Estimation Non - Paramétrique Estimations Fonctionnelles

Plan du Cours

Introduction : Problèmes concrets et motivations

Chapitre I : Estimateur à noyau pour variables réelles

- A - Histogrammes
- B - Estimations de la fonction de densité
- C - Estimations de la fonction de régression
- D - Applications

Chapitre II : Estimateur à noyau en dimension supérieur

- A - Le cas multivarié
- B - Le cas fonctionnel
- C - Applications

Introduction : Problèmes concrets et les motivations

Dans l'ensemble de ce cours nous supposons que pour chaque individu de l'échantillon on observe :

* une variable X
* un couple de variable (X, Y) où Y est la variable d'intérêt et X une variable explicative.

Nous supposerons que les échantillons sont tous constitués d'observations indépendantes et de même loi

* $(X_1, \dots, X_n) \overset{iid}{\sim} X$
* $((X_i, Y_i))_{i \in [\![1;n]\!]} \sim (X, Y)$.

A_ Données "galaxies" (library(MASS))

1_ Charger les données et décrire leurs natures

2_ Comment représenter la répartition des valeurs observées ? Faites varier le nombre de classes ou/et leurs bornes. Commenter les résultats obtenus.

3_ Quels défauts devrait-on tenter de gommer ?

* très sensible au choix du nombre de classes et/ou de leurs bornes
* constant par morceaux alors que la densité sous jacente est plutôt "lisse".

Objectifs : Proposer une méthode d'estimation de la densité d'une variable aléatoire réelle continue :

* qui ne fasse pas d'hypothèse sur la nature de cette densité (modèle non paramétrique)
* qui soit "lisse"
* qui ne dépende pas du choix des paramètres (telles que le nombre de classe ou leurs bornes).

Comme nous venons de le dire, nous nous attacherons dans le cadre de ce cours à proposer des méthodes d'estimations adaptés à des modèles non-paramétriques, c'est-à-dire pour lesquels on ne fait pas d'hypothèse a priori sur la nature de la fonction que l'on cherche à estimer mais simplement sur sa régularité. Cela permet d'avoir une approche exploratoire des données.

Une fois cette approche réalisée, on peut parfois observer que la densité semble de nature connue et que l'on peut la modéliser de manière paramétrique. Son estimation consiste alors à estimer chacun des paramètres à partir des données (cf cours de MS).

	Atouts	Défauts
Modèle paramétrique	• Bonne vitesse de convergence • Facile à interpréter	• Pas flexible • Nécessité d'avoir un a priori
Modèle non-paramétrique	• Pas d'hypothèse à priori • Plus flexible	• Vitesse de convergence plus lente • Plus difficile à interpreter

B - Données "pression"

1- Charger les données et décrire leurs natures.

2- Tracer un nuage de points représentant la pression en fonction de la température. Commenter le résultat obtenu ?

3- Comment estimer la fonction reliant la pression à la température ?

Il s'agit d'un modèle de régression non linéaire

$$Y = r(X) + \varepsilon \quad \text{où} \quad \mathbb{E}[\varepsilon | X] = 0$$

Alors $r(X) = \mathbb{E}[Y|X]$ car $\mathbb{E}[Y|X] = \mathbb{E}[r(X)|X] + \mathbb{E}[\varepsilon|X]$

$= r(X)$.

où r est une fonction mesurable que l'on cherche à estimer.

Si on est capable d'écrire r de manière paramétrique on peut ensuite s'en sortir en estimant chacun des paramètres (modèle paramétrique). Mais cela est souvent difficile et l'on préférerait ne pas faire d'hypothèse et utiliser seulement la régularité de la fonction r que l'on cherche à estimer.

Objectif : Trouver un estimateur non-paramétrique de r supposée continue :

* qui soit lisse
* qui ne fasse pas d'hypothèse sur la nature de r
* qui puisse être de nature très variée suivant la nature des données

C - Données Dopage

1 - Charger les données et décrire leurs natures
2 - Si on mesure le taux d'hématocrite d'un nouveau cycliste, comment lui attribuer une classe ? de manière pertinente ?

On a ici un problème de classification supervisée.
Pour chaque individu, on dispose
* X variable quantitative (donne de l'information sur la classe)
* Y variable qualitative (classe)

Objectif : Construire un classifieur à partir des données qui à toutes nouvelles valeurs x_0 de X associe une classe $\hat{y}_0$. Autrement dit, on veut être capable d'attribuer une classe à tout nouvel individu pour lequel on n'observe que X. On se sert bien sûr des données de l'échantillon

On va utiliser le Classifieur de Bayes :

$$\hat{y}_0 = \underset{y_0}{\operatorname{argmax}} \left(\mathbb{P}(Y = y_0 \mid X = x_0) \right)$$

C'est le meilleurs classifieurs possible. mais on ne peut l'utiliser directement puisque $y \longmapsto \mathbb{P}(Y = y \mid X = x_0)$ ne sont pas connues.
On peut toutes fois les estimer puisque :

$$\mathbb{P}(Y = y \mid X = x_0) = \mathbb{E}\left[\mathbb{1}_{\{Y = y_0\}} \mid X = x_0 \right].$$

$x_y : x_0 \longmapsto \mathbb{P}(Y = y \mid X = x_0)$ est en fait la fonction de régression du modèle suivant :

$$Z_y := \mathbb{1}_{\{Y = y\}}, \quad Z_y = x_y(X) + \varepsilon \text{ avec } \mathbb{E}[\varepsilon \mid X] = 0.$$

On peut donc pour tout y estimer $\hat{r}_y(x_0) = \widehat{\mathbb{P}(Y=y \mid X=x_0)}$

Objectif : Estimer, pour chaque classe y, $\hat{r}_y(x_0)$ (et donc $\mathbb{P}(Y=y \mid X=x_0)$) de manière non-paramétrique à l'aide d'un estimateur de la fonction de régression. Puis la classe pour laquelle la probabilité estimée est la plus forte.

$$\hat{y}_0 = \underset{y_0}{\operatorname{argmax}}\ \widehat{\mathbb{P}(Y=y_0 \mid X=x_0)} = \underset{y_0}{\operatorname{argmax}}\ \hat{r}_{y_0}(x_0)$$

D - Données Tecator

1 - Charger les données et tracer quelques courbes spectrométriq Commenter

2 - L'objectif de l'industriel est d'estimer le taux de graisse des morceaux de viandes à partir des courbes spectrométriques (économie de temps et d'argent)

On est dans le cas d'un modèle de régression sur variables fonctionnelles

$$Y = r(X) + \varepsilon \quad \text{où} \quad \mathbb{E}[\varepsilon \mid X] = 0$$

Y : variable aléatoire réelle

X : variable aléatoire fonctionnelle (courbe)

Comment peut-on généraliser les estimateurs définis au chapitre 1 au cas de variables explicatives qui sont des courbes

E- Données

1- On a pour chaque phonème un ensemble de log-périodogrammes des enregistrements sonores correspondant à ce phonème.

2- Charger les données et tracer quelques log-périodogrammes pour chaque phonème. Commenter.

Objectif : Etant donné un nouvel enregistrement, comment lui attribuer un phonème de manière pertinente ? Etendre au cas de variable explicative fonctionnelle la méthode de classification supervisée décrite plus tôt