Apprentissage statistique
Cadre de la classification supervisée
Problématique de choix de modèles

Jean-Michel Marin

Université de Montpellier
Institut Montpelliérain Alexander Grothendieck (IMAG)
Institut de Biologie Computationelle (IBC)

HMMA 303

Prédiction
Contexte

On dispose des données $(x_1, y_1), \ldots, (x_n, y_n)$ fournies par l'observation d'un système

On fait l'hypothèse qu'il y a une dépendance entre les comportements des couples $(x, y) \in E \times F$

On souhaite, à partir des données, extraire cette dépendance pour prédire y lorsque l'on observe x

Prédiction
Contexte

Example (filtrage de spams)

Dans une base de données de courriels classés *spam* ou *non spam* se trouve les fréquences relatives de 60 mots

On veut construire un filtre automatique de spam qui déciderait au vu des fréquences relatives des 60 mots si l'on a affaire à un spam ou non

n correspond au nombre de courriels dans la base de données d'apprentissage $(x_1, y_1), \ldots, (x_n, y_n)$

$E = [0,1]^{\otimes 60}$ et $F = \{\text{spam}, \text{non spam}\}$

Prédiction
Contexte

Example (reconnaissance de caractères manuscrits)

Un procédé de lecture optique produit des images 16×16 de chiffres manuscrits

On veut construire un système de reconnaissance automatique qui interprète une image quelconque

Si chaque pixel est binaire, alors $\mathbb{E} = \{0,1\}^{\otimes 256}$ et $\mathbb{F} = \{0,1,\ldots,9\}$

Prédiction
Contexte

Example (credit scoring)

On dispose de données sur les risques de crédit

x vecteur contenant des informations sur la personne ayant sollicité un crédit

y est binaire et code le fait qu'il y a eu ou pas un incident de paiement

Objectif : construire un prédicteur qui identifie les individus à risque

Prédiction
Contexte

Example (aide au diagnostic médical)

Des données relatives au risque d'infarctus du myocarde ont été collectées pour plusieurs centaines d'individus males d'une population

x est le vecteur de 7 coordonnées relatives à la pression sanguine, la consommation de tabac, le cholestérol, la présence d'antécédents familiaux, l'obésité, la consommation d'alcool et l'âge

y, binaire, code le fait que l'individu a ou non fait un infarctus

Objectif : construire un prédicteur qui détecte les sujets à risque

Prédiction
Formalisation

Hypothèse : $(X_1, Y_1), \ldots, (X_n, Y_n)$, la base de données d'apprentissage, constituent un n-échantillon du couple (X, Y) à valeurs dans $E \times F$:

- $(X_i, Y_i) \perp\!\!\!\perp (X_j, Y_j)$ pour tout $i \neq j$
- les couples (X_i, Y_i) ont la même loi de probabilités
- (x_i, y_i) est la réalisation de (X_i, Y_i)

Y est la variable à prédire, la réponse
X est le vecteur des variables prédictives

Prédiction
Formalisation

Definition

On appelle prédicteur de Y à partir de X toute fonction $g : \mathbb{E} \longrightarrow \mathbb{F}$

La qualité d'un prédicteur est mesurée par son coût moyen

Definition

Le coût moyen d'un prédicteur g est

$$C(g) = \mathbb{E}\left[h(Y, g(X))\right]$$

où h est une fonction de coût unitaire $h : \mathbb{F} \times \mathbb{F} \longrightarrow \mathbb{R}^+$ qui facture l'écart entre la valeur prédite et la valeur observée

Prédiction
Formalisation

Example (classification binaire)

$Y \in F = \{0, 1\}$

Supposons que $h(0,0) = h(1,1) = 0$
$h(0,1) = a > 0$ et $h(1,0) = b > 0$

Comparons deux classifieurs

- le hasard $g_1(x) = \begin{cases} 0 & \text{avec proba } 1/2 \\ 1 & \text{avec proba } 1/2 \end{cases}$
- celui consistant à affecter tout le monde à la classe 1
  $g_2(x) = 1$

Prédiction
Formalisation

Example (classification binaire - suite)

Calculons le coût moyen de g_1

$g_1(x) = \mathbf{1}_{U \geqslant 1/2}$ où U est une variable aléatoire suivant une loi uniforme sur $[0, 1]$

$$\begin{aligned} C(g_1) &= \mathbb{E}\left[h(Y, g_1(X))\right] \\ &= a\mathbb{P}(\{Y = 0\} \cap \{U \geqslant 1/2\}) + b\mathbb{P}(\{Y = 1\} \cap \{U < 1/2\}) \\ &= \frac{a}{2}\mathbb{P}(Y = 0) + \frac{b}{2}\mathbb{P}(Y = 1) \end{aligned}$$

Prédiction
Formalisation

Example (classification binaire - suite)

Calculons le coût moyen de g_2

$$\begin{aligned} C(g_2) &= \mathbb{E}\left[h(Y,1)\right] \\ &= a\mathbb{P}(Y=0) \end{aligned}$$

Prédiction
Formalisation

Example (classification binaire - suite)

Comparons les deux coûts

$$\begin{aligned} C(g_1) &\leqslant C(g_2) \\ \frac{a}{2}\mathbb{P}(Y=0)+\frac{b}{2}\mathbb{P}(Y=1) &\leqslant a\mathbb{P}(\{Y=0\} \\ \frac{a}{2}\mathbb{P}(Y=0)+\frac{b}{2}\mathbb{P}(Y=0) &\geqslant \frac{b}{2} \\ \mathbb{P}(Y=0) &\geqslant \frac{b}{a+b} \end{aligned}$$

Cas particulier du coût symétrique $a=b$

$$C(g_1) \leqslant C(g_2) \Longleftrightarrow \mathbb{P}(Y=0) \geqslant \frac{1}{2}$$

Prédiction
Prédicteur optimal

Definition

Le prédicteur optimal est celui qui a le coût moyen le plus faible

On se place dans le cas de la classification supervisée binaire $\mathrm{F} = \{0, 1\}$

On considère la fonction de coût élémentaire la plus générale $\mathrm{h}(0,0) = \mathrm{h}(1,1) = 0$, $\mathrm{h}(0,1) = \mathrm{a}$, $\mathrm{h}(1,0) = \mathrm{b}$

On cherche le classifieur g^* qui minimise le coût moyen

$$\mathrm{g}^* \in \arg\min_{\mathrm{g}} \mathrm{C}(\mathrm{g})$$

Prédiction
Prédicteur optimal

Nous remarquons que $C(g) = \mathbb{E}\left[h(Y, g(X))\right] = \mathbb{E}[\mathbb{E}[h(Y, g(X))|X]]$

Minimiser $C(g)$ est équivalent à minimiser
$\mathbb{E}[h(Y, g(X))|X = x]$ pour tout $x \in \mathbb{E}$

Le minimum point par point minimise également l'intégrale

On cherche donc à minimiser

$$\mathbb{E}[h(Y, g(X))|X = x] =$$

$$a\mathbf{1}_{g(x)=1}\mathbb{P}(Y = 0|X = x) + b\mathbf{1}_{g(x)=0}\mathbb{P}(Y = 1|X = x)$$

Prédiction
Prédicteur optimal

On obtient

$$g^*(x) = \begin{cases} 1 & \text{si} \quad \mathbb{P}(Y=1|X=x) > \frac{a}{a+b} \\ 0 & \text{si} \quad \mathbb{P}(Y=1|X=x) < \frac{a}{a+b} \\ 1 \quad \text{ou} \quad 0 & \text{si} \quad \mathbb{P}(Y=1|X=x) = \frac{a}{a+b} \end{cases}$$

Cas particulier coût symétrique $a = b$

$$g^*(x) = \begin{cases} 1 & \text{si} \quad \mathbb{P}(Y=1|X=x) > 1/2 \\ 0 & \text{si} \quad \mathbb{P}(Y=1|X=x) < 1/2 \\ 1 \quad \text{ou} \quad 0 & \text{si} \quad \mathbb{P}(Y=1|X=x) = 1/2 \end{cases}$$

Prédiction
Prédicteur optimal

$$g^*(x) = \mathbf{1}_{\mathbb{P}(Y=1|X=x)>1/2}$$

est appelé le classifieur de Bayes

Ne pas confondre avec la règle de Bayes

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

Prédiction
Prédicteur optimal

Lorsque $a = b = 1$,
$h(Y, g(X)) = 0$ si classification correcte
$h(Y, g(X)) = 1$ si erreur de classification

$\Longrightarrow C(g)$ est alors le taux d'erreur de classification

Le classifieur de Bayes est celui qui minimise le taux d'erreur de classification

Risque

Un algorithme d'apprentissage prend en entrée un échantillon $A_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ et rend en sortie un prédicteur $\hat{g}_{A_n} : E \longrightarrow F$

$\hat{g}_{A_n}$ est une fonction aléatoire, sa performance s'apprécie à la valeur de son risque R

$$\begin{aligned} R(\hat{g}_{A_n}) &= \mathbb{E}[C(\hat{g}_{A_n})] \\ &= \mathbb{E}_{A_n}\left[\mathbb{E}_{(X,Y)}\left[h(Y, \hat{g}_{A_n}(X)\right]\right] \end{aligned}$$

Stratégies d'estimation

Definition

Le coût empirique du prédicteur g, noté $C_{A_n}(g)$, est tel que

$$C_{A_n}(g) = \frac{1}{n} \sum_{i=1}^{n} h(Y_i, g(X_i))$$

Dans le cas de la classification supervisée binaire et pour la fonction de coût symétrique avec $a = b = 1$, le coût empirique correspond à la proportion de mal classés par g dans l'échantillon d'apprentissage

$$C_{A_n}(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{Y_i \neq g(X_i)}$$

Stratégies d'estimation

Il existe trois stratégies

Méthode 1 on minimise le coût empirique par rapport à $g \in \mathscr{G}$ où $\mathscr{G}$ est une certaine classe de fonction

Cette stratégie ne fait sens que si les fonctions appartenant à $\mathscr{G}$ sont de même niveau de complexité

Méthode 2 le seconde stratégie vise dans un premier temps à l'estimation de $\mathbb{P}_{Y|X=x}$ pour tout $x \in \mathbb{E}$, la loi de probabilité de $Y|X = x$

puis à utiliser cette estimation pour construire un classifieur, en singeant éventuellement le classifieur ayant le coût moyen minimal

Stratégies d'estimation

Pour estimer $\mathbb{P}(Y = 1|X = x)$ on procède comme suit

- on estime les densités conditionnelles de X sachant $Y = 0$ et $Y = 1$ par $\hat{f}_0(x)$ et $\hat{f}_1(x)$
- on estime la loi marginale de Y, $\pi_0 = \mathbb{P}(Y = 0)$ et $\pi_1 = \mathbb{P}(Y = 1)$, par $\hat{\pi}_0$ et $\hat{\pi}_1$
- on applique la règle de Bayes et on estime $\mathbb{P}(Y = 0|X = x)$ par

$$\hat{\mathbb{P}}(Y = 0|X = x) = \frac{\hat{\pi}_0 \hat{f}_0(x)}{\hat{\pi}_0 \hat{f}_0(x) + \hat{\pi}_1 f_1(x)}$$

On peut alors construire un classifieur à l'aide de $\hat{\mathbb{P}}(Y = 0|X = x)$ et $\hat{\mathbb{P}}(Y = 1|X = x) = 1 - \hat{\mathbb{P}}(Y = 0|X = x)$

Stratégies d'estimation

Méthode 3 on modélise directement $\mathbb{P}(Y = 1|X = x)$

C'est ce que l'on fait dans le cas de la régression logistique tout est conditionné à x

Problématique de sélection de modèles et complexité

Les méthodes d'apprentissage laissent souvent à l'utilisateur le soin du réglage de paramètres qui ont un impact sur la complexité des prédicteurs auxquels elles donnent accès

Une fois fixé les valeurs de ces paramètres, l'ensemble des prédicteurs potentiels est ce que l'on appelle un modèle

Problématique de sélection de modèles et complexité

Example (classification par régression)

$\mathbb{E} = \mathbb{R}^p$ et $\mathbb{F} = \{0, 1\}$ classification supervisée binaire et p variables prédictives quantitatives

On utilise la procédure suivante

- on estime un régresseur linéaire par la méthode des moindres carré

$$\hat{f}_{\mathcal{A}_n}(x) = \hat{\alpha}_0 + \hat{\alpha}_1 x^{(1)} + \ldots + \hat{\alpha}_p x^{(p)}$$

  à l'aide de $\mathcal{A}_n$
- on propose comme classifieur

$$\hat{g}_{\mathcal{A}_n}(x) = 1 \Longleftrightarrow \hat{f}_{\mathcal{A}_n}(x) \geqslant 1/2$$

Problématique de sélection de modèles et complexité

Example (classification par régression - suite)

Pour toute sorte de raison, on peut envisager d'appliquer cette procédure à des parties + ou - restreintes de l'ensemble des variables

Fixer un sous-ensemble de régresseurs c'est choisir un modèle

Une mesure de complexité est par exemple le nombre de variables

Problématique de sélection de modèles et complexité

Example (méthode des k-plus-proches voisins)

$\mathbb{E} = \mathbb{R}^d$ et $\mathbb{F} = \{0, 1\}$

Elle fonctionne comme suit

- pour tout $x \in \mathbb{E}$ on note $\mathcal{V}_{A_n,k}(x)$ l'ensemble contenant les k-plus-proches voisins de x au sens d'une distance bien choisie
- $\hat{g}_{A_n,k}(x)$ est la classe majoritaire dans $\mathcal{V}_{A_n,k}(x)$

En cas d'égalité, on prend la classe majoritaire sur les $k-1$-plus-proches voisins

L'entier k joue sur le comportement du classifieur : plus k est grand moins le classifieur est complexe

Risque et complexité

Le coût empirique sous-estime le risque pour les modèles complexes

Avec le coût empirique, on choisit toujours le modèle le plus complexe

Donner trop de flexibilité au modèle conduit à des classifieurs de mauvaise qualité, c'est le phénomène de sur-apprentissage (overfitting)

Méthodes pratiques de sélection de modèles

Idée de base ne pas se fier au coût empirique calculé sur l'échantillon qui a servi à faire l'estimation/la prédiction

$$C_{\mathcal{A}_n}(\hat{g}_{\mathcal{A}_n}) = \frac{1}{n}\sum_{i=1}^{n} h(Y_i, \hat{g}_{\mathcal{A}_n}(X_i))$$

est un mauvais estimateur de

$$R(\hat{g}_{\mathcal{A}_n}) = \mathbb{E}_{\mathcal{A}_n}\left[\mathbb{E}_{(X,Y)}\left[h(Y, \hat{g}_{\mathcal{A}_n}(X)\right]\right]$$

On voudrait choisir un modèle, ie un niveau de complexité en fonction du risque

On veut déterminer le modèle de risque le plus faible et pour cela nous devons estimer le risque

Méthodes pratiques de sélection de modèles
Méthode de validation

Lorsque l'on a beaucoup de données (n très grand) un usage courant est de séparer les données en 2 parties à peu près égales

- $\mathscr{A}$ données d'apprentissage à l'aide desquelles sont construits les prédicteurs $\mathscr{A} \subset A_N$, $\#\mathscr{A} \approx n/2$
- $\mathscr{T}$ données de test qui servent à évaluer la qualité des prédicteurs à l'aide de

$$C_{\mathscr{T}}(\hat{g}_{\mathscr{A}}) = \frac{1}{\#\mathscr{T}} \sum_{(x,y)\in\mathscr{T}} h(y, \hat{g}_{\mathscr{A}}(x))$$

$\mathscr{T} \subset A_n$ et $\#\mathscr{T} \approx n/2$

Méthodes pratiques de sélection de modèles
Méthode de la validation croisée

Dans le cas précédent, on peut renverser les rôles joués par $\mathscr{A}$ et $\mathscr{T}$, on croise

De manière générale, lorsque la masse des données est insuffisante on découpe notre jeu de données en plus de 2 parties. on utilise la méthode de la validation croisée à K ensembles (K-fold-cross-validation)

Objectif : estimer les risques associés aux différents modèles et comparer les risques pour choisir un modèle

Méthodes pratiques de sélection de modèles
Méthode de la validation croisée

Algorithme

- on divise les données en K parties disjointes (partition) de tailles égales (si possible) $B_1, \ldots, B_K$ de tailles $n_1, \ldots, n_K$ avec $\sum_{i=1}^{K} n_i = n$

- pour chaque $i \in \{1, \ldots, K\}$
  on utilise $A_{-i} = A_n / B_i$ pour calculer $\hat{g}_{A_{-i}}(x) \in F$
  on calcule $\hat{R}_i = \frac{1}{n_i} \sum_{(x,y) \in B_i} h(y, \hat{g}_{A_i}(x))$

- on calcule l'estimateur final du risque

$$\hat{R} = \frac{1}{K} \sum_{i=1}^{K} \hat{R}_i$$

Méthodes pratiques de sélection de modèles
Méthode de la validation croisée

Si n est grand relativement à $p = \dim(E)$ (nombre de variables prédictives) on prend $K = 2$

Si n est de taille moyenne, on prend $K = 5$ ou $K = 10$

Si n est petit, on prend $K = n$, leave-one-out cross validation