2 =$
$(\frac{1}{2\pi})^{\frac{p}{2}} \int_{R^p} \Sigma_{h=1}^n \Sigma_{l=1}^n a_h a_l e^{i<\mathbf{x}_h - \mathbf{x}_l, \mathbf{t}>} e^{-\frac{1}{2}\|\mathbf{t}\|^2} d\mathbf{t} =$
$(\frac{1}{2\pi})^{\frac{p}{2}} \int_{R^p} |\Sigma_{h=1}^n a_h e^{i<\mathbf{x}_h, \mathbf{t}>}|^2 e^{-\frac{1}{2}\|\mathbf{t}\|^2} d\mathbf{t} \geq 0.$

c) noyau **laplacien** $K(\mathbf{x},\mathbf{y}) = exp - \lambda \|\mathbf{x} - \mathbf{y}\|$ (même méthode, $e^{-|t|}$ est la f.c. de la loi de Cauchy; plus généralement, tout noyau de type positif est la transformée de Fourier d'une mesure positive: théorème de **Bochner**).

C) Choix d'un noyau.

Le théorème suivant nous permet d'estimer l'erreur de prédiction d'une règle de classification linéaire par son erreur apparente.

Théorème (Devroye-Wagner).
Notons:
$\mathcal{A}_n = \{(\mathbf{X}_i, Y_i) \in R^p \times \{0, 1\}, i = 1...n\}$ un ensemble d'apprentissage,
$g_n(\mathbf{x}) = 1_{<w,\Phi(\mathbf{x})>+b>0}$ une règle de classification linéaire,
$\varepsilon_{g_n} = P(g_n(\mathbf{X}) \neq Y)$ l'erreur de prédiction de g_n ($(\mathbf{X}, Y)$ indépendant de l'ensemble d'apprentissage $\mathcal{A}_n$),
$\varepsilon_{g_n,\mathcal{A}_n} = \frac{1}{n}\Sigma_{i=1}^{n} 1_{g_n(\mathbf{X_i})\neq Y_i}$l'erreur apparente .
Alors $P(|\Sigma_{i=1}^{n} \alpha_i \varepsilon_{g_n,\mathcal{A}_n} - \varepsilon_{g_n}| > a) \leq 8n^p e^{-\frac{na^2}{32}}$ (convergence en probabilité de l'erreur apparente vers l'erreur de prédiction quand $n \to +\infty$) [démonstration à l'aide de la théorie de Vapnik-Chervonenkis].

Il reste à choisir la famille $\{g_n\}$ (par un choix du noyau Φ) telles que ε_{g_n} converge vers l'erreur de Bayes ε^*.

On pourra effectuer une estimation "adaptative" des paramètres du noyau (comme le noyau gaussien) par validation croisée...

Exercice: ensemble de Highleyman (fin).

1°) Réseaux de neurones. Par validation croisée, déterminer le nombre optimal de neurones (avec 1 couche cachée).

2°) Modèles SVM: noyaux polynômiaux (le degré 2 suffit), noyau gaussien (estimation du paramètre par validation croisée).

Appendice: méthodologie de Lagrange.

Soient F, G_1,...,G_n des fonctions à valeurs réelles, $\mathbf{G} = (G_1, ...G_n)$.

1)On veut minimiser $F(\mathbf{x})$ avec la contrainte $\mathbf{G}(\mathbf{x}) = 0$.

a) On minimise $F(\mathbf{x}) - \Sigma_{i=1}^{n} \alpha_i G_i(\mathbf{x})$ en $\mathbf{x}$**;** les α_i sont les **multiplicateurs de Lagrange**. Soit $\hat{\mathbf{x}}(\alpha)$ une solution:
$F(\hat{\mathbf{x}}(\alpha)) - < \alpha, \mathbf{G}(\hat{\mathbf{x}}(\alpha)) > \leq F(\mathbf{x}) - < \alpha, \mathbf{G}(\mathbf{x}) > \forall \alpha \in R^n$, $\forall \mathbf{x}$.

b) On résout le système $\mathbf{G}(\hat{\mathbf{x}}(\alpha)) = 0$ en α**.** Soit $\hat{\alpha}$ une solution:
$F(\hat{\mathbf{x}}(\hat{\alpha})) = F(\hat{\mathbf{x}}(\hat{\alpha})) - < \hat{\alpha}, \mathbf{G}(\hat{\mathbf{x}}(\hat{\alpha})) > \leq F(\mathbf{x}) - < \widehat{\alpha}, \mathbf{G}(\mathbf{x}) >= F(\mathbf{x})$
pour $\mathbf{G}(\mathbf{x}) = 0$, donc $\hat{\mathbf{x}}(\hat{\alpha})$ est une solution du problème initial.

2)On veut minimiser $F(\mathbf{x})$ avec $\mathbf{G}(\mathbf{x}) \in (R^+)^n$, notée $\mathbf{G}(\mathbf{x}) \geq 0$.

a) On minimise $F(\mathbf{x}) - < \alpha, \mathbf{G}(\mathbf{x}) >$ en $\mathbf{x}$, avec $\alpha \in (R^+)^n$**.**
Soit $\hat{\mathbf{x}}(\alpha)$ une solution:
$F(\hat{\mathbf{x}}(\alpha)) - < \alpha, \mathbf{G}(\hat{\mathbf{x}}(\alpha)) > \leq F(\mathbf{x}) - < \alpha, \mathbf{G}(\mathbf{x}) > \forall \alpha \geq 0$, $\forall \mathbf{x}$.

b) On minimise $< \alpha, \mathbf{G}(\hat{\mathbf{x}}(\alpha)) >$ en α avec $\alpha \geq 0$.
Soit $\hat{\alpha}$ une solution:
$F(\hat{\mathbf{x}}(\hat{\alpha})) - < \hat{\alpha}, \mathbf{G}(\hat{\mathbf{x}}(\hat{\alpha})) > \leq F(\mathbf{x}) - < \hat{\alpha}, \mathbf{G}(\mathbf{x}) > \forall \mathbf{x}$;
$< \hat{\alpha}, \mathbf{G}(\hat{\mathbf{x}}(\hat{\alpha})) > \leq 0$ car$< \alpha, \mathbf{G}(\hat{\mathbf{x}}(\alpha)) >= 0$ pour $\alpha = 0$,
donc $F(\hat{\mathbf{x}}(\hat{\alpha})) \leq F(\hat{\mathbf{x}}(\hat{\alpha})) - < \hat{\alpha}, \mathbf{G}(\hat{\mathbf{x}}(\hat{\alpha})) >$;
$F(\mathbf{x}) - < \hat{\alpha}, \mathbf{G}(\mathbf{x}) > \leq F(\mathbf{x})$ pour $\mathbf{G}(\mathbf{x}) \geq 0$, car $\hat{\alpha} \geq 0$;
ainsi $F(\hat{\mathbf{x}}(\hat{\alpha})) \leq F(\mathbf{x})$ pour $\mathbf{G}(\mathbf{x}) \geq 0$.