ARF

Apprentissage bayésien
Estimation de densité

Cours 2

Nicolas Baskiotis
nicolas.baskiotis@lip6.fr

Master 1 DAC

équipe MLIA, Laboratoire d'Informatique de Paris 6 (LIP6)
Université Pierre et Marie Curie (UPMC)

S2 (2014-2015)

Plan

Notions et notations

Rappel

- Univers Ω, Espace probabiliste $(\Omega, \mathcal{A}, P)$
- Variable aléatoire (v.a.) : $X : \Omega \to \mathbb{R}$
  notation : $P(X = 1) = 0.3$
- Fonction de densité $p(x)$ (ou de masse $P(x)$), fonction de répartition $F(x)$
  $p(x) \geq 0,\ \int p(x)dx = 1, F(b) - F(a) = p(a \leq X \leq b) = \int_a^b p(x)dx$

- Espérance, variance :
  $\mathbb{E}[X] = \int xp(x)dx, Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$
- Densité jointe, indépendance, conditionnement, marginalisation :
  Soit X, Y deux v.a. et leur densité jointe : $p(x, y)$
  - trouver $p(x) \to$ marginalisation : $p(x) = \int_y dp(x, y)$
  - indépendance : $p(x, y) = p(x)p(y)$
  - conditionnement : $p(x|y) = p(x, y)/p(y)$

  $\Rightarrow$ Bayes : $p(y|x) = p(x|y)p(y)/p(x)$

Quelques lois et bornes de convergence

Loi faible/forte des grands nombres

Soit $X_1, \ldots, X_m$ v.a. tirer de la même loi, de même espérance μ et variance, et la moyenne empirique $\bar{X}_m = \frac{1}{m}\sum_{i=1}^{m} X_i$, alors

- $\forall \epsilon > 0$, $\lim_{m\to\infty} Pr(|\bar{X}_m - \mu| \leq \epsilon) = 1$ (faible)
- $Pr(\lim_{m\to\infty} \bar{X}_m = \mu) = 1$ (forte)

Théorème central limite

X_i v.a. iid, de moyenne μ, variance σ, alors $Z_m = \frac{\bar{X}_m - \mu}{\sigma/\sqrt{(m)}} \to \mathcal{N}(0,1)$.

Bornes usuelles

- Gauss-Markov : pour $X \geq 0, \epsilon > 0$, $Pr(X \geq \epsilon) \leq \frac{\mu}{\epsilon}$
- Tchebychev : $Pr(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$

$\Rightarrow$ si $\bar{X}_m = \frac{1}{m}\sum_1^m X_i$, $\mathbb{E}(\bar{X}_m) = \mu$, $Var(\bar{X}_m) = \frac{\sigma^2}{m}$, donc $Pr(|\bar{X}_m - \mu| \geq \epsilon) \leq \frac{\sigma^2}{m\epsilon^2}$

- Hoeffding : $X_i \in [a,b]$,$Pr(|\bar{X}_m - \mu| \geq \epsilon) \leq 2\exp\left(-\frac{2m\epsilon^2}{(b-a)^2}\right)$

Plan

Classification binaire

Rappel

- Deux classes : $\mathcal{Y} = \{y_+, y_-\}$
- une description dans $\mathcal{X}$ des exemples
- une distribution $p(X)$ des exemples
- objectif : prendre une décision sur la classe d'un exemple x donné

$\Rightarrow$ on cherche une fonction $f : \mathcal{X} \to \mathcal{Y}$ (classifieur)

- on notera souvent $\hat{y}$ la décision prise sur un exemple x, $\hat{y} = f(x)$

Comment évaluer un classifieur ?

Fonction de perte

- Notion d'erreur, de perte associée à une décision $f(x)$
- Erreur simple : à chaque fois qu'on se trompe, on compte 1

$\Rightarrow$ fonction de perte : $\ell(f(x), y) = \begin{cases} 1 & \text{si } f(x) \neq y \\ 0 & \text{sinon} \end{cases}$ *0-1 loss*

- Risque associé : $R(y_i|x) = \sum_j l(y_i, y_j) P(y_j|x) = 1 - P(y_i|x)$
- $R = \int R(f(x)|x) p(x) dx$
- Peut-on toujours avoir un risque nul ? souvent ?

Première approche

Le plus simple

Si on dispose de $P(y = y_+)$ et $P(y = y_-)$, probabilités a priori :

- elles décrivent notre connaissance générique du problème
- peuvent dépendre des situations
- on peut décider y_+ si $P(y_+) > P(y_-)$, y_- dans le cas contraire

Première approche

Le plus simple

Si on dispose de $P(y = y_+)$ et $P(y = y_-)$, probabilités a priori :

- elles décrivent notre connaissance générique du problème
- peuvent dépendre des situations
- on peut décider y_+ si $P(y_+) > P(y_-)$, y_- dans le cas contraire

Problèmes

- Toujours la même décision
- On ne tient pas compte de la description.
- Evaluation du risque : $R = min(P(y_+), P(y_-))$

Dans un monde idéal (bayésien)

Si on dispose ...

de $P(y)$ (probabilité a priori) et de $p(x|y)$:

- $p(y,x) = p(y|x)p(x) = p(x|y)p(y)$
- $p(x) = p(x|y_+)p(y_+) + p(x|y_-)p(y_-)$
- $p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{p(x|y_+)P(y_+)+p(x|y_-)P(y_-)}$

Alors

En observant x, on peut convertir $p(y)$ en *probabilité a posteriori* $p(y|x)$.

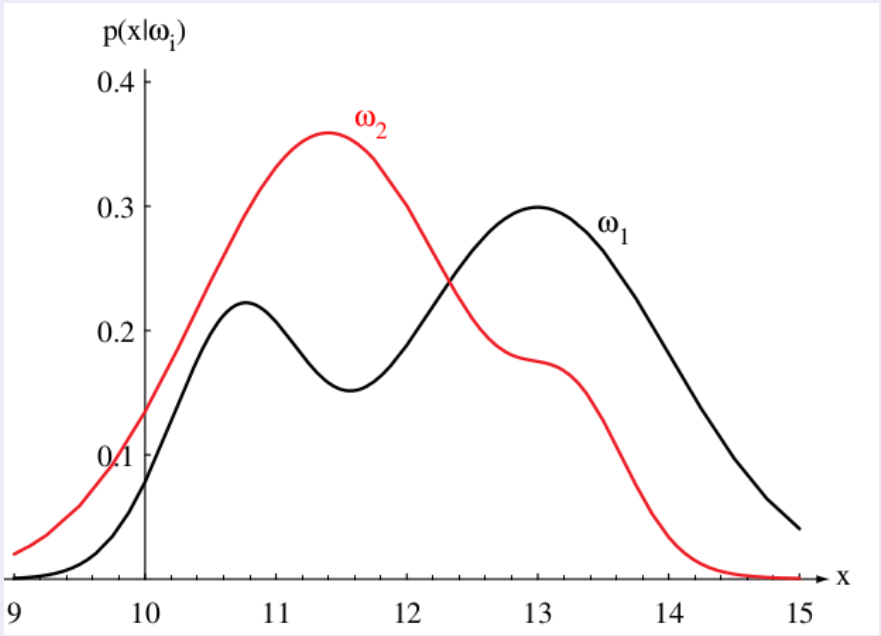
- On appelle $p(x|y)$ la vraissemblance de y par rapport à x.
- décision bayésienne : choisir y_+ si $p(y_+|x) > p(y_-|x)$, le contraire sinon

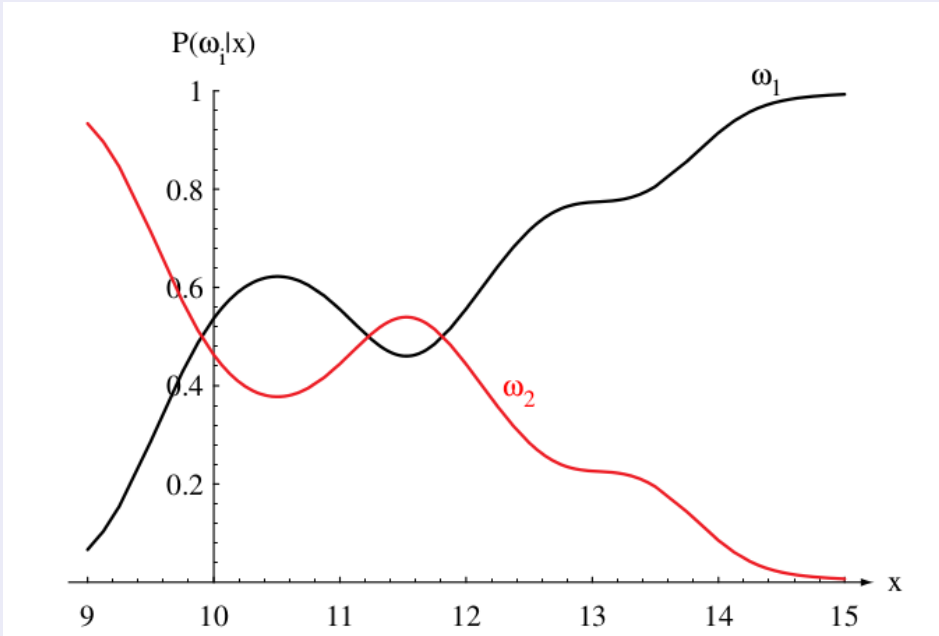
$\Rightarrow f(x) = argmax_y p(y|x)$

- $p(x)$ est-il important ?

Exemple

pour $p(y_+) = 2/3$ et $p(y_-) = 1/3$





(Duda et al. 00)

Probabilité de l'erreur

Caclul de l'erreur

- $P(erreur|x) = \begin{cases} P(y_+|x) & \text{si on décide } y_- \\ P(y_-|x) & \text{si on décide } y_+ \end{cases}$
- $P(erreur) = \int P(erreur|x)p(x)dx$
- $P(erreur|x) = min(P(y_+|x), P(y_-|x))$
- $P(erreur|x) = min(P(x|y_+)P(y_+), P(x|y_-)P(y_-))$
- Si $p(x|y_+) = p(x|y_-)$?
- Si $P(y_+) = P(y_-)$?

Risque bayésien

- $R = \int R(f(x)|x)p(x)dx$
- On peut montrer que c'est le meilleur classifieur possible (cf TD)
- alors est-ce que c'est fini ?

$\Rightarrow$ Malheuresement non, $p(x|y)$ rarement disponible . . .

Que faire ?

Apprentissage paramètrique, bayésien : estimation de $p(x, y)$

- attention ! $x \in \mathcal{X}$, de dimension d
- en vérité : $p(x|y) = p(x_1, x_2, \ldots, x_d|y)$
- dans le cas binaire ($x_i \in \{0, 1\}$), $2 * 2^d$ paramètres !!
- une solution simple : *naive bayes*, considérer chaque dimension indépendante

$\Rightarrow$ $p(x|y) = p(x_1|y)p(x_2|y)\ldots p(x_d|y)$, $2 * d$ paramètres.

- ou poser des lois a priori, estimation de paramètres des lois $\rightarrow$ estimation bayésienne, maximum de vraissemblance
- modèles graphiques, recherche d'indépendance entre dimension, ...

Ou s'en affranchir (en partie)

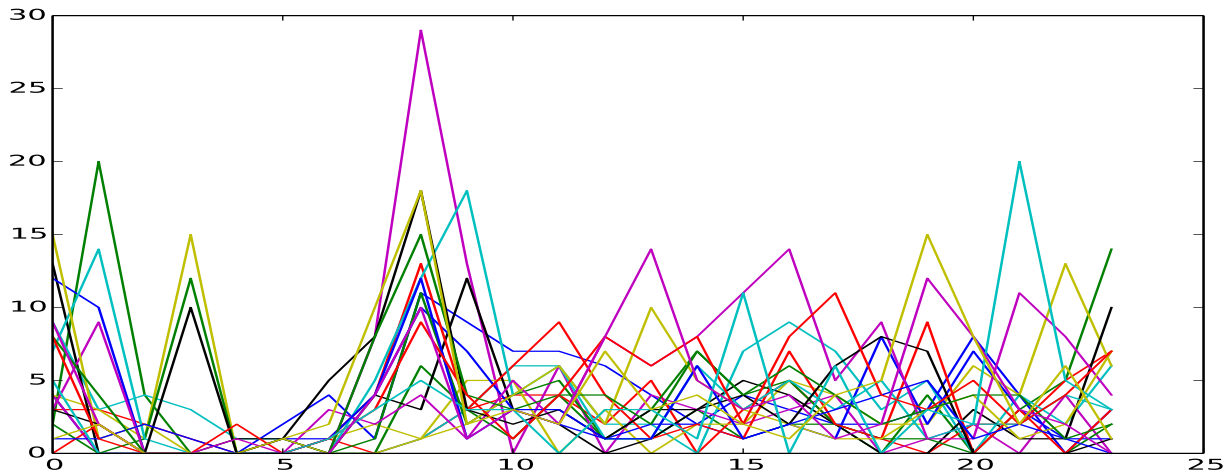
- C'est la suite de ce cours !

Plan

Estimation d'histogramme

Contexte

- Pour la classification ou pour la regression (y continu ou réel)
- $\mathcal{X}$ est de dimension d, attributs continus sont discrétisés : m_i nombre de cas pour l'attribut i
- nombre total de cases : $\prod_{i=1}^{d} m_i$
- objectif : calculer la sortie moyenne pour chaque case (pour la densité, nombre d'échantillons dans la case)
- prédiction : la sortie correspondante à la case du nouvel exemple

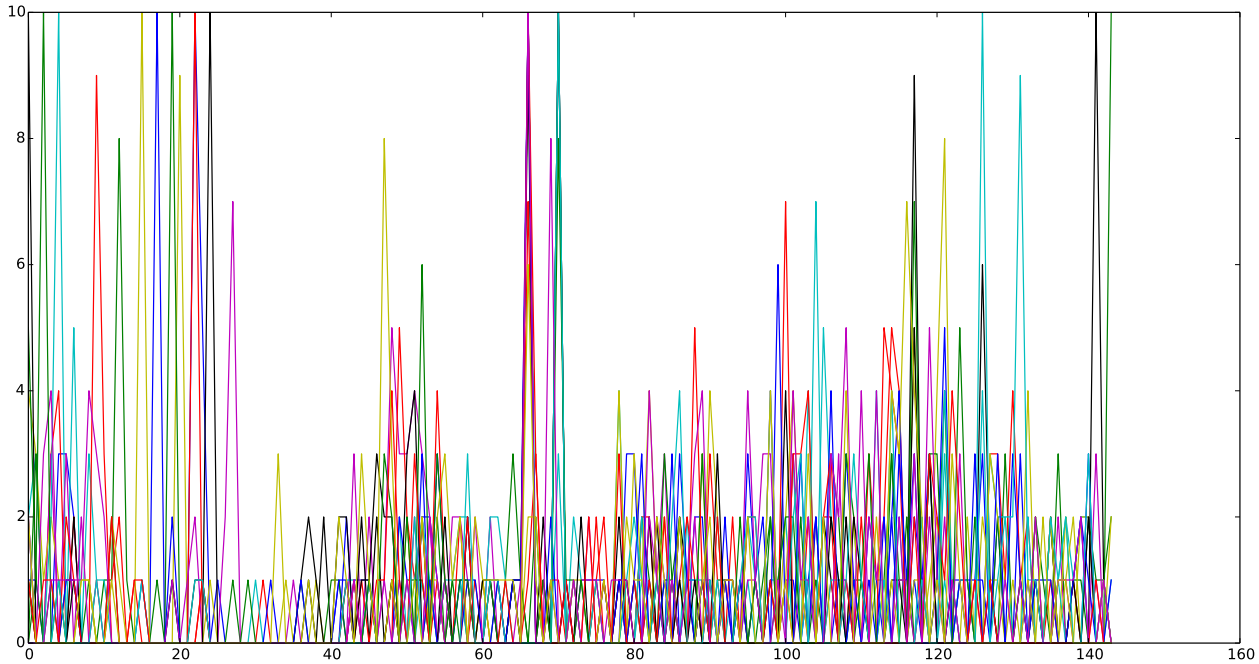


Nombre de velibs empruntés par heure sur 20 stations

Estimation d'histogramme

Problème : explosion combinatoire

- Quand le nombre de dimension augmente, le nombre de cases augmente exponentiellement
- Beaucoup de cases sans aucun échantillon
- Et même celles qui en ont, un nombre faible $\rightarrow$ peu représentatif



Nombre de velibs empruntés par 1/4 d'heure sur 20 stations

Estimation non paramètrique

Idée générale

- Soit une région $\mathcal{R}$ de l'espace,
- P la probabilité qu'un exemple x appartienne à cette région, $P = \int_{\mathcal{R}} p(x)dx$
- $x_1, \ldots, x_n$ iid, P_k la probabilité d'avoir k parmi n dans $\mathcal{R}$
- $P_k = C_n^k P^k (1-P)^{n-k}$, $\mathbb{E}[k] = nP$

Raffinement

- Si $p(x)$ est continue et que $\mathcal{R}$ est petit, que $p(x)$ ne varie presque pas dans $\mathcal{R}$

$\Rightarrow$ $\int_{\mathcal{R}} p(x)dx \simeq p(x)V$, V volume de $\mathcal{R}$

$\Rightarrow$ $p(x) \simeq \frac{k/n}{V}$

- avec $\{\mathcal{R}_1, \mathcal{R}_2, \ldots\}$ des régions pour $1, 2, \ldots$ échantillons, k_n le nombre d'échantillons dans $\mathcal{R}_n$ et V_n le volume, on a $p_n(x) = \frac{k_n/n}{V_n}$

Fenêtre de Parzen

Principe

- $\mathcal{R}_n$ est un hypercube, chaque côté de longueur h_n
- $V_n = h_n^d$
- $\phi(x) = \begin{cases} 1 & \text{si } |x^i| \leq 1/2 \\ 0 & \text{sinon} \end{cases}$
- ϕ définie un hypercube unitaire centré à l'origine.
- $\phi(\frac{x-x'}{h_n}) = 1$ ssi x' est dans l'hypercube de volume V_n centré en x.

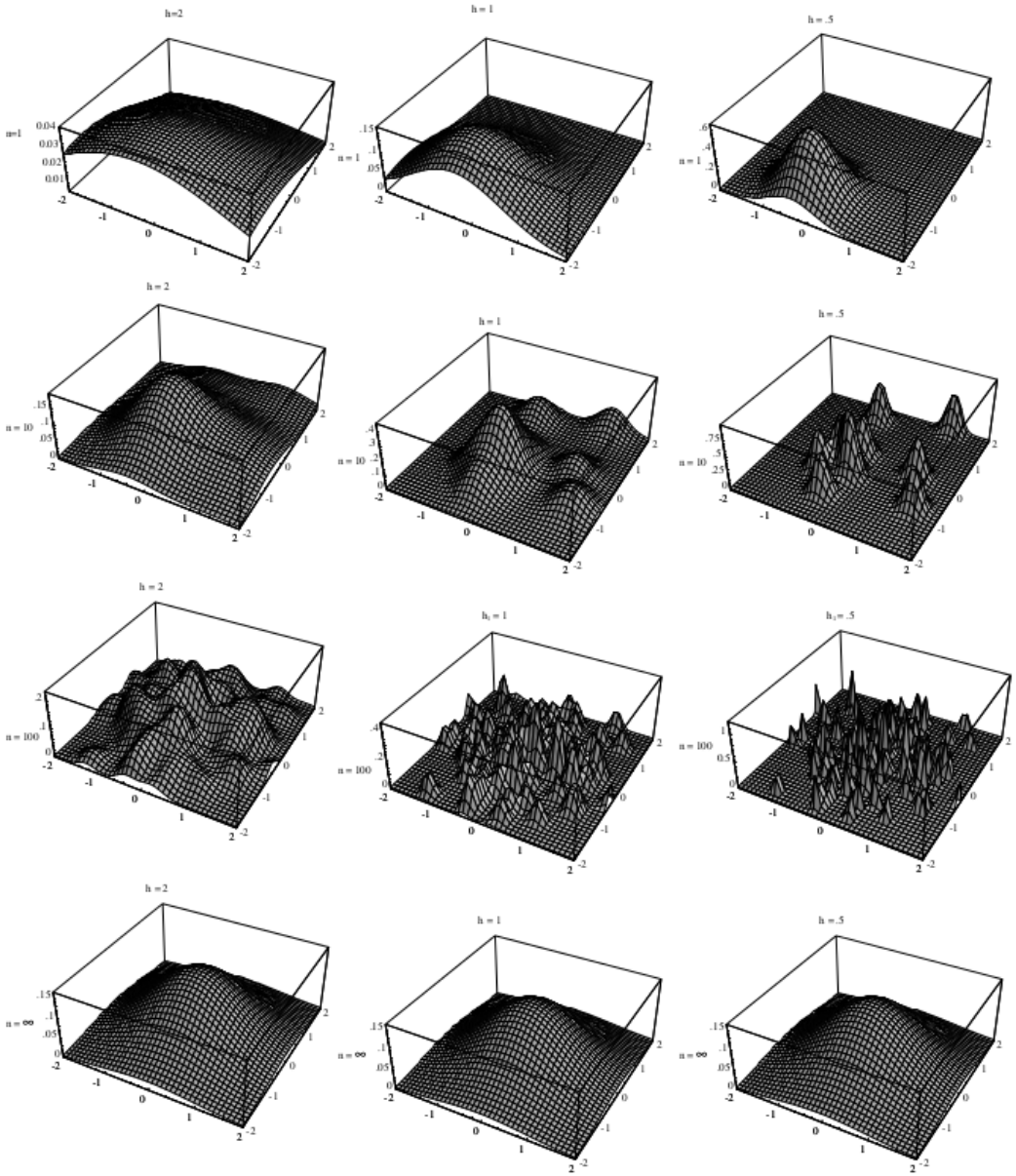
Conséquence

- $k_n = \sum_{i=1}^{n} \phi((x - x_i)/h_n)$
- $p_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{V_n} \phi(\frac{x-x_i}{h_n})$
- simplification : $\delta_n(x) = \frac{1}{V_n} \phi(x/h_n) \rightarrow p_n(x) = \frac{1}{n} \sum_{i=1}^{n} \delta_n(x - x_i)$

Discussion

Effet de h_n

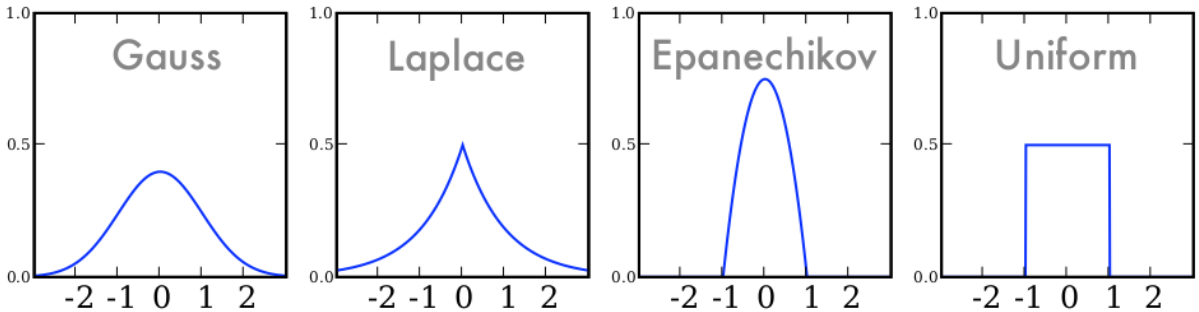
- h_n grand
  $\rightarrow \delta_n$ peu sensible, paysage homogène
- h_n petit
  $\rightarrow \delta_n$ tend vers un pic de Dirac.
- compromis entre petite résolution et grande variabilité



Discussion

Pourquoi se limiter à des hypercubes ?

- ϕ peut être plus générale (noyaux)
- conditions nécessaires :
  - $\phi(x) \leq 0$
  - $\int \phi(x)dx = 1$



$(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2}$ $\frac{1}{2}e^{-|x|}$ $\frac{3}{4}\max(0, 1-x^2)$ $\frac{1}{2}\chi_{[-1,1]}(x)$

Estimateur de Watson-Nadaraya

De la densité à la classification

- Classification binaire : $p(x|y_+)$ et $p(x|y_-)$, et pour la suite $y_+ = 1, y_- = -1$
- $p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{\frac{1}{n_y}\sum_{y_i=y}\phi(x-x_i)\frac{n_y}{n}}{\frac{1}{n}\sum_i \phi(x-x_i)}$
- $p(y_+|x) - p(y_-|x) = \frac{\sum_j y_j \phi(x-x_j)}{\sum_i \phi(x-x_i)} = \sum_j y_j \frac{\phi(x-x_j)}{\sum_i \phi(x-x_i)}$
- directement adaptable à la regression

Plus proches voisins (k-nearest Neighbors)

Principe

- plutôt que de prendre en compte un noyau ou la distance, prendre en compte le voisinage (immédiat ou non) du point
- un paramère : k le nombre de voisins à prendre en compte
- $p(y|x) = \frac{1}{k} \sum_{j} y_j$ k- plus proches

Discussion

- Parzen : travail sur le volume, pas de controle sur le nombre de points considérés
- Knn : volume libre, mais nombre de points fixe
- dans tous les cas :
  - complexité grande des algorithmes (possible d'utiliser des arbres de partitionement (KD-tree) et autres heuristiques pour accélerer)
  - des paramètres à choisir . . .
- Comment choisir les paramètres ?

Plan

Sélection de modèles

Problèmatique

- Très souvent, il faut fixer des paramètres aux algorithmes d'apprentissage
  - profondeur de l'arbre
  - nombre de voisins dans les k-nn
  - longueur de l'hypercube, paramètre des noyaux dans les fenêtres de Parzen
- quels effets ont ses paramètres ?
  - ils déterminent généralement le pouvoir expressif du modèle
  - combien le modèle va coller aux données et faire peu d'erreurs sur les données d'apprentissage
  - ou au contraire faire plus d'erreurs mais généraliser

$\Rightarrow$ ils calibrent le *sur-apprentissage* ou le *sous-apprentissage*

- compromis entre l'apprentissage par cœur et l'apprentissage uniforme

Sélection de modèles empirique

Choisir le paramètrage en fonction des données

- évaluer les différents paramètrages en fonction de l'évaluation des modèles
- utiliser des données pour évaluer les modèles
- Mais pas n'importe lesquelles !!

Evaluer un modèle

- Problème : il ne faut jamais évaluer un modèle sur l'ensemble d'apprentissage (pourquoi ?)
- vocabulaire :
  - ensemble d'apprentissage
  - ensemble de calibration (optionel, dépend des algos)
  - ensemble de test
- Mais comment éviter un biais lors de la construction de ses ensembles ?

Validation croisée

Principe

- Partitionner les données en k sous-ensembles
- apprendre le modèle sur $k-1$ sous-ensembles
- évaluer le modèle sur le dernier sous-ensemble
- répeter l'opération k-fois, sur toutes les combinaisons possibles, en guardant les sous-ensembles fixes.
- la performance moyenne est la moyenne des k évaluations : $\frac{1}{k}\sum_{i=1}^{k} \ell(p(X_i|X/X_i))$.

Discussion

- Cas particulier : si $k = n-1 \rightarrow$ *leave-one-out*
- Vaut-il mieux k grand ou petit ?
- Si on dispose de beaucoup (beaucoup) de données, est-ce toujours intéressant ? Et dans le cas de peu (très peu) de données ?
- Inconvénients ?