Apprentissage bayésien
Estimation de densité

Cours 2
ARF   Master DAC

Nicolas Baskiotis
nicolas.baskiotis@lip6.fr
http://webia.lip6.fr/~baskiotisn

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

S2 (2016-2017)

Plan

1. Rappel MAPSI/Probabilités
2. Classification bayesienne
3. Estimation de densité
4. Sélection de modèles

Notions et notations

Rappel

- Univers Ω, Espace probabiliste $(\Omega, \mathcal{A}, P)$
- Variable aléatoire réelle (v.a.r.) : $X : \Omega \rightarrow \mathbb{R}$
  notation : $P(X = 1) = 0.3$, loi de X (ou mesure de probabilité) : P_X,
  fonction de répartition F_X : $F_X(b) = P_X(X < b)$
- Dans le cas continue, fonction de densité p_X :
  $p_X(x) \geq 0,\ \int_x p_X(x)dx = 1, F_X(b) - F_X(a) = p_X(a \leq X \leq b) = \int_a^b p_X(x)dx$
  abus de notation : $p_X \ \rightarrow p$
- Espérance, variance :
  $\mathbb{E}[X] = \int_x x p_X(x)dx \ \ Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$
- Densité jointe, indépendance, conditionnement, marginalisation :
  Soit X, Y deux v.a. et leur densité jointe : $p_{X,Y}(x, y)$
  - trouver $p_X \rightarrow$ marginalisation : $p_X(x) = \int_y p_{X,Y}(x, y)dy$
  - indépendance : $p_{X,Y}(x, y) = p_X(x)p_Y(y)$
  - 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 $X_1, \dots, X_m$ 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

Formalisation

- Deux classes : $\mathcal{Y} = \{y_+, y_-\}$
- un ensemble $\mathcal{X} \subseteq \mathcal{R}^d$ de représentation des exemples (d la dimension)
- un exemple : $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathcal{X}$
- objectif : prendre une décision sur la classe d'un exemple $\mathbf{x} \in \mathcal{X}$

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

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

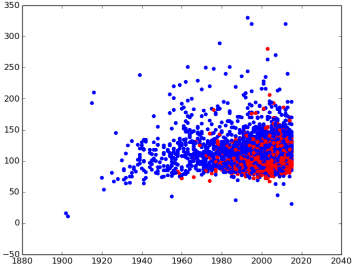
Films et avis

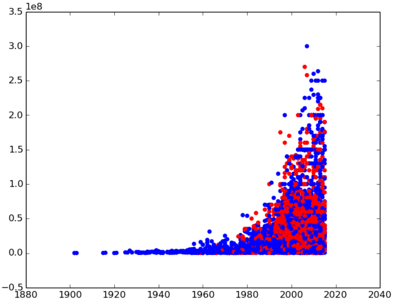
- Deux classes : j'aime (y_+) et je n'aime pas (y_-)
- Un film décrit par : (année, budget, durée,nationalité) (4 dimensions, $\mathcal{X} = \mathcal{R}^4$)
- Une fonction de prédiction : $f(\mathbf{x}) = y_+$ si $x_1 \geq 2000$ sinon y_-

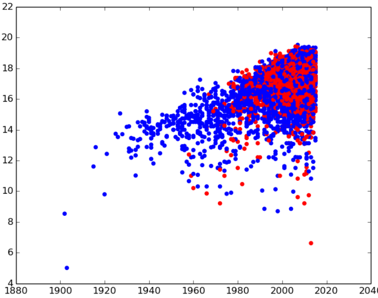
Classification binaire

Sur la base imdb . . . :

- Année vs Durée
- Année vs Budget







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
- Quelle est le risque de se tromper ?

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
- Quelle est le risque de se tromper ?

Problèmes

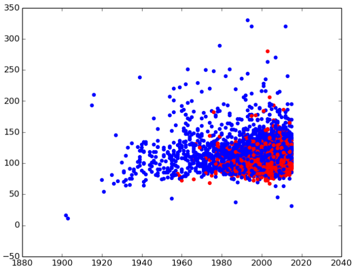
- Toujours la même décision
- On ne tient pas compte de la description $\mathbf{x} \in \mathcal{X}$.
- Evaluation du risque : $R = min(P(y_+), P(y_-))$

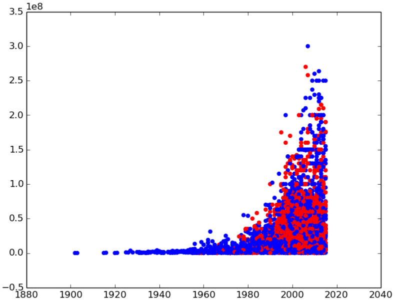
Comment faire mieux ?

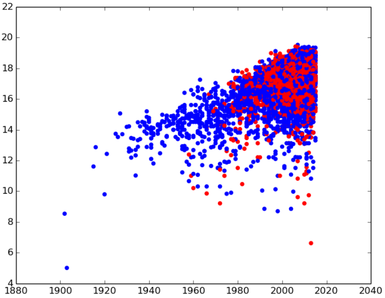
Classification binaire

Sur la base imdb . . . :

- Année vs Durée
- Année vs Budget







Dans un monde idéal (bayésien)

Si on dispose ...

de $P(y)$ (probabilité a priori) et de $p(\mathbf{x}|y)$:

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

Alors

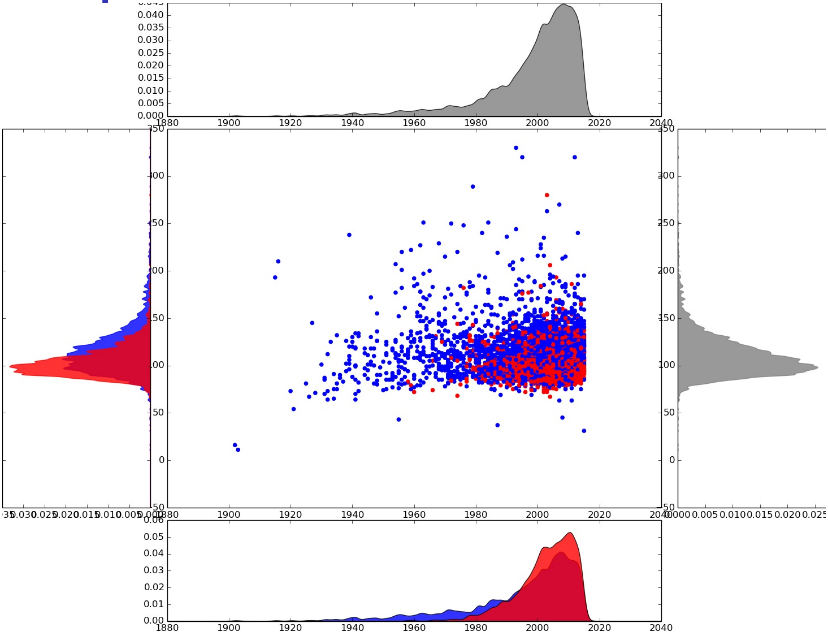
En observant $\mathbf{x}$, on peut étudier la *probabilité a posteriori* $p(y|\mathbf{x})$.

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

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

- $p(\mathbf{x})$ est-il important ?

Exemple imdb



Comment évaluer l'erreur d'un classifieur?

Fonction de perte : quantifié une erreur

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

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

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

Probabilité de l'erreur

Caclul de l'erreur

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

Risque bayésien : $\int R(f(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x}$

- Classifieur bayésien : f qui minimise le risque
- On peut montrer que c'est le meilleur classifieur possible (cf TD)
- alors est-ce que c'est fini ?

$\Rightarrow$ Malheureusement non, $p(\mathbf{x}|y)$ rarement disponible ...

Que faire ?

Apprentissage paramètrique, bayésien : estimation de $p(\mathbf{x}, y)$

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

$\Rightarrow$ $p(\mathbf{x}|y) = p(x_1|y)p(x_2|y)\ldots p(x_d|y)$, $2 * d$ 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 de densité

Contexte

- Estimer la densité d'une variable aléatoire (ou la loi jointe de plusieurs)
- à partir d'un ensemble de réalisation : un échantillon d'exemples.
- Exemple : distribution de films en fonction de l'année, du budget, . . .
- En classification : si on peut évaluer les densités de variables aléatoires
  $\Rightarrow$ classifieur bayésien pour la classification !

Applications multiples

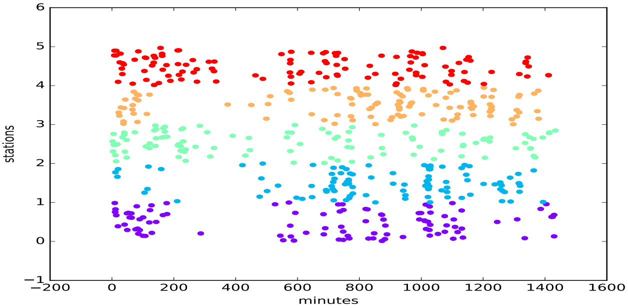
- Estimation de file d'attentes, d'occupation de lieux,
- Estimation des pics des stocks, des pics de pollution

Estimation de densité : Vélibs

Velibs pris à une station

- Données : $\{\text{time}_i, \text{station}_i\}$ $\; i \in \{1, \ldots, N\}$ les logs des vélibs
- On veut évaluer la probabilité qu'un vélib soit emprunté à une station donnée durant un intervalle de temps donné

$\Rightarrow$ estimation de densité de la variable aléatoire X_s à valeur dans $[0, 60 * 24]$ (temps exprimé en minute) pour la station s

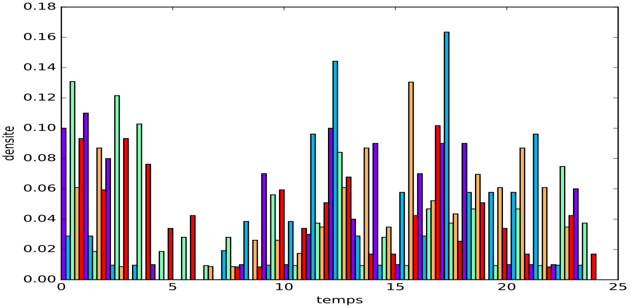


Velibs empruntés pour 5 stations sur une journée

Méthode par histogramme

Estimation par une variable aléatoire discrète

- Correspond à une discrétisation des valeurs de la v.a.
- Choix d'un pas de discrétisation : toutes les heures par exemple $\Delta = 60$
- On compte le nombre d'observations qui tombe dans chaque créneau $n_i, \; i \in \{0, 1, \ldots, 23\}$
- L'estimation est alors : $p_i = \frac{n_i}{N\Delta}$

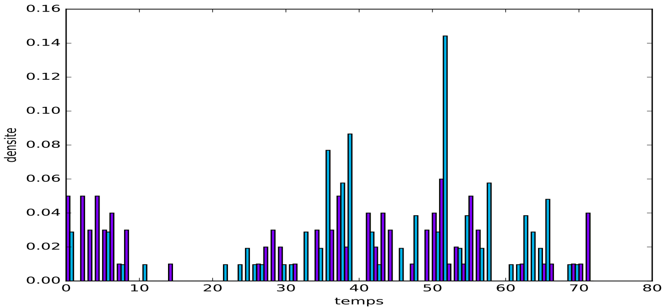


Nombre de velibs empruntés par heure sur 5 stations

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 20 minutes sur 2 stations

Estimation non paramètrique par noyaux

Idée générale

- Soit une région $\mathcal{R}$ de l'espace,
- P la probabilité qu'un exemple x appartienne à cette région, $P = \int_{\mathcal{R}} p(\mathbf{x})d\mathbf{x}$
- $x_1, \ldots, x_n$ 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 $p(\mathbf{x})$ 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

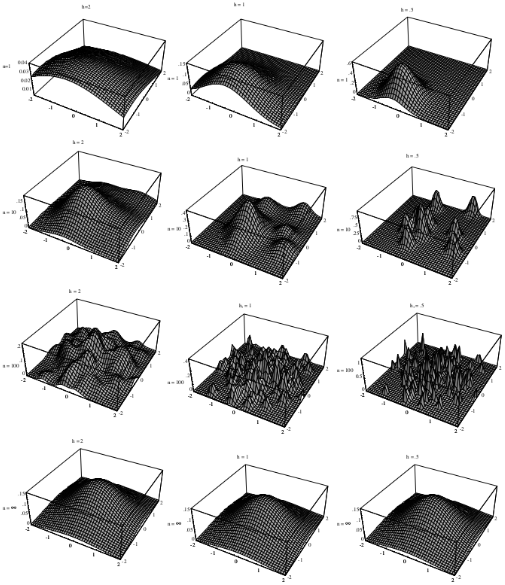
Principe

- $\mathcal{R}_n$ 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



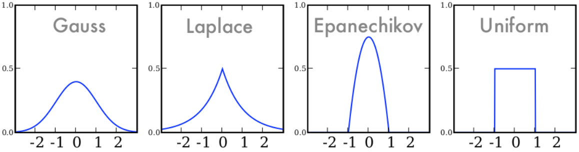
Effet de h_n

- 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

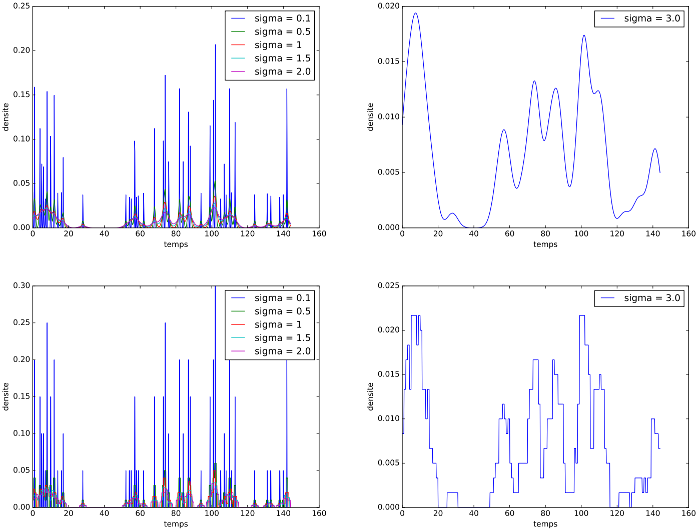
Pourquoi se limiter à des hypercubes ?

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

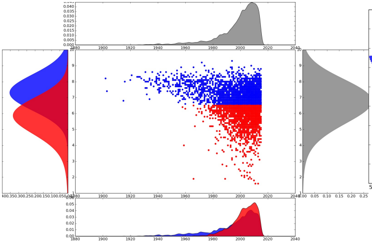


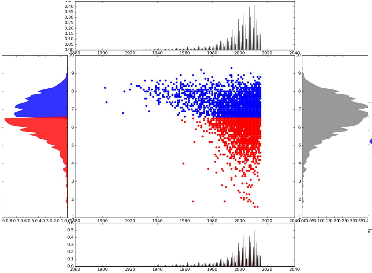
$(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2}$ $\quad \frac{1}{2}e^{-|x|}$ $\quad \frac{3}{4}\max(0, 1 - x^2)$ $\quad \frac{1}{2}\chi_{[-1,1]}(x)$

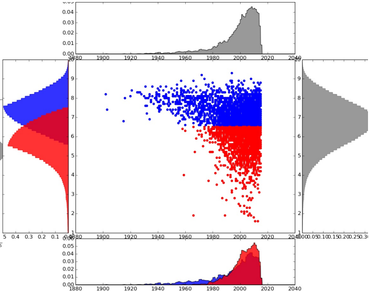
Discussion : exemples velib

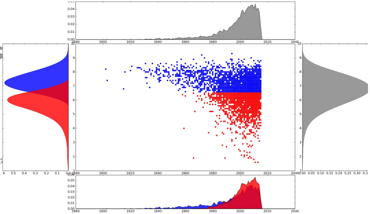


Discussion : exemples imdb









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
- $p(y|x) = \frac{1}{k}\sum_{j \ k\text{- plus proches}} y_j$

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 ?