APPRENTISSAGE STATISTIQUE

CHAPITRE 1 – L'apprentissage statistique : pourquoi , comment ?

A-/ Introduction

Trois grandes familles de modèles statistiques :
- les réseaux de neurones
- les machines à vecteurs de support
- les cartes auto-adaptatives

B-/ Quelques définitions concernant les modèles

Le terme **modèle** est utilisé pour une équation paramétrée permettant de calculer la valeur d'une grandeur à modéliser à partir d'autres grandeurs appelées **variables** ou **facteurs**.
On distinguera les **modèles statiques** des **modèles dynamiques**, et les **modèles linéaires en leurs paramètres** des **modèles non linéaires en leurs paramètres**.

a) Modèles statiques

Un **modèle statique** est une fonction paramétrée notée g (x,w) , où x est le vecteur dont les composantes sont les valeurs des variables, et où w est le vecteur des paramètres du modèle.

1) Modèles statiques linéaires en leurs paramètres

Un modèle statique est linéaire en ses paramètres s'il est une combinaison linéaire de fonctions non paramétrées des variables ; il est de la forme

$$g(x,w)=\sum_{i=1}^{p} w_i f_i(x)$$

où f i est une fonction connue, non paramétrée, ou à paramètres connus. Ce modèle peut encore s'écrire sous la forme d'un produit scalaire :

$$g(x,w)=w.f(x)$$

où f(x) est le vecteur dont les composantes sont les fonctions $f_i(x)$.

Les polynômes, par exemple, sont des modèles linéaires en leurs paramètres : les fonctions $f_i(x)$ sont les monômes des variables x. Les polynômes sont néanmoins non linéaires en leurs variables.

On appelle **modèle linéaire** un modèle qui est linéaire en ses paramètres et en ses variables. Les modèles linéaires sont donc de la forme :

$$g(x,w)=\sum_{i=1}^{p} w_i x_i = w.x$$

Un **modèle affine** est un modèle linéaire qui contient une constante additive :

$$g(x,w)=w_0+\sum_{i=1}^{p-1} w_i x_i$$

2) Modèles statiques non linéaires en leurs paramètres

Nous étudierons particulièrement dans cet ouvrage les modèles non linéaires en leurs paramètres qui sont de la forme

$$g(x,w)=\sum_{i=1}^{p} w_i f_i(x,w')$$

	Classification (*k* plus proches voisins)	Prédiction (modèles linéaires)
Dilemme biais-variance gouverné par	$\frac{\text{Nombre d'exemples}}{\text{Nombre de plus proches voisins}}$	$\frac{\text{Nombre de paramètres}}{\text{Nombre d'exemples}}$
Limite inférieure de l'erreur de généralisation	Limite de Bayes	Variance du bruit

Tableau 1-1. Dilemme biais-variance pour la classification par la méthode des plus proches voisins et pour la prédiction par des modèles linéaires ou polynomiaux

le **dilemme biais-variance**, est la nécessité de trouver le meilleur compromis possible entre la capacité du modèle à apprendre les exemples d'apprentissage et sa capacité à généraliser à des situations non apprises.

C-/ Éléments de théorie de l'apprentissage

Cette section présente quelques résultats théoriques fondamentaux concernant l'apprentissage supervisé, pour la prédiction et la classification.
On présentera tout d'abord un formalisme général pour la modélisation par apprentissage. On introduira ensuite le classifieur de Bayes, et l'on en démontrera les propriétés. Enfin, on prouvera que le dilemme biais-variance est un phénomène général.

a) Fonction de perte, erreur de prédiction théorique

on définit donc une fonction dite « fonction de perte »

$$\pi[y^p, g(x,w)] \geq 0$$

où y p est la valeur souhaitée et g(x, w) est la valeur prédite par le modèle, dont les paramètres sont les composantes du vecteur w, étant donné le vecteur de variables x

Une distance naturelle, très fréquemment utilisée, est l'erreur quadratique de modélisation :

$$\pi[y^p, g(x,w)] = [y^p - g(x,w)]^2$$

On peut modéliser les résultats des mesures y^p comme des réalisations d'une variable aléatoire Y^p , et les vecteurs des variables x comme des réalisations d'un vecteur aléatoire X.
Alors les valeurs de la fonction de perte π deviennent elles-mêmes des réalisations d'une variable aléatoire Π , fonction de Y^p et de X, et il est naturel de caractériser la performance du modèle par l'espérance mathématique de Π , ou erreur de prédiction théorique, que nous noterons P^2

$$P^2 = E_\Pi = \iint \pi[y^p, g(x,w)]\, p_{y^p,X}\, dy^p\, dx$$

où $p_{y^p,X}$ est la probabilité conjointe de la variable aléatoire Y^p et du vecteur aléatoire X

L'erreur de prédiction théorique peut alors s'écrire :

$$P^2 = E_X[E_{Y^p|X}(\Pi)]$$

où $E_{Y^p|X}(\Pi)$ désigne l'espérance mathématique de la variable aléatoire $\Pi(Y^p|X)$, c'est à dire l'espérance mathématique de la fonction perte pour les prédictions effectuées par le modèle pour un vecteur de variables x données .

---

démonstration :

La probabilité conjointe peut s'écrire $p_{Y^p,X} = p_{Y^p}(y^p|x)\,p_x$

L'erreur de prédiction théorique s'écrit donc

$$P^2 = \iint \pi[y^p, g(x,w)]\,p_{y^p,X}\,dy^p dx = \int [\int \pi[y^p, g(x,w)]\,p_{y^p}(y^p|x)\,dy^p]\,p_X dx = E_X[E_{Y^p|X}(\Pi)]$$

Le meilleur modèle est le modèle pour lequel l'erreur de prédiction théorique est minimum. Appliquons cette propriété successivement à deux tâches : la prédiction et la classification.

1) Prédiction

On pose $\pi[y^p, g(x,w)] = [y^p - g(x,w)]^2$

Alors le meilleur modèle possible pour la **fonction de régression** de la grandeur à modéliser est

$$f(x) = E_{Y^p|X}$$

---

démonstration :

Rappelons que l'espérance mathématique de la fonction de perte est donnée par :

$$E_{Y^p|X}(\Pi) = \int (y^p - g(x,w))^2\,p_{y^p}(y^p|x)\,dy^p$$

Son minimum est obtenu pour le modèle f(x) tel que

$$0 = \left(\frac{dE_{Y^p|X}}{dg(x,w)}\right)_{g(x,w)=f(x)}$$

$$0 = \left(\frac{d\int (y^p - g(x,w))^2\,p_{y^p}(y^p|x)\,dy^p}{dg(x,w)}\right)_{g(x,w)=f(x)}$$

$$0 = 2\int (y^p - f(x))\,p_{y^p}(y^p|x)\,dy^p$$

$$0 = 2\int y^p\,p_{y^p}(y^p|x)\,dy^p - 2f(x)\int p_{y^p}(y^p|x)\,dy^p$$

La première intégrale n'est autre que l'espérance mathématique de Y p étant donné x ; la seconde est égale à 1 par définition de la densité de probabilité. On obtient ainsi : $f(x) = E_{Y^p|X}$

---

2) Classification : règle de Bayes et classifieur de Bayes

Considérons à présent un problème de classification à deux classes A et B. Affectons l'étiquette $y^p=+1$ à tous les exemples de la classe A et l'étiquette $y^p=-1$ à tous les exemples de la classe B.

Comme nous l'avons fait plus haut, nous cherchons une fonction g(x, w) qui permettra d'affecter à la classe A tous les éléments pour lesquels sgn[g(x, w)] = +1, et à la classe B tous les éléments pour lesquels sgn[g(x, w)] = −1.
Cette fonction doit être telle que l'erreur de prédiction théorique soit minimale

règle de décision de Bayes
Pour la classification, on ne cherche pas à approcher les
valeurs des résultats de mesures, mais à classer correctement des objets. On utilise donc une autre fonction de perte, mieux adaptée à ce problème :

$$\pi[y^p, sgn(g(x,w))]=0 \; si \; y^p=sgn(g(x,w))$$
$$\pi[y^p, sgn(g(x,w))]=1 \; si \; y^p\neq sgn(g(x,w))$$

donc la fonction de perte vaut 1 s'il y a eu erreur de classement pour l'objet décrit par x et 0 sinon

L'espérance mathématique de la variable aléatoire discrète Π n'est autre que la probabilité pour que le classifieur considéré commette une erreur de classification pour un objet décrit par x ; en effet :

$$E_{\Pi}(x)=1*Pr_{\Pi}(1|x)+0*Pr_{\Pi}(0|x)=Pr_{\Pi}(1|x)$$

La variable aléatoire Π est fonction de Y^p . Son espérance mathématique peut donc s'écrire :

$$E_{\Pi}(x)=\pi(+1,sgn(g(x,w)))Pr_{Y^p}(1|x)+\pi(-1,sgn(g(x,w)))Pr_{Y^p}(-1|x)$$

La probabilité d'appartenance d'un objet à une classe C connaissant le vecteur de variables x qui décrit cet objet, notée $Pr_{Y^p}(C|x)$ est appelée **probabilité à posteriori** de la classe C pour l'objet décrit par x.

On remarque que $E_{\Pi}(x)$ ne peut prendre que deux valeurs :

$$E_{\Pi}(x)=Pr_{Y^p}(1|x) \; si \; sgn(g(x,w))=-1$$
$$E_{\Pi}(x)=Pr_{Y^p}(-1|x) \; si \; sgn(g(x,w))=1$$

Supposons que la probabilité a posteriori de la classe A au point x soit supérieure à celle de la classe B :

$$Pr_{Y^p}(1|x)>Pr_{Y^p}(-1|x)$$

Rappelons que l'on cherche la fonction g(x,w) pour laquelle la probabilité d'erreur de classification au point x , c'est à dire $E_{\Pi}(x)$,soit minimum. La fonction g(x,w) pour laquelle $E_{\Pi}(x)$ est minimum est donc celle telle que sgn(g(x,w))=+1 puisque dans ce cas $E_{\Pi}(x)=Pr_{Y^p}(-1|x)$ qui est la plus petite des deux valeurs possibles.

A l'inverse si $Pr_{Y^p}(-1|x)>Pr_{Y^p}(1|x)$ la fonction g(x,w) qui garanti le plus petit taux d'erreur en x est telle que $sgn(g(x,w))=-1$

En résumé, le meilleur classifieur possible est celui qui, pour tout x, affecte l'objet décrit par x à la classe dont la probabilité a posteriori est la plus grande en ce point.

Cette règle de décision (dite règle de Bayes) garantit que le nombre d'erreurs de classification est minimal ; pour pouvoir la mettre en œuvre, il faut calculer (ou estimer) les probabilités a posteriori des classes.

Classifieur de Bayes

Le classifieur de Bayes utilise, pour le calcul des probabilités a posteriori, la formule de Bayes : étant donné un problème à c classes C i (i = 1 à c), la probabilité a posteriori de la classe C i est donnée par la relation

$$Pr(C_i|x) = \frac{p_x(x|C_i) Pr_{C_i}}{\sum_{j=1}^{c} p_X(x|C_i) Pr_{C_i}}$$

où $p_X(x|C_j)$ est la densité de probabilité du vecteur x des variables observées pour les objets de la classe C_j et Pr_{C_j} est la probabilité a priori de la classe C_j , c'est à dire la probabilité pour qu'un objet tiré au hasard appartienne à C_j .

Si toutes les classes ont la même probabilité a priori 1/c, la règle de Bayes revient à classer l'objet inconnu x dans la classe pour laquelle x a la plus grande vraisemblance : c'est une application de la méthode du maximum de vraisemblance.

3) Dilemme Biais-Variance