Master PRIM
18 Décembre 2007
9-12h

Reconnaissance des formes

Epreuve correspondant au cours de Marie-Odile Berger
Documents distribués en cours autorisés

1 Estimation d'une densité de probabilité

On considère un ensemble de n mesures indépendantes $x_1, ..., x_n$ d'une variable aléatoire. On note $f(x)$ la densité de cette loi. On s'intéresse ici au problème de calculer f dans le cas paramétrique (on connaît le modèle de loi qui est suivi) et non paramétrique (la forme de f est inconnue a priori). On rappelle que, étant donné un ensemble A, la probabilité de cet ensemble est définie par $p(A) = \int_A f(x)dx$. Pour simplifier les choses, on suppose que les mesures x_i sont à valeur réelle.

1.1 Estimation au maximum de vraisemblance

On suppose que f suit une loi gaussienne de densité $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-1/2(\frac{x-m}{\sigma})^2}$. Il suffit donc d'estimer les paramètres m et σ pour déterminer f complètement. Utiliser la méthode du maximum de vraisemblance pour calculer m et σ en fonction des $x_1, ..., x_n$

Quelles sont en général les difficultés de l'"estimation paramétrique?

1.2 Estimation non paramétrique

On souhaite calculer f sans connaître a priori sur sa forme.

1. La figure 1 donne un exemple des x_i acquis (ne pas tenir compte des flèches numérotées de 1 à 4). Dessiner sur votre copie un exemple de fonction f compatible avec cet exemple.

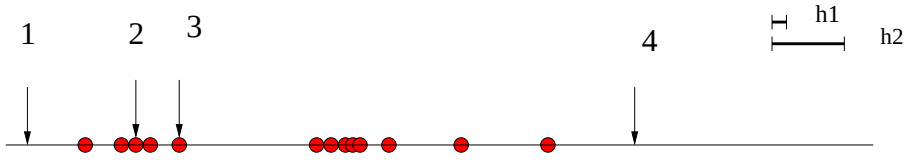

Figure 1: Un exemple des x_i acquis

2. La méthode la plus rudimentaire pour calculer une densité est de calculer un histogramme des échantillons. R étant subdivisé en intervalles réguliers de longueur h, on compte pour chaque intervalle le nombre d'échantillons p tombant dans cette intervalle. En calculant la probabilité de cet intervalle, montrer que on peut approcher $f(x)$ par $f(x) \approx \frac{1}{h}\frac{p}{n}$.
Montrez l'influence concrète du choix du paramètre h en dessinant approximativement la distribution obtenue avec cette méthode en utilisant respectivement des subdivisions de taille h_1 et h_2 comme indiqué dans la figure 1. Qu'en déduisez vous sur le choix de h?

3. On choisit maintenant d'adapter la méthode précédente en utilisant des intervalles de taille différente pour le comptage. On choisit, pour tout x pour lequel on veut calculer $f(x)$ le plus petit intervalle centré autour de x qui contienne au moins r échantillons (r étant une donnée définie a priori). On note h la longueur de cet intervalle (h dépend maintenant de x).

   Calculer la densité aux points marqués d'une flèche pour r=3 (il ne s'agit que d'un calcul approximatif). Comparer avec les valeurs fournies par la méthode de l'histogramme.
   Quels sont les avantages de la méthode proposée ?

2 Fusion de donnees

On considère le problème général de la fusion de données: étant données deux mesures m_1 et m_2 appartenant à R^p d'une même grandeur acquises par des moyens differents, on souhaite obtenir une meilleure estimation de cette grandeur prenant en compte les deux mesures. Ces deux mesures ont pour covariance associée les matrices Λ_1 et Λ_2.

Pour $p > 1$, on souhaite réaliser une estimation aux moindres carrés en recherchant le vecteur m_{new} minimisant

$$(m_{new} - m_1)^t \Lambda_1^{-1} (m_{new} - m_1) + (m_{new} - m_2)^t \Lambda_2^{-1} (m_{new} - m_2)$$

1. Montrer que ce critère est bien adapté au problème
2. Si, on choisit pour métrique une norme significative du problème, $||Z||^2 = Z^t \Lambda^{-1} Z$ où Λ est la matrice de covariance des mesures, montrer que la solution de l'équation $Z = A\theta$ aux moindres carrés est donnée par

$$\hat{\theta} = (A^t \Lambda^{-1} A)^{-1} A^t \Lambda^{-1} Z$$

3. Utiliser le résultat précédent pour montrer que l'estimation obtenue est

$$m_{new} = (\Lambda_1^{-1} + \Lambda_2^{-1})^{-1} (\Lambda_1^{-1} m_1 + \Lambda_2^{-1} m_2)$$

4. Calculer la covariance sur l'estimation m_{new}?
5. Applications numérique: Calculer l'estimation obtenue dans les deux cas suivants
   - $m_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $m_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\Lambda_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Lambda_2 = \begin{pmatrix} 1 & 0 \\ 0 & 10 \end{pmatrix}$
   - $m_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $m_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\Lambda_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Lambda_2 = \begin{pmatrix} 10 & 0 \\ 0 & 1 \end{pmatrix}$

   Expliquer en quoi ces résultats sont bien conformes à l'intuition.

3 Estimation robuste

- L'analyse en composantes principales est elle un processus robuste (i.e est-elle fortement influencée par la présence de données très erronées parmi les données)? Si oui, expliquez pourquoi. Si non, expliquer aussi pourquoi et dire comment on pourrait rendre cette méthode robuste et quelles sont les difficultés prévisibles.
- Le point de rupture d'un estimateur est la proportion d'observations incorrectes qu'un estimateur peut supporter sans donner de valeurs très erronées Expliquer pourquoi le point de rupture de l'estimateur moyenne est 0. Quel est le point de rupture de l'estimateur mediane?

  On considère l'estimateur suivant d'une grandeur p: $\hat{p}_h = \sum_{i=1}^{h} r_i(p)$ où les résidus $r_i(p)$ sont ordonnés par valeurs croissantes (pour fixer les idées, $p = (a, b)$ est une droite à estimer et $r_i(p) = |y_i - ax_i + b|$ est le résidu associé à chaque mesure (x_i, y_i). Quel est le point de rupture de cet estimateur?