

Apprentissage bayésien Estimation de densité

Nicolas Baskiotis

nicolas.baskiotis@sorbonne-universite.fr

équipe MLIA,
Institut des Systèmes Intelligents et de Robotique (ISIR)
Sorbonne Université

S2 (2022-2023)

Plan

- 1 Rappel MAPSI/Probabilités
- 2 Classification bayésienne
- 3 Estimation de densité par histogramme
- 4 Estimation de densité par noyaux
- 5 Estimation de densité et classification

Notions et notations

Rappel

- Univers Ω , Espace probabiliste (Ω, \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 :

$$\mathbb{E}[X] = \int_x x p_X(x)dx \quad \text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- 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 : $p(x|y) = p(x, y)/p(y)$

⇒ 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

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

Théorème central limite

X_i v.a. iid, de moyenne μ , variance σ^2 , alors $Z_m = \frac{\bar{X}_m - \mu}{\sigma / \sqrt{m}} \rightarrow \mathcal{N}(0, 1)$.

Bornes usuelles

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

1 Rappel MAPSI/Probabilités

2 Classification bayésienne

3 Estimation de densité par histogramme

4 Estimation de densité par noyaux

5 Estimation de densité et classification

Classification binaire

Formalisation

- Deux classes : $\mathcal{Y} = \{y_+, y_-\}$
 - un ensemble $\mathcal{X} \subseteq \mathbb{R}^d$ de représentation des exemples (d la dimension)
 - un exemple : $\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathcal{X}$
 - objectif : prendre une décision sur la classe d'un exemple $\mathbf{x} \in \mathcal{X}$
- ⇒ 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, $\mathcal{X} = \mathbb{R}^4$)
- Une fonction de prédiction : $f(\mathbf{x}) = y_+$ si $x_1 \geq 2000$ sinon y_-

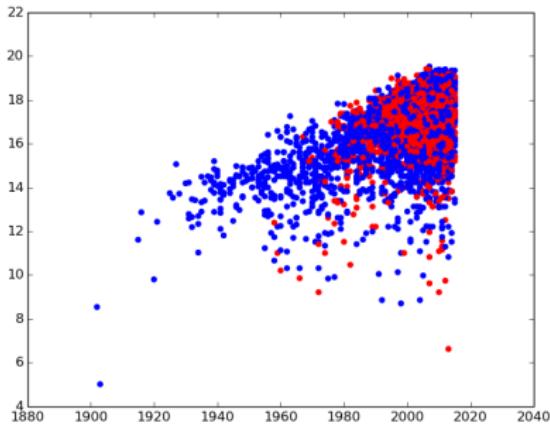
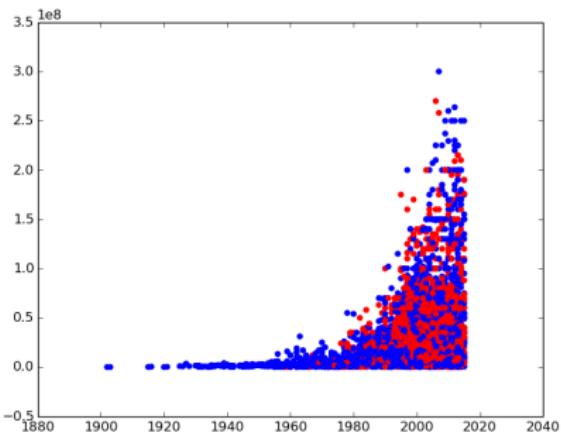
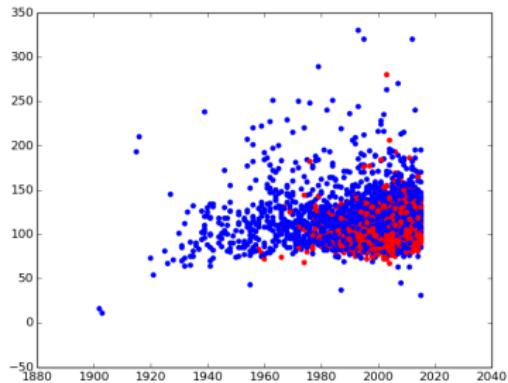
Tweets de Trump

- Deux classes : le mot *Trump* apparaît dans le tweet ou non
- Un tweet peut être décrit par son heure, le jour de la semaine, le mois, ...

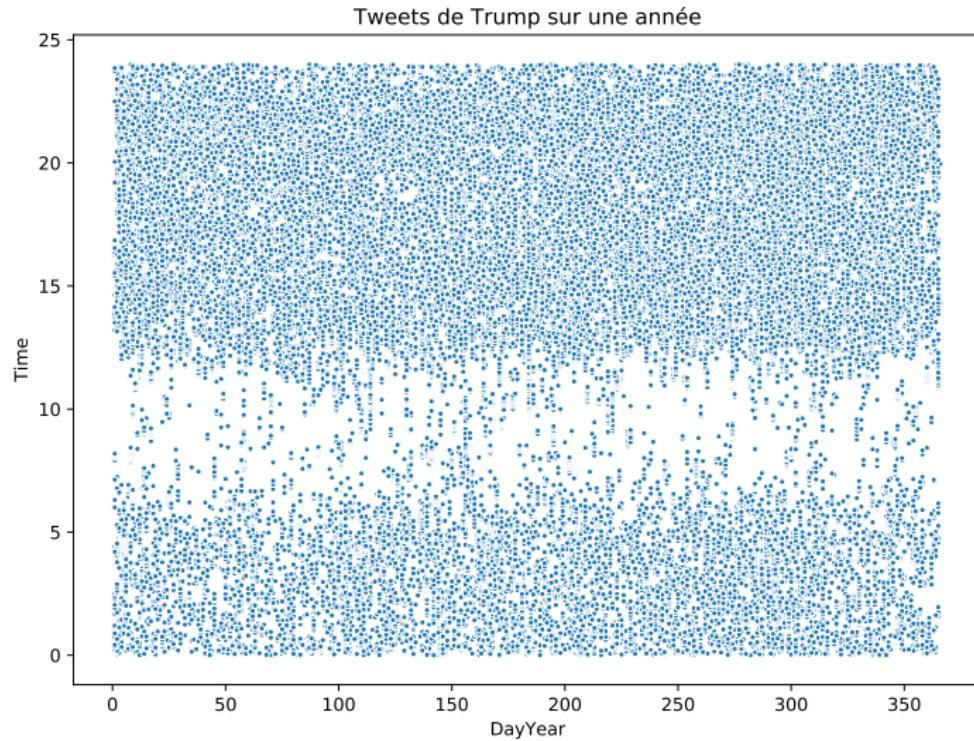
Classification binaire : IMDB

Sur la base imdb . . . :

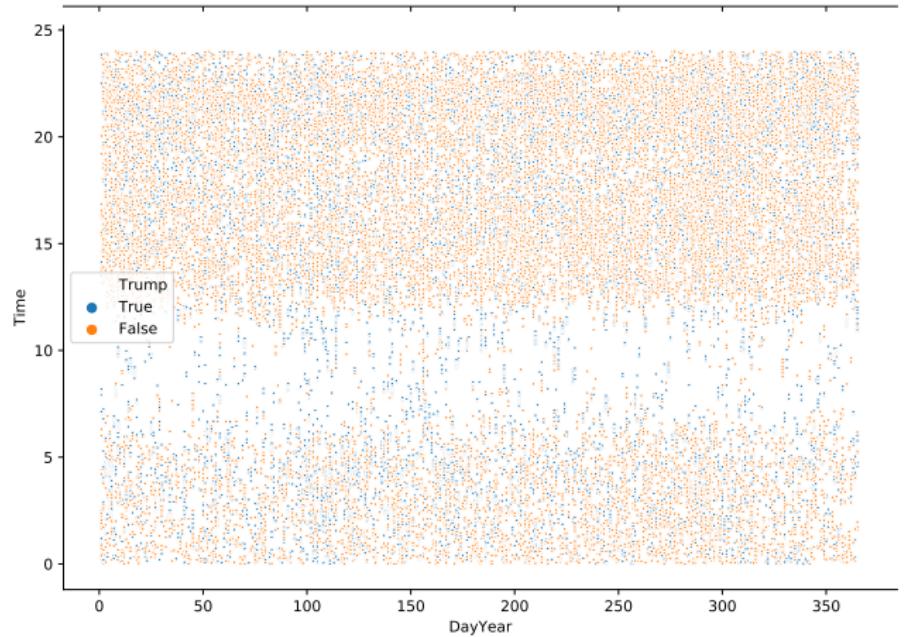
- Année vs Durée
- Année vs Budget



Classification binaire : tweets



Classification binaire : tweets



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
- Quel 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
- Quel est le risque de se tromper ?

Problèmes

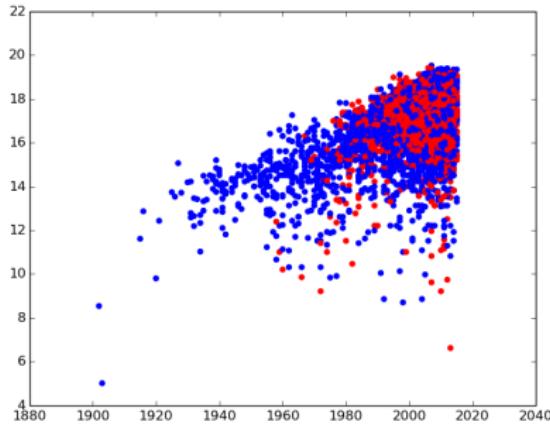
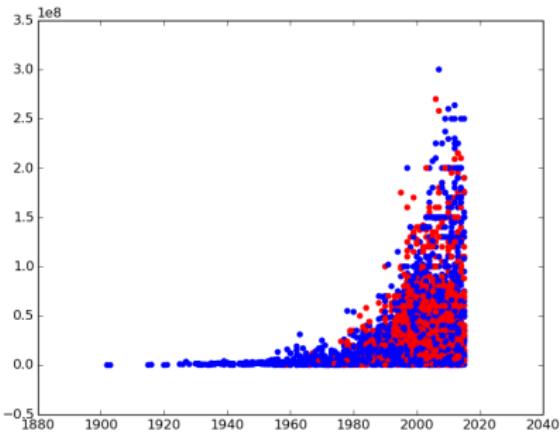
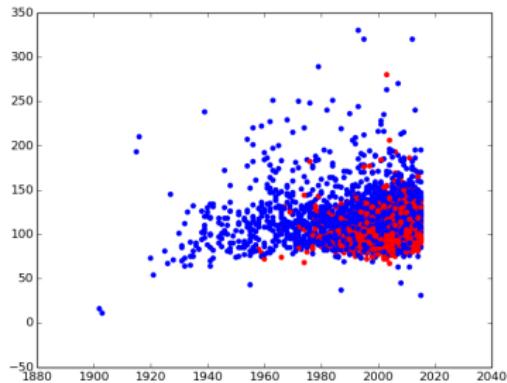
- Toujours la même décision
- On ne tient pas compte de la description $x \in \mathcal{X}$.
- Évaluation du risque : $R = \min(P(y_+), P(y_-))$

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 $p(x|y)$:

Dans un monde idéal (bayésien)

Si on dispose ...

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

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

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

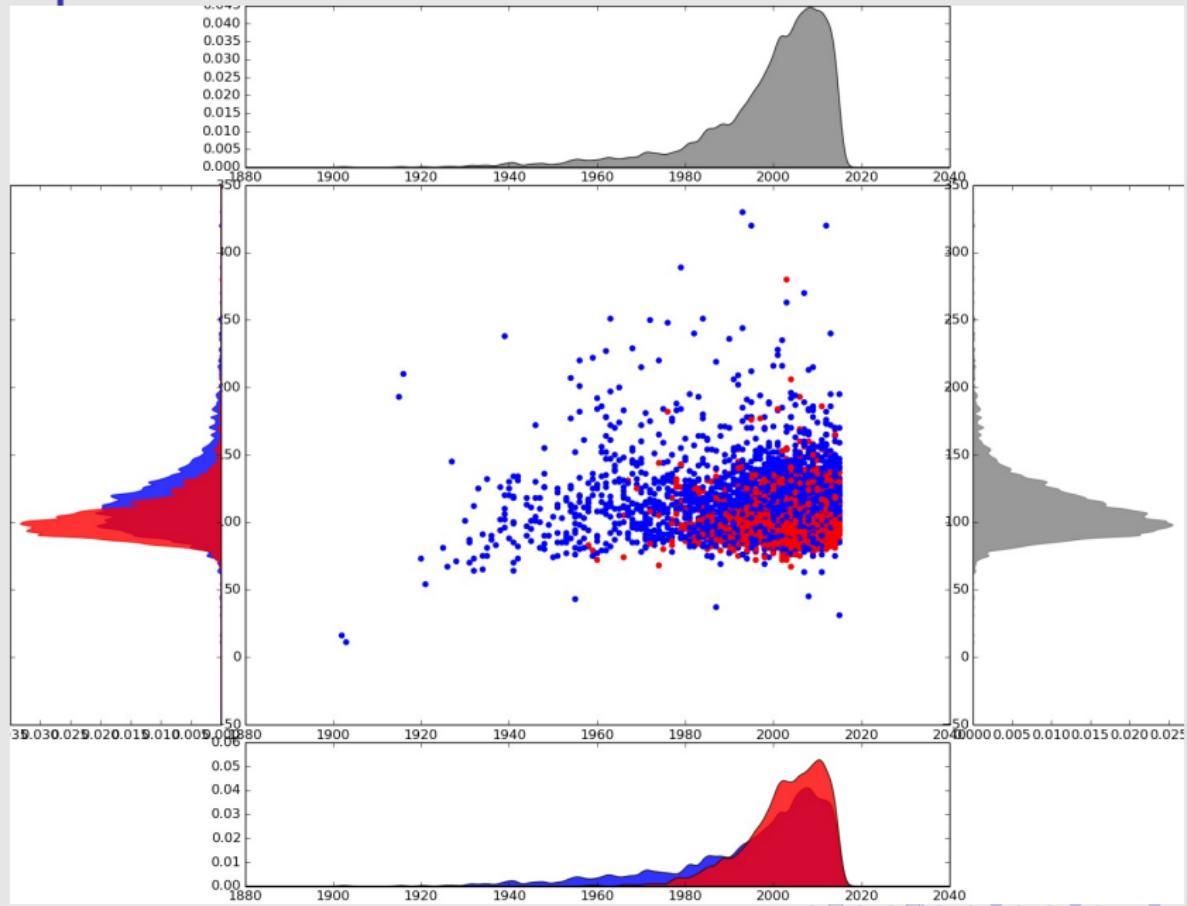
En observant \mathbf{x} , on peut étudier la *probabilité a posteriori* $p(y|\mathbf{x})$.

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

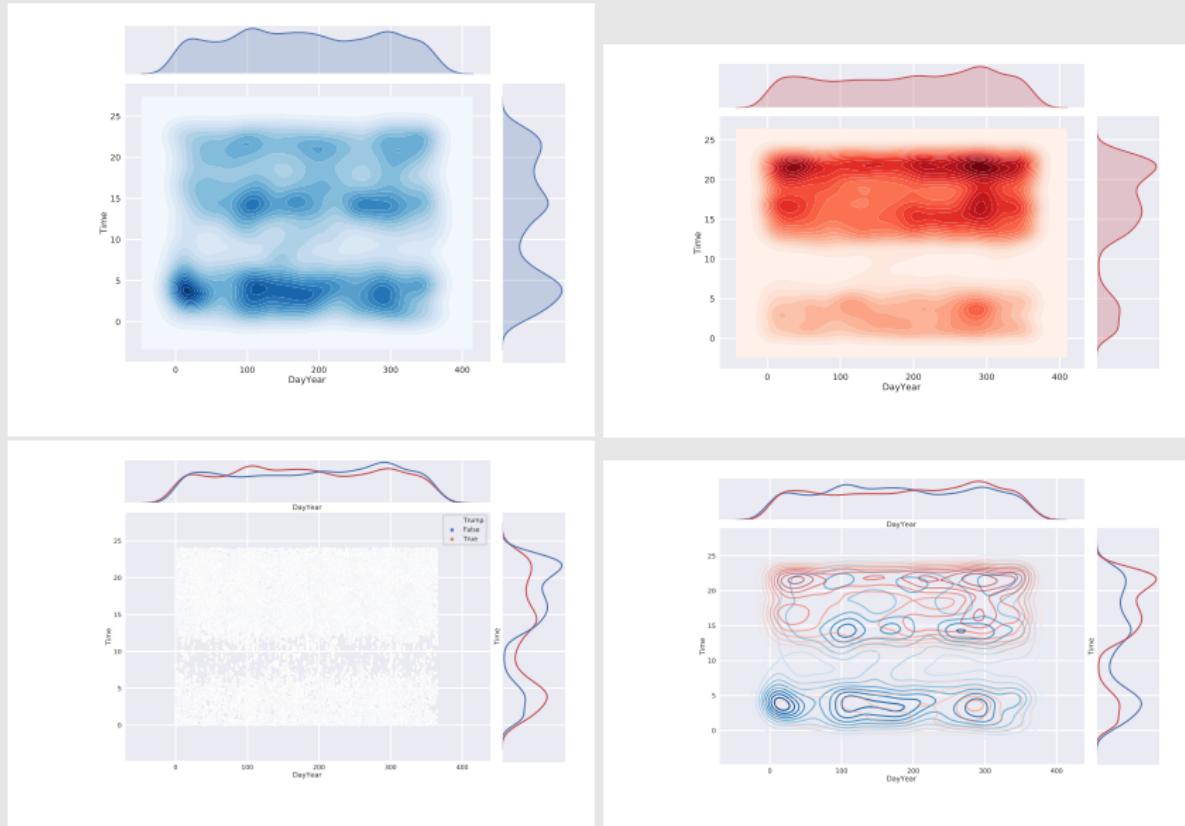
$$\Rightarrow f(\mathbf{x}) = \operatorname{argmax}_y p(y|\mathbf{x}) = \operatorname{argmax}_y \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})}$$

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

Exemple imdb



Exemple Tweets



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

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

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

Calcul 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 ?

Probabilité de l'erreur

Calcul 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 ?
- ⇒ Malheureusement non, $p(\mathbf{x}|y)$ rarement disponible ...

Que faire ?

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

- attention ! $\mathbf{x} \in \mathcal{X}$, de dimension d plutôt grand (voir très grand!)
- en vérité : $p(\mathbf{x}|y) = p(x_1, x_2, \dots, x_d|y)$
- dans le cas binaire ($x_i \in \{0, 1\}$), $2 * 2^d$ paramètres !!
- une solution simple : *naive bayes*, considérer chaque dimension indépendante
⇒ $p(\mathbf{x}|y) = p(x_1|y)p(x_2|y) \dots p(x_d|y)$, $2 * d$ paramètres.
- ou poser des lois a priori, estimation de paramètres des lois → estimation bayésienne, maximum de vraisemblance
- modèles graphiques, recherche d'indépendance entre dimension, ...

Où s'en affranchir (en partie)

- C'est la suite de ce cours !

Plan

1 Rappel MAPSI/Probabilités

2 Classification bayésienne

3 Estimation de densité par histogramme

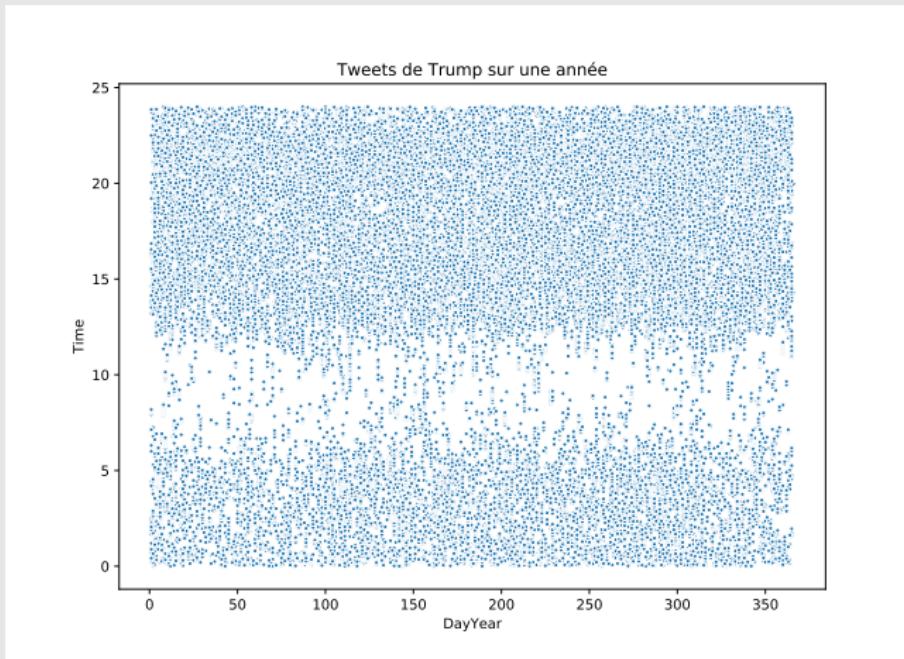
4 Estimation de densité par noyaux

5 Estimation de densité et classification

Le problème

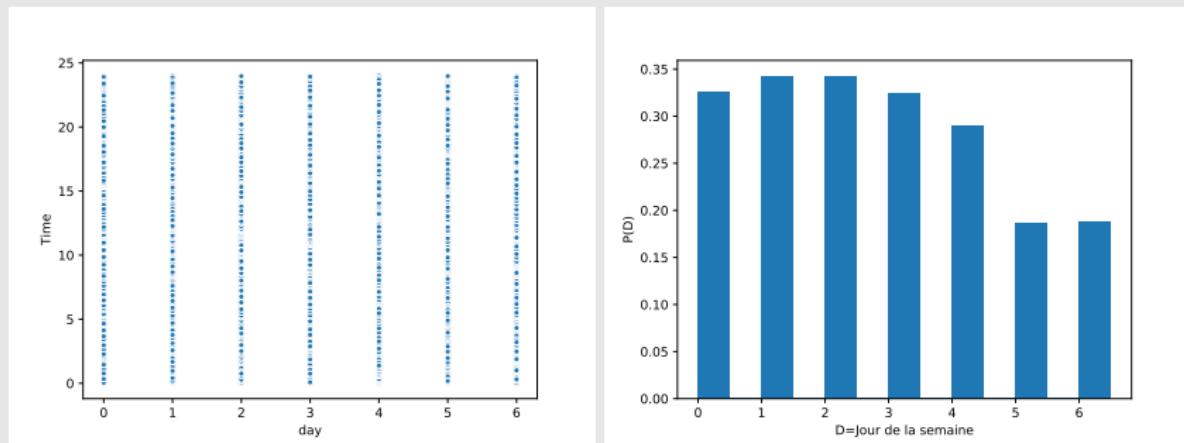
Données

Un échantillon est décrit selon des caractéristiques (dimensions) continues/catégorielles/ordinales
⇒ Quelle est la loi sous-jacente de la distribution des données ?



Lorsque les variables sont discrètes

Soit la dimension d indiquant le jour de la semaine, on observe :



Estimation par histogramme:

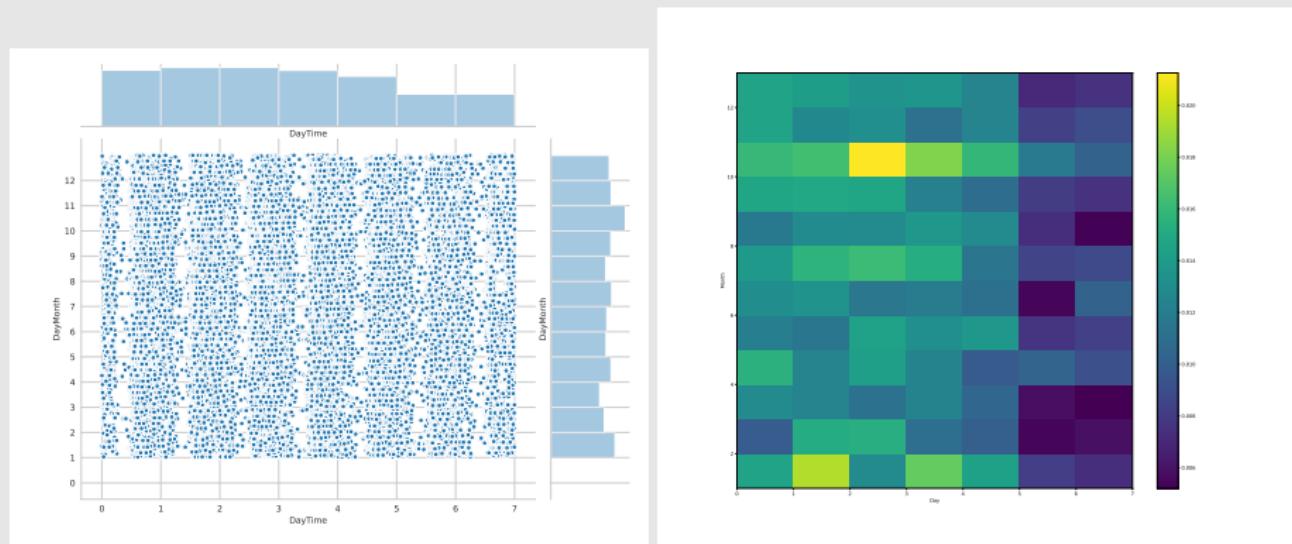
Soit $E = \{\mathbf{x}_i\}_{i=1}^N$ un échantillon de N tweets, x_i^d indique le numéro du jour de la semaine du i -ème tweet :

- $P(D = k) = \frac{|\{\mathbf{x}_i | x_i^d = k\}|}{N} = \frac{\sum_{i=1}^N \mathbf{1}_{x_i^d = k}}{N}$
- Si l'échantillonnage est i.i.d, l'estimation de la v.a. discrète converge vers la loi avec N .

Généralisation à plusieurs variables

Soit la dimension d indiquant le jour de la semaine et m le mois :

$$P(D = d, M = m) = \frac{|\{x_i | x_i^d = k, x_i^m = m\}|}{N} = \frac{\sum_{i=1}^N \mathbf{1}_{x_i^d = d, x_i^m = m}}{N}$$



Lorsque les variables sont continues

Ex : $x_i^t \in [0, 7]$ indique le moment h du i -ème tweet dans la semaine (unité d'un jour).

Estimation par une variable aléatoire discrète H

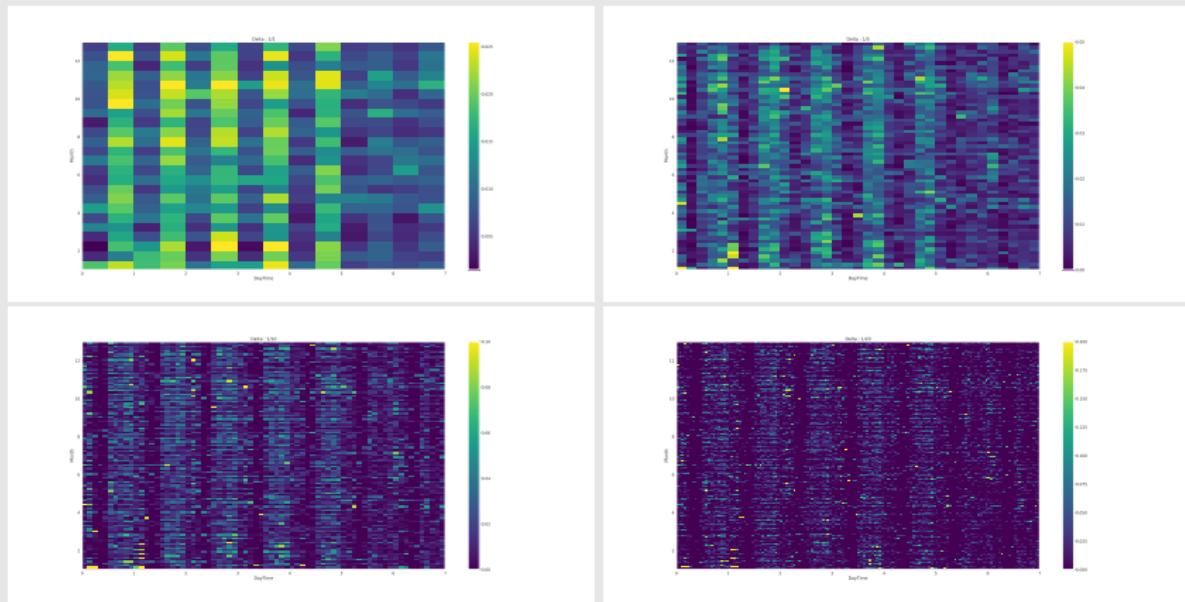
- Discréétisation des valeurs de la v.a.
- Choix d'un pas de discréétisation : 0.5 jour par exemple $\Delta = 0.5$. Cela définit $7/\Delta$ créneaux T_j .
- L'estimation discrète est alors : $P(H \in T_j) = \frac{|\{\mathbf{x}_i | x_i^t \in T_j\}|}{N}$
- Densité de probabilité associée : $p(h \in T_j) = \frac{P(H \in T_j)}{\Delta^d}$ (d la dimension, ici 1).
- Importance de la discréétisation : petit \rightarrow sur-apprentissage, trop grand \rightarrow sous-apprentissage



Limites de la méthode des histogrammes

En grande dimension d :

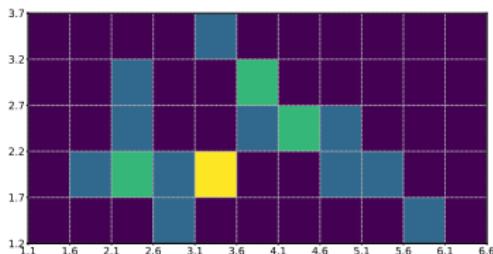
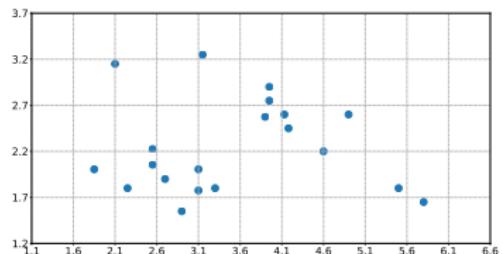
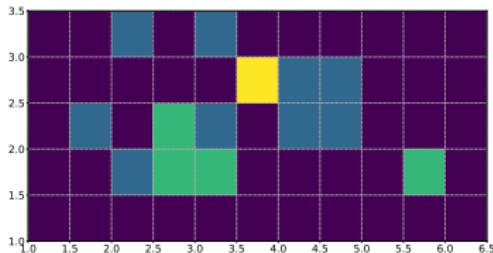
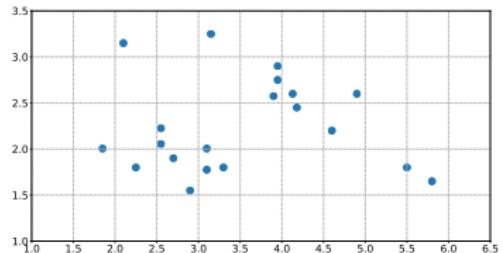
- la taille du modèle (le nombre de cases) augmente exponentiellement en $(\frac{D}{\Delta})^d$
- Beaucoup de cases sans aucun échantillon
- Celles qui en ont en ont un nombre faible → peu représentatives



Limites de la méthode des histogrammes

Effet de bords

Déplacer légèrement la discréétisation peut provoquer de grands changements d'estimation.

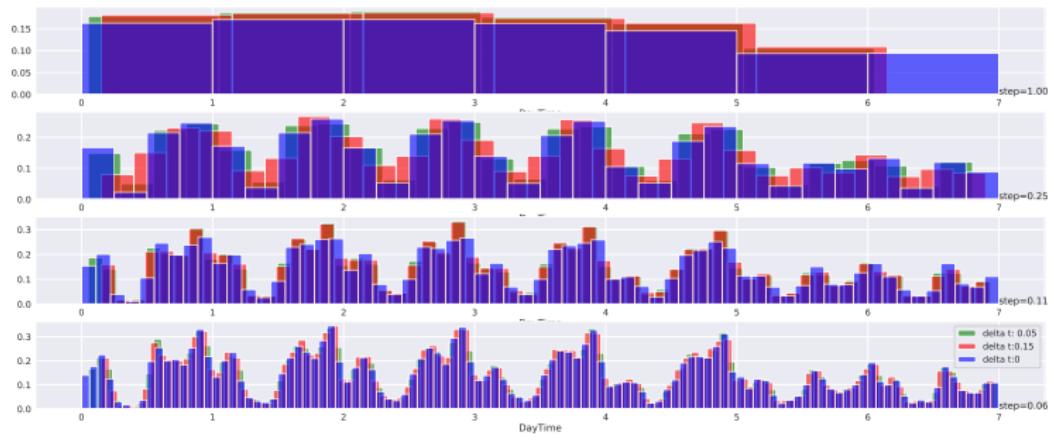


Exemple données artificielles

Limites de la méthode des histogrammes

Effet de bords

Déplacer légèrement la discréétisation peut provoquer de grands changements d'estimation.



Données Tweets

Plan

1 Rappel MAPSI/Probabilités

2 Classification bayésienne

3 Estimation de densité par histogramme

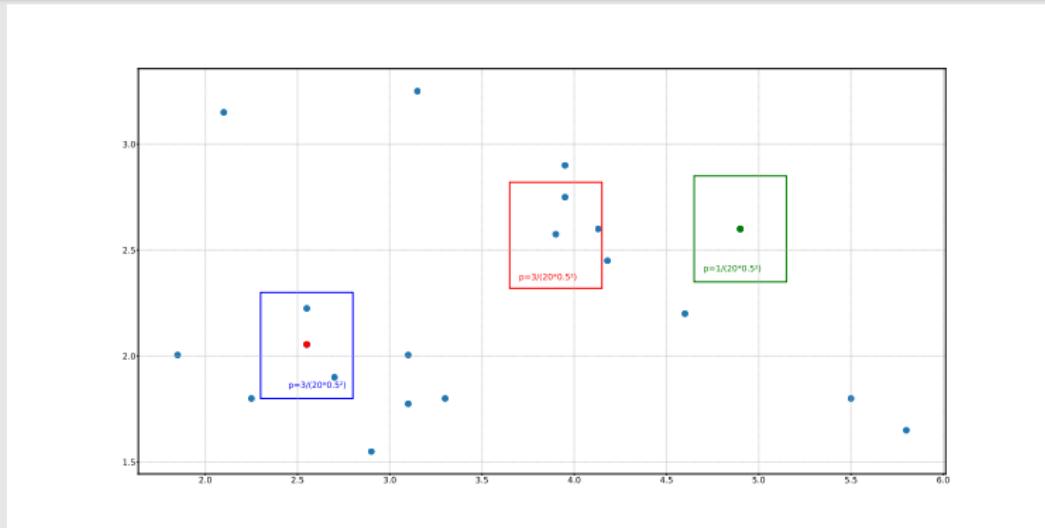
4 Estimation de densité par noyaux

5 Estimation de densité et classification

Estimation non paramétrique par noyaux

Intuition : histogramme centré au point d'intérêt

- Plutôt que de décider d'une discréttisation a priori, l'estimation est faite en centrant une fenêtre autour du point d'intérêt x_0 (dans un espace de dimension d)
 - Soit \mathcal{R} l'hypercube centré en x_0 de longueur r et $p(x_0)$ la densité à estimer
 - Hypothèse : densité constante autour du point
 - Probabilité discrète qu'un point soit dans l'hypercube : $P_{\mathcal{R}} = \int_{\mathcal{R}} p(x_0) d\mathbf{x} = r^d p(x_0)$
- ⇒ Donc $p(x_0) = \frac{P_{\mathcal{R}}}{r^d}$



Estimation non paramétrique par noyaux

Intuition : histogramme centré au point d'intérêt

- Plutôt que de décider d'une discréttisation a priori, l'estimation est faite en centrant une fenêtre autour du point d'intérêt \mathbf{x}_0 (dans un espace de dimension d)
- Soit \mathcal{R} l'hypercube centré en \mathbf{x}_0 de longueur r et $p(\mathbf{x}_0)$ la densité à estimer
- Hypothèse : densité constante autour du point
- Probabilité discrète qu'un point soit dans l'hypercube : $P_{\mathcal{R}} = \int_{\mathcal{R}} p(\mathbf{x}_0) d\mathbf{x} = r^d p(\mathbf{x}_0)$

$$\Rightarrow \text{Donc } p(\mathbf{x}_0) = \frac{P_{\mathcal{R}}}{r^d}$$

Justification mathématique :

- Soit X la v.a. du nombre de \mathbf{x}_i dans \mathcal{R} pour un échantillon de N points
- $P(X = k) = C_N^k P_{\mathcal{R}}^k (1 - P_{\mathcal{R}})^{N-k}$, $\mathbb{E}[X] = NP_{\mathcal{R}}$, donc $P_{\mathcal{R}} = \frac{\mathbb{E}[X]}{N}$

$$\Rightarrow p(\mathbf{x}_0) \simeq \frac{k/N}{r^d}$$

Fenêtre de Parzen

Formalisation pour un échantillon de taille N

- \mathcal{R} est un hypercube, chaque côté de longueur r
 - $V = r^d$, d la dimension de l'espace de représentation
 - $\phi(\mathbf{x}) = \begin{cases} 1 & \text{si } |x^i| \leq 1/2 \\ 0 & \text{sinon} \end{cases}$ fonction indicatrice de l'hypercube unitaire
 - ϕ définit un hypercube unitaire centré à l'origine.
- $\Rightarrow \phi\left(\frac{\mathbf{x}_0 - \mathbf{x}}{r}\right) = 1$ ssi \mathbf{x} est dans l'hypercube de volume V centré en \mathbf{x}_0 .

Conséquence

- Nombre d'échantillons dans l'hypercube : $k = \sum_{i=1}^N \phi\left(\frac{(\mathbf{x}_0 - \mathbf{x}_i)}{r}\right)$
- Densité estimée :

$$p(\mathbf{x}_0) = \frac{1}{N} \sum_{i=1}^N \frac{1}{V} \phi\left(\frac{\mathbf{x}_0 - \mathbf{x}_i}{r}\right)$$

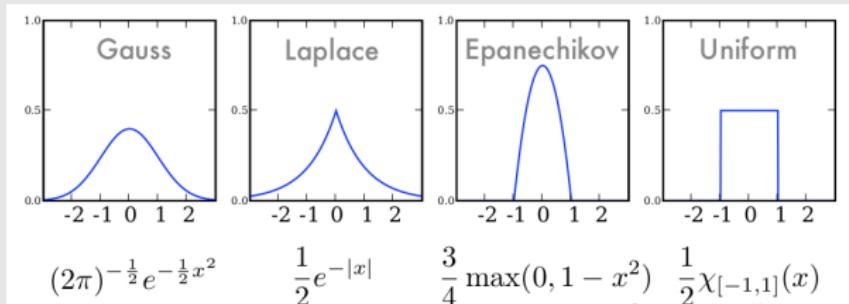
- On note $\delta(\mathbf{x}) = \frac{1}{V} \phi\left(\frac{\mathbf{x}}{r}\right)$, alors

$$p(\mathbf{x}_0) = \frac{1}{N} \sum_{i=1}^N \delta(\mathbf{x}_0 - \mathbf{x}_i)$$

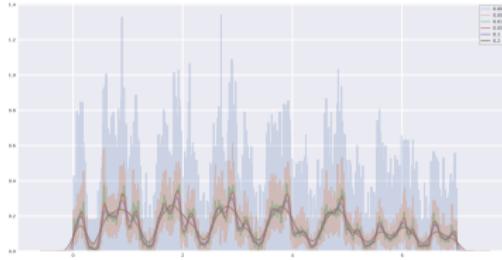
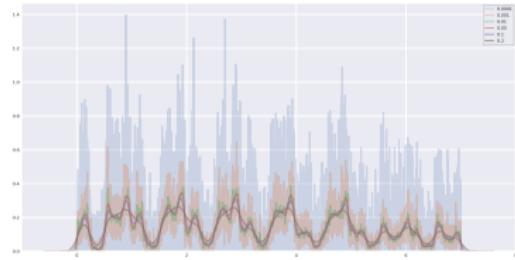
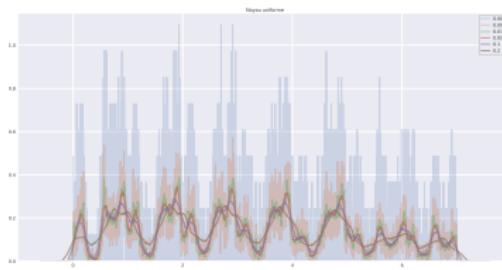
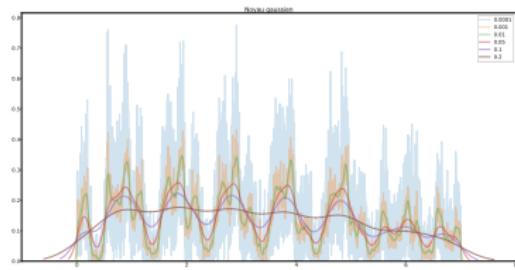
Discussion

Pourquoi se limiter à des hypercubes ?

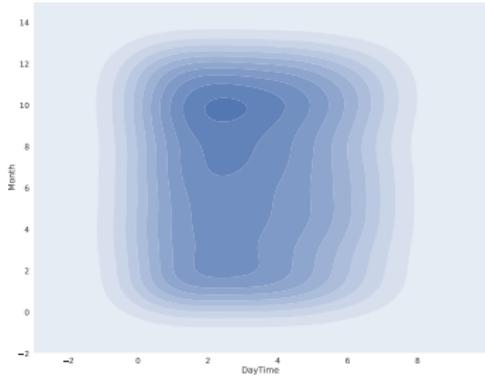
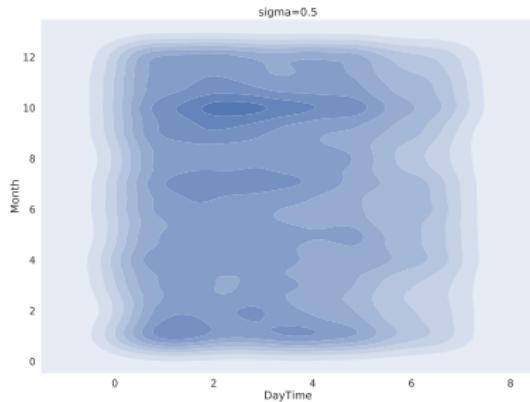
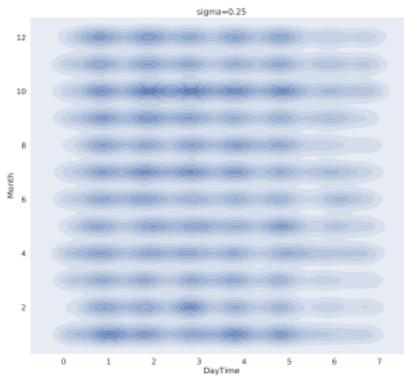
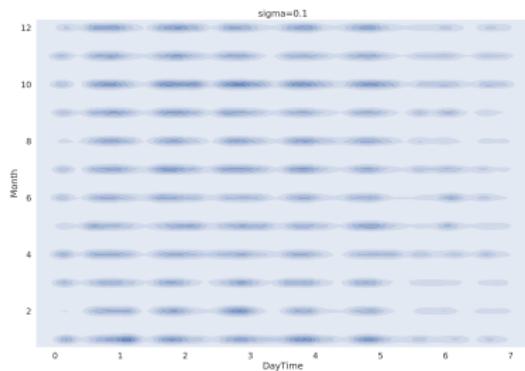
- ϕ peut être plus générale (noyaux)
- permet de pondérer différemment selon la distance au point d'estimation
- conditions nécessaires :
 - ▶ $\phi(x) \geq 0$
 - ▶ $\int \phi(x)dx = 1$
 - ▶ $\phi(x)$ symétrique
 - ▶ $\phi(x)$ maximale en 0 et double monotonie.



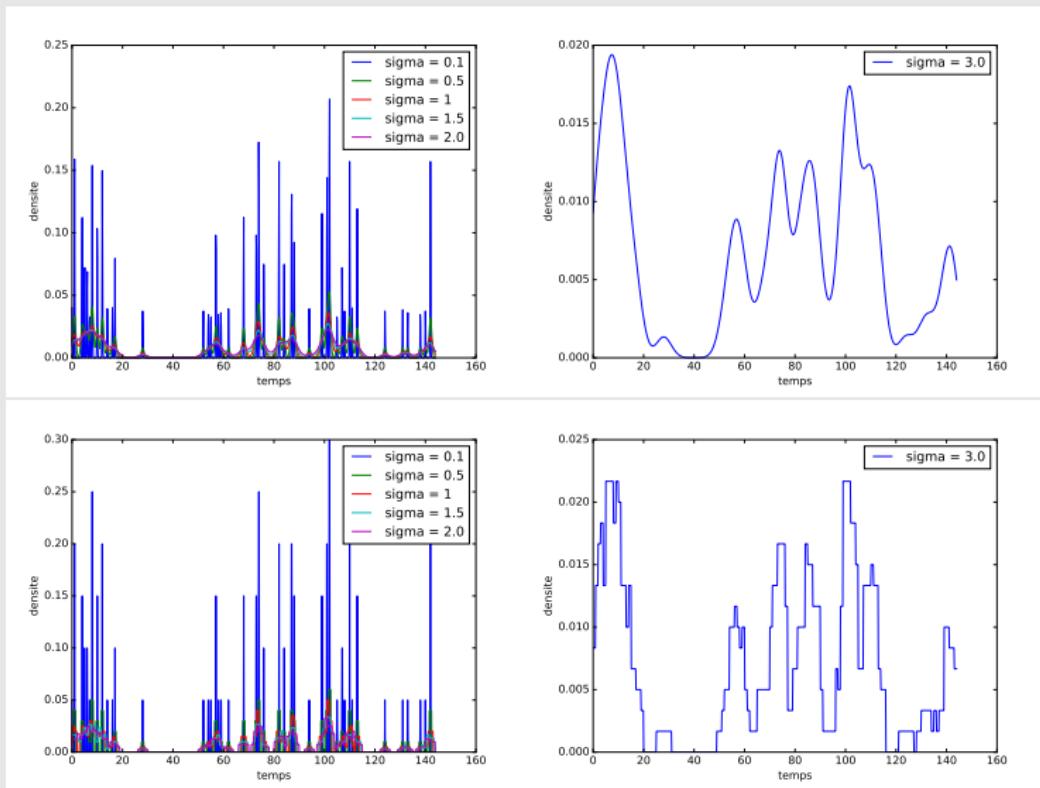
Discussion : exemples Tweets



Discussion : exemples Tweets



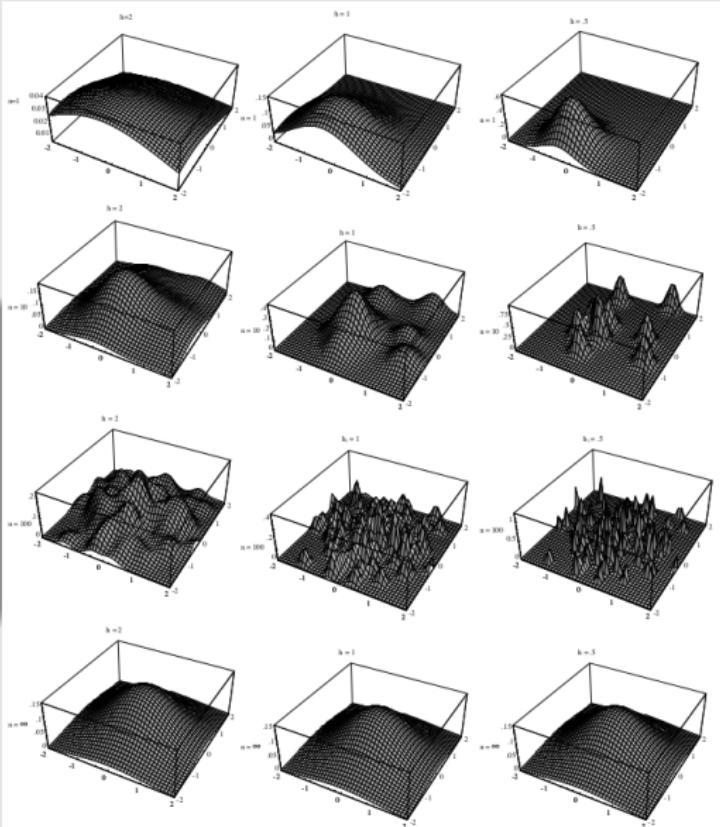
Discussion : exemples velib



Discussion

Effet de r

- r grand
→ δ peu sensible,
paysage homogène
- r petit
→ δ tend vers un pic de Dirac.
- compromis entre petite résolution
et grande variabilité

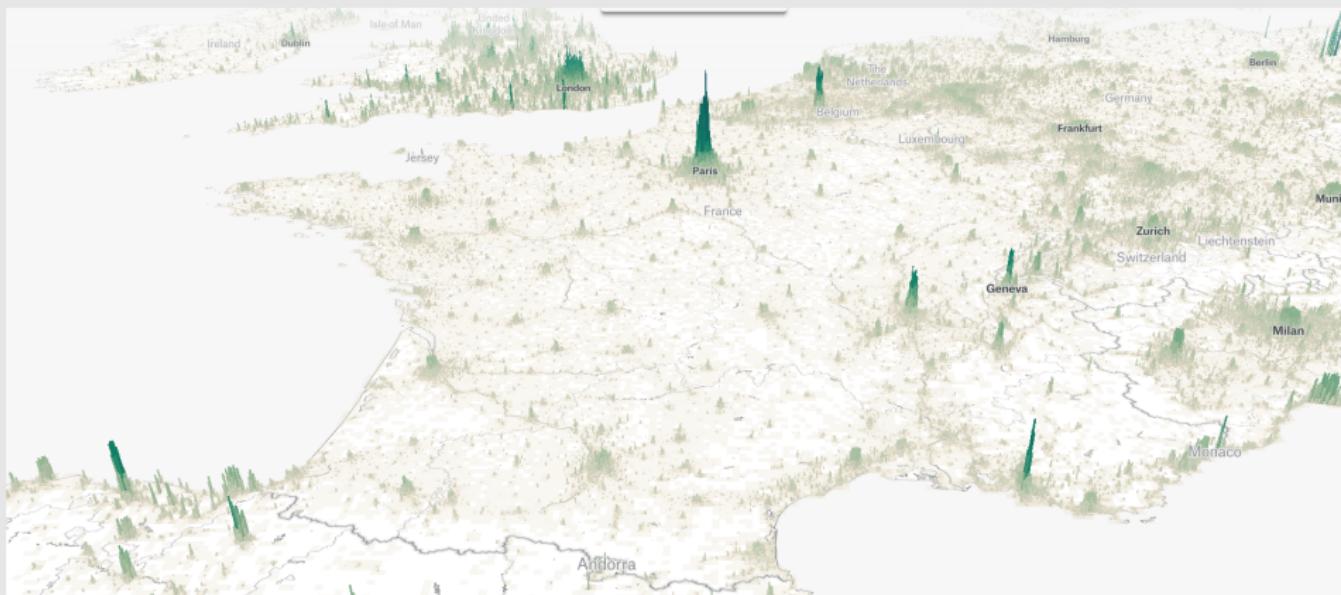


Duda et. al. 01

Évaluation qualitative du modèle

Visualisation de la densité estimée

- Discrétisation de l'espace
- Évaluation en chaque point discret de l'espace de la densité
 - ⇒ comme pour les histogrammes, mais ici le modèle est continu !
 - ⇒ En plus, très belles visualisations



Évaluation quantitative du modèle

Important pour la sélection de l'hyper-paramètre (le rayon du noyau)

- Mesure d'évaluation : vraisemblance !
- Soit $E = \{\mathbf{x}_i\}_{i=1}^N$ un échantillon

$$L(\theta; E) = p_\theta(E) = \prod_{i=1}^N p_\theta(\mathbf{x}_i)$$

(ou plutôt log-vraisemblance ...)

- Problème : vraisemblance maximale pour un rayon infiniment petit
 - Pic de Dirac le plus performant (mais nul en généralisation)
- ⇒ Partition en ensemble d'apprentissage/test (et/ou validation croisée)

Estimation de densité : conclusion

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.

Applications multiples

- Estimation de file d'attentes, d'occupation de lieux, des stocks, des pics de pollution
- Réponse à un problème plus général : lissage des données, traitement de données de capteurs :
On a accès qu'à la réalisation d'une variable aléatoire à certains pas de temps, mais la mesure qui nous intéresse est une mesure continue ...

Deux méthodes principales :

- Histogramme : rapide, mais coûteuse en mémoire, effet de bords
 - Par noyaux : lent, plus flexible, sensibilité de la taille du noyau
- ⇒ Dans tous les cas, la dimension doit être faible ...

Plan

1 Rappel MAPSI/Probabilités

2 Classification bayésienne

3 Estimation de densité par histogramme

4 Estimation de densité par noyaux

5 Estimation de densité et classification

Estimateur de Nadaraya-Watson

De la densité à la classification

On dispose

- d'un label y_i pour chaque \mathbf{x}_i , $y_i \in \{-1, 1\}$
- du nombre d'exemples positifs n_+ , du nombre d'exemples négatifs n_- .
- d'un noyau δ pour l'estimation de densité.

L'objectif de la classification binaire est de déterminer $p(\mathbf{x}|y = 1)$ et $p(\mathbf{x}|y = -1)$

$$\bullet p(\mathbf{x}|y = 1) = \frac{1}{n_+} \sum_{i|y_i=1} \delta(\mathbf{x} - \mathbf{x}_i), \quad p(\mathbf{x}|y = -1) = \frac{1}{n_-} \sum_{i|y_i=-1} \delta(\mathbf{x} - \mathbf{x}_i)$$

$$\bullet p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})} = \frac{\frac{1}{n_y} \sum_{i|y_i=y} \delta(\mathbf{x} - \mathbf{x}_i)^{\frac{n_y}{n}}}{\frac{1}{n} \sum_i \delta(\mathbf{x} - \mathbf{x}_i)} = \frac{\sum_{i|y_i=y} \delta(\mathbf{x} - \mathbf{x}_i)}{\sum_i \delta(\mathbf{x} - \mathbf{x}_i)}$$

$$\Rightarrow p(y_+|\mathbf{x}) - p(y_-|\mathbf{x}) = \frac{\sum_{j=1}^N y_j \delta(\mathbf{x} - \mathbf{x}_j)}{\sum_{i=1}^N \delta(\mathbf{x} - \mathbf{x}_i)} = \frac{1}{\sum_{i=1}^N \delta(\mathbf{x} - \mathbf{x}_i)} \sum_{j=1}^N y_j \delta(\mathbf{x} - \mathbf{x}_j)$$

- directement adaptable à la régression
- Intuition : moyennage sur le voisinage local du point à estimer (en pondérant par la distance)

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 paramètre : k le nombre de voisins à prendre en compte
- $p(y|x) = \frac{1}{k} \sum_{j, x_j \in \{k\text{- plus proches}\}} y_j$

Discussion

- Parzen : travail sur le volume, pas de contrôle 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 partitionnement (KD-tree) et autres heuristiques pour accélérer)
 - ▶ des paramètres à choisir ...
- Comment choisir les paramètres ?