méthodes paramétriques, méthodes non paramétriques

Thierry Chateau

Pascal Institute

2017

Plan

1. Introduction

2. Méthodes paramétriques
   - Maximum de vraisemblance

3. Méthodes non paramétriques
   - Choix d'un estimateur
   - KDE : méthode du noyau
   - Méthode des KPPV

Introduction

- Bayes est un classifieur optimal (sous certaines hypothèses) qui peut être appliqué si $P(\omega_i)$ et $p(\mathbf{x}|\omega_i)$ sont connus.
- Dans ce cadre, apprendre $P(\omega_i)$ et $p(\mathbf{x}|\omega_i)$ est une opération indispensable.
- On supposera que l'on dispose d'un nombre suffisant d'expériences connues (vecteur de paramétres et classe associée) afin d'avoir une vision globale du système.
- L'apprentissage consiste à obtenir $P(\omega_i)$ et $p(\mathbf{x}|\omega_i)$ à partir des expériences.
- Plusieurs méthodes sont alors utilisées.

Introduction

- Les méthodes paramétriques :
  - $p(\mathbf{x}|\omega_i)$ est décrite par une fonction littérale de k paramétres (exemple : loi gaussienne)
  - L'apprentissage consiste alors à trouver les k paramétres qui estiment au mieux l'observation de la densité faite à travers les expériences.
- Les méthodes non paramétriques
  - On cherche une méthode générale permettant, à partir des expériences, d'obtenir :

$$p(\mathbf{x}|\omega_i) = f(\mathbf{x}, \text{échantillons})$$

Méthodes paramétriques

- $p(\mathbf{x}|\omega_i)$ est décrite par une fonction littérale de k paramétres (exemple : loi gaussienne)
- L'apprentissage consiste alors à trouver les k paramétres qui estiment au mieux l'observation de la densité faite à travers les expériences.

Méthodes paramétriques

- On suppose que la loi de densité de probabilité $p(\mathbf{x}|\omega_i)$ est connue (sa forme littérale) :
  - loi normale,
  - loi de Poisson,
  - loi gamma,
  - ...
- L'estimation des paramétres décrivant la loi peut se faire de plusieurs façons :
  - méthode du maximum de vraisemblance,
  - méthode d'estimation Bayesienne,
  - ...

Méthodes paramétriques

- **Rq :** Présentation d'une méthode : maximum de vraisemblance dans le cas de l'apprentissage supervisé

Maximum de vraisemblance

- Soient E_1, E_2, E_c, c ensembles d'éhantillons (expériences) représentant les c classes recherchées :
  - échantillons tiré indépendamment
  - expériences représentatives de $p(\mathbf{x}|\omega_i)$
- **Hypothèse :** la forme paramétrique de $p(\mathbf{x}|\omega_i)$ est connue, déterminée par un vecteur de paramétres θ_i.
- par exemple, $p(\mathbf{x}|\omega_i)$ est une loi normale $N(\boldsymbol{\mu}_i, \Sigma_i)$, donc $\theta_i = \{\boldsymbol{\mu}_i, \Sigma_i\}$.
- Si les formes sont définies dans $\boldsymbol{R}^d$, alors :

$$\theta_i = \{\mu_1, \mu_2, ..\mu_d, \sigma_{22}, \sigma_{23}, .., \sigma_{2d}, \sigma_{33}, \sigma_{34}, .., \sigma_{3d}, ..., \sigma_{dd}\}$$

Maximum de vraisemblance

- **Notation** : pour indiquer que $p(\mathbf{x}|\omega_i)$ dépend de θ_i, on le notera : $p(\mathbf{x}|\omega_i, \theta_i)$
- **But** : on cherche à estimer les paramétres de la loi de densité de probabilité (θ_i)
- **Hypothèse** : les echantillons de E_i ne donnent pas d'informations sur $\theta_j,\ j \neq i$
- D'où un raisonnement indépendant sur chaque classe (suppression de l'indice i).

$$E = \{\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, .., \mathbf{X}_n\}$$

Maximum de vraisemblance

- la probabilité d'obtenir le tirage E, si la classe suit la loi définie par le vecteur de paramétres $\boldsymbol{\theta}$) est :

$$P(E, \boldsymbol{\theta}) = \prod_{k=1}^{n} P(\mathbf{X}_k | \boldsymbol{\theta})$$

- Cherchons la valeur du vecteur $\boldsymbol{\theta}$, noté $\hat{\boldsymbol{\theta}}$, qui maximise $P(E, \boldsymbol{\theta})$

Maximum de vraisemblance

- si $\boldsymbol{\theta} = \{\theta_1, \theta_2, .., \theta_p\}$:

$$\nabla_{\boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial}{\partial \theta_1} \\ \frac{\partial}{\partial \theta_2} \\ . \\ \frac{\partial}{\partial \theta_p} \end{pmatrix}$$

- On définit l'expression $l(\boldsymbol{\theta}) = \log[P(E|\boldsymbol{\theta})]$:

$$l(\boldsymbol{\theta}) = \sum_{k=1}^{n} \log[P(\mathbf{X}_k|\boldsymbol{\theta})]$$

Maximum de vraisemblance

- Alors,

$$\nabla_{\boldsymbol{\theta}} l = \nabla_{\boldsymbol{\theta}} \sum_{k=1}^{n} \log[P(\mathbf{X}_k|\boldsymbol{\theta})]$$

- $P(E|\theta)$ sera maximum si :

$$\nabla_{\boldsymbol{\theta}} l = \mathbf{0}$$

Les méthodes non paramétriques

- Méthode paramétrique $\rightarrow$ connaissance de l'expression littérale de $p(\mathbf{x}|\omega_i)$:
  - les formes classiques sont unimodales (exception : possibilité de mixture de gaussiennes)
  - hypothèse forte (si erronnée, cela peut conduire à de très mauvais résultats)
- Les méthodes non paramétriques prennent en compte les échantillons et leur répartition spatiale dans l'espace des paramétres.
- la conséquence et une estimation de $p(\mathbf{x}|\omega_i)$ plus proche de la réalité

Les méthodes non paramétriques

Principe de base

- Estimation de $p(\mathbf{x}|\omega_i)$ ou $p(\omega_i|\mathbf{x})$
- Soit $\mathcal{D}$ un domaine inclu dans l'espace des attributs ($\mathcal{D} \subset \boldsymbol{R}^d$) et pouvant être considéré comme un voisinage du vecteur de paramétres $\mathbf{x}$ pour lequel on cherche $p(\mathbf{x}|\omega_i)$.
- Hypothèse : $p(\mathbf{x}|\omega_i) \approx$ const. sur $\mathcal{D}$
- alors :

$$p(\mathbf{x} \in \mathcal{D}) = \int_{\mathcal{D}} p(\mathbf{x}'|\omega_i) d\mathbf{x}' \approx p(\mathbf{x}|\omega_i) \int_{\mathcal{D}} d\mathbf{x}'$$

Les méthodes non paramétriques

- Si $V(\mathcal{D})$ est l'hypervolume de $\mathcal{D}$:

$$p(\mathbf{x} \in \mathcal{D}) \approx p(\mathbf{x}|\omega_i)V(\mathcal{D})$$

- Finalement :

$$\forall \mathbf{x} \in \mathcal{D} p(\mathbf{x}|\omega_i) \approx \frac{p(\mathbf{x} \in \mathcal{D})}{V(\mathcal{D})}$$

- Si t échantillons, sur n au total, se trouvent dans le domaine $\mathcal{D}$, alors :

$$p(\mathbf{x} \in \mathcal{D}) \approx \frac{t}{n}$$

Les méthodes non paramétriques

- et : $\boxed{p(\mathbf{x}|\omega_i) \approx \frac{\frac{t}{n}}{V(\mathcal{D})}}$
- Soit $\mathbf{x}_0$ une mesure et $D(\mathbf{x}_0)$ un domaine entourant $\mathbf{x}_0$; on cherche à estimer $\hat{P}(\mathbf{x}_0|\omega)$. Dans ce cas,

$$\hat{P}(\mathbf{x}_0|\omega) \approx \frac{\frac{t}{n}}{V(\mathcal{D}(\mathbf{x}_0))}$$

- Les variantes de cette méthode portent sur la définition du voisinage $D(\mathbf{x}_0)$:
  - type de fonction
  - lié à n
  - lié à la notion de proximité

Choix d'un estimateur

- Bon estimateur $\rightarrow$ estimateur convergeant :

$$\lim_{\frac{t}{n} \rightarrow \infty} \hat{P}(\mathbf{x}_0|\omega) = P(\mathbf{x}_0|\omega)$$

$$\lim_{\frac{t}{n} \rightarrow \infty} \frac{\frac{t}{n}}{V(\mathcal{D}(\mathbf{x}_0))} = P(\mathbf{x}_0|\omega)$$

- L'estimateur revient à lisser la probabilité sur le voisinage de $\mathbf{x}_0$
- Pour s'approcher de la vraie valeur, il faut diminuer $\mathcal{D}(\mathbf{x}_0)$

Choix d'un estimateur

- Attention, si $\mathcal{D}(\mathbf{x}_0)$ est trop faible, de nombreux domaines ne vont pas avoir d'échantillon, donc $p(\mathbf{x}|\omega_i) = 0$.
- Il est donc nécessaire de lier le nombre d'échantillons et la taille du domaine. (n et $\mathcal{D}(\mathbf{x}_0)$, noté par la suite $\mathcal{D}_n(\mathbf{x}_0)$) :

$$\hat{P}(\mathbf{x}_0|\omega) = \frac{\frac{t_n}{n}}{V_n(\mathcal{D}(\mathbf{x}_0))}$$

Choix d'un estimateur

On en déduit 3 conditions nécessaires pour l'estimateur :

$$\hat{p}_n(\mathbf{x}_0|\omega) \rightarrow p(\mathbf{x}_0|\omega) \text{ si :}$$

1. $\lim_{n\rightarrow\infty} V(\mathcal{D}_n(\mathbf{x}_0)) = 0$
2. $\lim_{n\rightarrow\infty} t_n = +\infty$
3. $\lim_{n\rightarrow\infty} \frac{t_n}{n} = 0$

Choix d'un estimateur

Deux types de méthodes sont alors envisagées :

1. Lier $V(\mathcal{D}_n(\mathbf{x}_0))$ à n : il s'agit de la méthode du noyau (fenêtres de Parzen)
2. fixer t_n en fonction de n et faire croître $V(\mathcal{D}_n(\mathbf{x}_0))$ jusqu'à qu'il contienne t_n échantillons : il s'agit de la méthode des t_n plus proches voisins

Méthode du noyau

Méthode des fenêtres de Parzen

- Point clef : choix de la fonction de voisinage $\mathcal{D}_n(\mathbf{x}_0)$
- exemple avec un hypercube de coté h_n :

$$V(\mathcal{D}_n(\mathbf{x}_0)) = (h_n)^d$$

- on travaille dans $\boldsymbol{R}^d$
- On définit une fonction $\varphi(\mathbf{u})$, égale à 1 dans l'hypercube de coté 1 centré à l'origine (dans $\boldsymbol{R}^d$) :

Méthode du noyau

$$\begin{cases} \varphi(\mathbf{u}) = 1 \text{ si } |u_j| \leq 0.5 \; j = 1, .., d \\ \varphi(\mathbf{u}) = 0 \text{ sinon} \end{cases}$$

Si $\mathcal{D}_n(\mathbf{x}_0)$ est un hypercube de coté h_n, alors :

$$\mathbf{x} \in \mathcal{D}_n(\mathbf{x}_0) \Leftrightarrow \varphi\left(\frac{\mathbf{x}_0 - \mathbf{x}}{h_n}\right) = 1$$

Méthode du noyau

Le nombre d'échantillons t_n se trouvant dans le domaine $\mathcal{D}_n(\mathbf{x}_0)$ s'obtient par la formule :

$$t_n = \sum_{i=1}^{n} \varphi \left(\frac{\mathbf{x}_0 - \mathbf{x}_i}{h_n} \right)$$

où $\mathbf{x}_i$ est l'échantillon i.

$$\hat{P}_n(\mathbf{x}_0|\omega) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{V(\mathcal{D}_n(\mathbf{x}_0))} \varphi \left(\frac{\mathbf{x}_0 - \mathbf{x}_i}{h_n} \right)$$

La fonction φ s'appelle **Noyau de l'estimateur**

Choix du noyau

- le choix du noyau doit respecter une condition de normalité :

$$\begin{cases} \varphi(\mathbf{x}) \geq 0 \ \forall \mathbf{x} \in \boldsymbol{R}^d \\ \int_{\boldsymbol{R}^d} \varphi(\mathbf{x}) d\mathbf{x} = 1 \end{cases}$$

- Exemples de noyaux
  - noyau cubique,
  - noyau triangulaire,
  - noyau normal,
  - noyau exponentiel,
  - ...

Méthode des t_n p.p.v

Méthode des t_n plus proches voisins

- idée : adaptation de la taille du voisinage de $\mathbf{x}_0$
- On veut englober un nombre fixé (t_n) d'echantillons parmi les n échantillons définissant la classe afin d'avoir suffisamment déchantillons pour contribuer à la définition de $\hat{p}_n(\mathbf{x}_0)$.
- Soit $\mathcal{D}_r(\mathbf{x}_0)$ un domaine de volume unité centré en $\mathbf{x}_0$:

$$V[\mathcal{D}_r(\mathbf{x}_0)] = 1$$

- Soit $\mathcal{D}(\mathbf{x}_0, \alpha)$ le domaine homothétique de $\mathcal{D}_r(\mathbf{x}_0)$ de centre $\mathbf{x}_0$ et de rapport d'homothétie α.

Méthode des t_n p.p.v

- Alors :

$$V[\mathcal{D}(\mathbf{x}_0, \alpha)] = \alpha^d$$

- La méthode des t_n plus proches voisins consiste à faire croître α jusqu'à ce que $\mathcal{D}(\mathbf{x}_0, \alpha)$ englobe t_n échantillons.
- L'estimateur de la densité de probabilité est alors donné par :

$$\hat{p}_n(\mathbf{x}_0) = \frac{\frac{t_n}{n}}{V[\mathcal{D}(\mathbf{x}_0, \alpha)]}$$

Méthode des t_n p.p.v

- Cet estimateur converge vers la vraie valeur de $p_n(\mathbf{x}_0)$ si, par exemple :

$$t_n = t_0 * \sqrt{n} \text{ ou } t_n = t_0 * logn$$

- avec t_0 : paramètre à ajuster !!

Estimation des probabilités à postériori

- Dans les méthodes non paramètriques, on cherche à estimer directement les $p(\mathbf{x}|\omega_i)$ à partir des échantillons.
- Les échantillons sont donc suffisants pour estimer :
  - les probabilités à priori $P(\omega_i)$,
  - les lois de densité de probabilité $p(\mathbf{x}|\omega_i)$.
- **Démonstration :**
- On suppose que l'on dispose de n échantillons représentant les différentes classes possibles. De plus, on suppose que K_i échantillons représentent la classe ω_i, donc :

$$n = \sum_{i=1}^{c} K_i$$

Estimation des probabilités à postériori

- On définit un volume V, autour de $\mathbf{x}$, contenant k échantillons, dont k_i sont étiquetés ω_i. On estime alors que :

$$P(\omega_i) = \frac{K_i}{n}$$

$$p(\mathbf{x}|\omega_i) = \frac{\frac{k_i}{K_i}}{V}$$

d'où :

$$p(\mathbf{x}|\omega_i).P(\omega_i) = \frac{\frac{k_i}{n}}{V}$$

Estimation des probabilités à postériori

- En appliquant la règle de Bayes :

$$p(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i).P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)} \approx \frac{\frac{\frac{k_i}{n}}{V}}{\sum_{j=1}^{c} \frac{\frac{k_i}{n}}{V}}$$

- Finalement la densité de probabilités est donnée par le rapport du nombre d'échantillons de la classe ω_i présents dans le volume par le nombre total d'échantillons présents dans le volume

$$p(\omega_i|\mathbf{x}) \approx \frac{k_i}{k}$$

Méthodes basée sur les échantillons

- Les méthodes précédentes supposent le calcul de $p(\mathbf{x}|\omega_i)$, puis l'application de la règle de Bayes.
- Il existe des méthodes qui utilisent directement les échantillons en tant que tels pour définir les classes et la méthode de décision associée :
  - Régle de décision du plus proche voisin,
  - Régle de décision des k plus proches voisins

Régle du plus proche voisin

- la méthode suppose que l'on dispose d'une mesure de distance dans l'espace des paramètres.
- la forme inconnue est alors classée dans la classe de l'échantillon le plus proche.

Régle des k plus proches voisin

- Extension de la méthode précédente,
- Soit une forme inconnue **x** à classer :
- On mesure la distance de **x** à tous les échantillons
- On sélectionne les q plus proches échantillons
- On affecte **x** à la classe majoritaire parmi les q plus proches échantillons