



# Machine Learning Cours 1

## Master 1 DAC

Nicolas Baskiotis

[nicolas.baskiotis@sorbonne-universite.fr](mailto:nicolas.baskiotis@sorbonne-universite.fr)

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

S2 (2023-2024)

# Plan

- 1 Organisation de l'UE
- 2 Introduction
- 3 Les problématiques générales
- 4 Classification bayésienne
- 5 Estimation de densité par histogramme
- 6 Estimation de densité par noyaux
- 7 Estimation de densité et classification



# Informations administratives

## Créaux

- Cours : Mercredi 16h-18h
- TD/TME :
  - ▶ groupe 1 : Mardi 8h30-12h45 (Nicolas Thome)
  - ▶ groupe 2 : Jeudi 8h30-12h45 (Nicolas Baskiotis)
  - ▶ groupe MLL : Mercredi 8h30-12h45 (Stéphane Doncieux)

## Supports et références

- Site du master : <http://dac.lip6.fr/master/ml/>
- Beaucoup de références on-line, beaucoup de ebooks et de livres, cf site
- en cas de questions/problèmes  
→ Mattermost : channel ML (en priorité) → {nicolas.baskiotis, nicolas.thome, stephane.doncieux}@sorbonne-universite.fr

## Évaluation

- CC : travail en TME, projet et partiel
- Examen
- Pour MLL : que partiel/examen et TME.

# Plan

- 1 Organisation de l'UE
- 2 Introduction**
- 3 Les problématiques générales
- 4 Classification bayésienne
- 5 Estimation de densité par histogramme
- 6 Estimation de densité par noyaux
- 7 Estimation de densité et classification



## Classification de documents

### E-mails en spam, shopping, travail, . . .

Supprimer tous les spams maintenant (les messages se trouvant dans le dossier Spam depuis plus de 30 jours sont automatiquement supprimés)

<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Tatianna	Re: Para os homens - Vai lhe interessar muito!	01:50
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	comebuy	Téléphones les plus compétitifs de Comebuy	22:38
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Francois	100 raisons de jouer sur Majestic	27 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Fund Investigation Bureau	TREAT AS URGENT RIGHT AWAY	27 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Mrs Elizabeth Johnson	Hello My Beloved One.	27 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Evelyn	Re: Amigo, não está satisfeito com o tamanho? Isto pode te ajudar!	27 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Amanda, Amanda (2)	Re: Amigo, o que vc faria com 10cm a mais?	26 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Groupe Partouche	Et encore un gagnant au Megapot !	26 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Carli, Joshua Daniel	N/A	26 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	RCH Tournoi	Votre Semaine avec 100000 en Tout	26 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Jemmy Kiamet	Nicolas Baskiotis F-E...E...L-L-N G...H...O...R...N-Y-?-...-G-E-T...L...A...I...D...-...N...O...W!	26 janv.
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Jean-Pierre	Les meilleurs casinos pour les joueurs français	25 janv.

Principale

Réseaux sociaux

Promotions



<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	CollierPrenom	Annonce ⓘ Spécial St Valentin - 3 Jours Seulement - 15% de Réduction !	×
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	SoftLayer.com	Annonce ⓘ Get a Secure Cloud - We've secured the public cloud with private servers, private networks, and full private clouds.	×
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Booking.com	Last-minute deals for Montréal and London. Get them before they're gone!	28/12/2014
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Voyages-sncf.com	DERNIERE MINUTE NOUVEL AN : profitez des meilleurs prix !	26/12/2014
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Impossible	Year's End Clearance - Up to 20% off Film and Accessories	26/12/2014
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Booking.com	Nicolas - you qualify for at least 20% off places to stay	26/12/2014
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Communauté d'entraide Gr.	Nicolas, des questions sur vos produits ?	25/12/2014
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Deloitte	Trouvez l'offre de recrutement de votre	25/12/2014

gmail.com

# En texte toujours

## Reconnaissance de chiffres

8 2 9 4 4 6 4 9 7 0 9 2 9 5 1 5 9 1 0 3  
2 3 5 9 1 7 6 2 8 2 2 5 0 7 4 9 7 8 3 2  
1 1 8 3 6 1 0 3 1 0 0 1 1 2 7 3 0 4 6 5  
2 6 4 7 1 8 9 9 3 0 7 1 0 2 0 3 5 4 6 5

## Ou de captcha

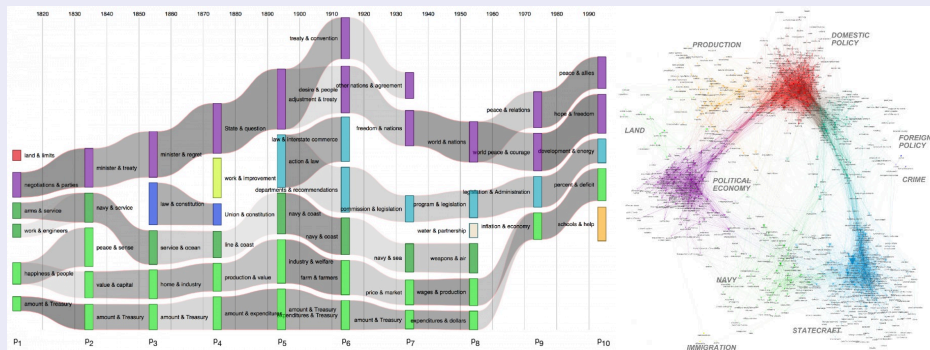
[Yann et al. 08], Newcastle University



Characters under typical distortions	Recognition rate
	~100%
	96+%
	100%
	98%
	~100%
	95+%

# Sur des documents

## Détection de thèmes (topic detection)



Analyse de 255 discours de l'état de l'union, États-Unis

[Rule et al, 2014]

Et plein d'autres applications : traduction, détection de plagiat, résumé automatique, ...

⇒ U.E. Traitement Automatique du Langage

# En image

## Détection de visages

(opencv)



## Mais aussi ...

(betafaceapi.com)



Score: 0.42  
X: 398.67  
Y: 29.66  
Width: 26.79  
Height: 26.79  
Angle: -5.45

age : 37 (16%), gender : male, race : white, chin size : average, color background : 4c5042 (15%), color clothes middle : 3295eb (48%), color clothes sides : 38a9f5 (96%), color eyes : ac8066, color hair : fbf2ea (80%), color mustache : a56855 (65%), color skin : dbb5a1, eyebrows corners : extra low, eyebrows position : average, eyebrows size : extra thin, eyes corners : low, eyes distance : average, eyes position : average, eyes shape : extra round, glasses rim : no, hair beard : none, hair color type : blond (80%), hair forehead : yes, hair length : none, hair mustache : thick, hair sides : very thin, hair top : short, head shape : average, head width : extra narrow, mouth corners : low, mouth height : extra thin, mouth width : extra small, nose shape : extra straight, nose width : wide, teeth visible : no [collapse]



Score: 0.57  
X: 216.66  
Y: 155.08  
Width: 28.34  
Height: 28.34  
Angle: 0.95

age : 46 (23%), gender : male, race : white, chin size : extra small, color background : 0c0c0d (36%), color beard : 4a2617 (50%), color clothes middle : a22e55 (82%), color clothes sides : a54031 (74%), color eyes : 966a58, color hair : 655348 (77%), color skin : b98f78, eyebrows corners : average, eyebrows position : extra high, eyebrows size : extra thin, eyes corners : average, eyes distance : close, eyes position : extra low, eyes shape : extra thin, glasses rim : no, hair beard : short, hair color type : brown light (77%), hair forehead : no, hair length : short, hair mustache : none, hair sides : thin, hair top : short, head shape : rect, head width : extra wide, mouth corners : average, mouth height : extra thin, mouth width : average, nose shape : average, nose width : extra narrow, teeth visible : no [collapse]

# En image

## Catégorisation et organisation automatique





# En image

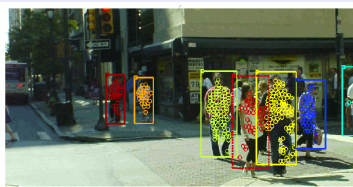
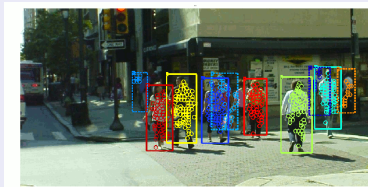
## Détection d'objets

teradeep.com, Purdue University



## Tracking

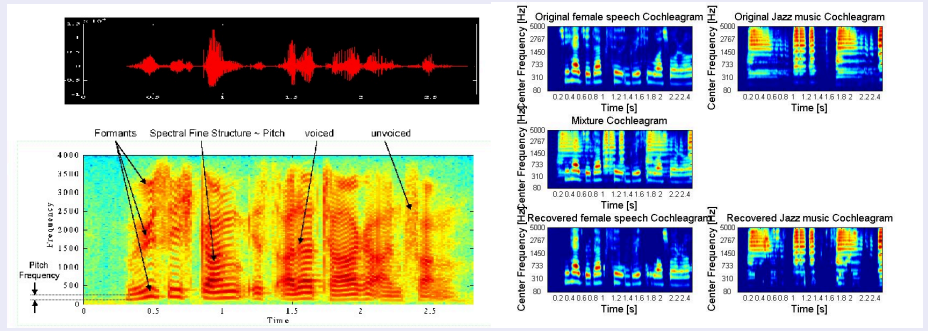
[Fragkiadaki et al. 12], Pennsylvania University



# Et l'audio ...

## Reconnaissance de la parole, séparation de sources

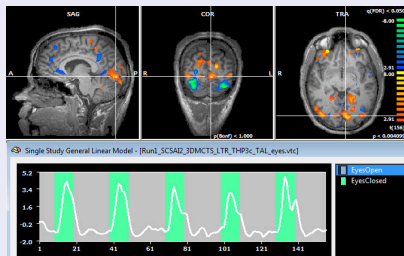
<http://markus-hauenstein.de>



Mais aussi débruitage, transcription musicale, reconnaissance du locuteur, classification/identification de musiques...

# Interface cerveau-machine (BCI)

## Classification d'actions, de pensées



## Contrôle

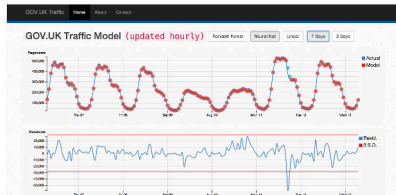


# Objets connectés

## Traqueurs d'activité



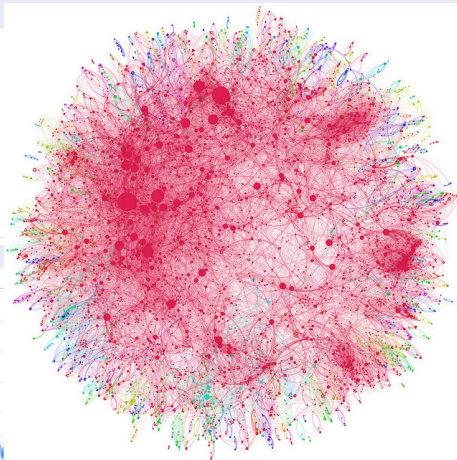
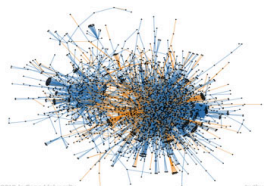
## Surveillance vidéo, monitoring consommation électrique, sécurité réseau



# Réseaux sociaux

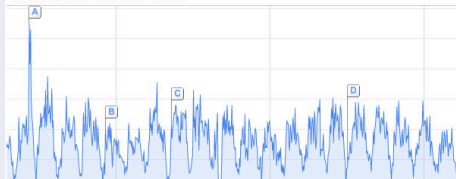
## Détection de communauté, phénomènes de diffusions, classification

Largest Diffusion Network



Meme Activity

Zoom: 1' 5' 1h 1d 5d 1m 3m 6m 1y Max



# Matchmaking

de profils, sites de rencontre



Experts, CV - Emplois, Jeux



Linked in  
viadeo



# Systèmes de recommandation

## De musiques, de films, de produits, d'amis

### Similar Artists



- 1 Bob Dylan
- 2 Radiohead
- 3 Led Zeppelin
- 3 The Rolling Stones
- 5 Pink Floyd
- 6 David Bowie
- 7 The Who
- 8 John Lennon



Recommendation Engine Home New Music My History Web: musicandmovies.com Log Out

Search All

Movies Music Action Arts

 Umbrellas of Cherbourg, The Jacques Demy Drama   1964   Unrated Select	 Brokedown Palace Jonathan Kaplan Drama   1998   PG-13 Select	 West Beirut Ziad Doueiri Drama   1998   PG-13 Select	 Suspect Peter Yates Thriller   1997   R Select	 Heights Jeremy Kagan Drama   2004   R Select	 Babylon A.D. Matthew Kassovitz Action   2008   PG-13 Select
---	---	---	---	---	--

amazon.com

Recommended for You

Amazon.com has new recommendations for you based on [items](#) you purchased or told us you own.



[The Little Big Things: 162 Ways to Pursue EXCELLENCE](#)



[Fascinate: Your 7 Triggers to Persuasion and Captivation](#)



[Sherlock Holmes \[Blu-ray\]](#)



[Alice in Wonderland \[Blu-ray\]](#)

facebook

write friends | click anywhere to drag | mouse wheel to zoom at mouse pointer | double-click nodes to visit profile pages

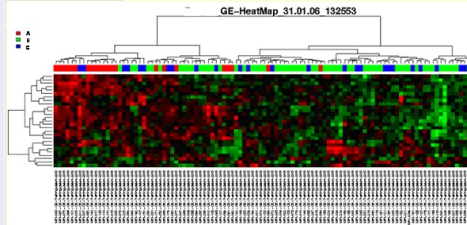


# Et dans d'autres sciences

## Biologie IZBI, Leipzig University

### Gene Signal Value Visualization - Gene Expression Heatmap

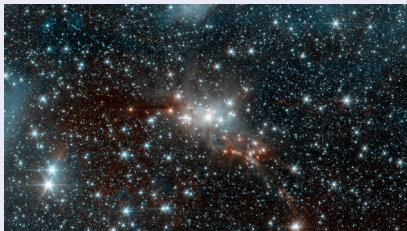
This form draws the heatmap of Gene Expression signals determined by a selected Experiment Group and a selected Gene Group.



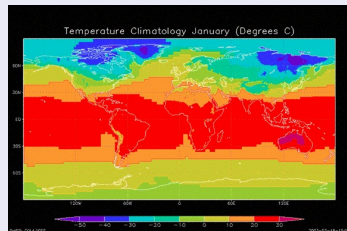
## Économie



## Astronomie

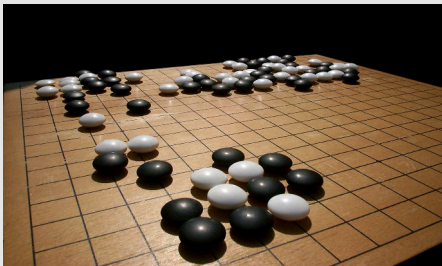


## Climatologie (complétion données)





# Dans les jeux et la robotique



# Dans l'industrie



# Le Machine Learning défini par lui même

NI

Donne moi une courte définition du Machine Learning



Le Machine Learning est une technique d'apprentissage automatique utilisée en informatique pour créer des modèles capables de prédire et de résoudre des problèmes complexes en utilisant des données en grande quantité. Ces modèles sont capables d'apprendre par eux-mêmes en s'ajustant en fonction des données qu'on leur fournit, sans qu'on ait besoin de les programmer explicitement pour effectuer chaque tâche.



NI

Qu'est ce que l'apprentissage automatique ?



L'apprentissage automatique est un domaine de l'informatique qui s'intéresse à la création de modèles informatiques capables d'apprendre et de s'améliorer de manière autonome en utilisant des données en grande quantité. Ces modèles sont souvent utilisés pour résoudre des problèmes complexes, comme la reconnaissance vocale, la reconnaissance d'images ou encore la prédiction de tendances financières.



(ChatGPT d'OpenAI : <https://chat.openai.com/chat>)

## En quelques mots

- Trouver des structures, des régularités dans des observations.
- Prédire de nouvelles observations.

## Touche à beaucoup de domaine, interdisciplinarité très forte

- Statistiques : théorie de l'apprentissage, fouille de données, inférence
- Informatique : IA, vision, RI
- Ingénierie : signal, contrôle, robotique
- Science cognitive, psychologie, neuroscience, épistémologie
- Économie : théorie de la décision, théorie des jeux

## L'apprentissage artificiel

- étudie les algorithmes qui améliorent leur performance sur une tâche donnée en fonction de leur expérience.
- fondements mathématiques, informatiques et applications concrètes des systèmes qui apprennent, raisonnent et agissent.

# Quand appliquer l'apprentissage ?

## Lorsque :

- l'expertise humaine est absente
- impossible d'expliquer cette expertise
- les solutions sont dynamiques
- les solutions doivent être adaptées à beaucoup de cas spécifiques
- la taille du problème est trop grand pour que l'humain puisse le résoudre

# Plan

- 1 Organisation de l'UE
- 2 Introduction
- 3 Les problématiques générales**
- 4 Classification bayésienne
- 5 Estimation de densité par histogramme
- 6 Estimation de densité par noyaux
- 7 Estimation de densité et classification

# Les grandes familles

## Apprentissage supervisé

- Classification
- Régression
- Forecasting
- Complétion de données
- Ranking
- Recommandation

## Apprentissage non supervisé

- Clustering
- Apprentissage de représentation, de dictionnaire
- Analyse de séquences
- Représentation hiérarchique
- Détection d'anomalies

## Apprentissage par renforcement

- Apprendre à jouer
- Apprendre à interagir avec l'environnement

# Apprentissage supervisé

## Données du problème

- Une représentation  $X$  des objets de l'étude
- Une sortie d'intérêt  $y$  qui peut être numérique, catégorielle, structurée, complexe (label, réponse, étiquette, ...)
- Un ensemble d'exemples, d'échantillons, sous leur représentation  $X$  et avec leur sortie connue  $\{(x_1, y_1), \dots, (x_n, y_n)\}$

## Objectifs

- Prédire de manière précise la sortie  $y$  pour un nouvel exemple  $x$  non vu
- Comprendre quels facteurs influencent la sortie
- Évaluer la qualité de nos prédictions



# Apprentissage non supervisé

## Données du problème

- Une représentation  $X$  des objets de l'étude
- Un ensemble d'exemples, d'échantillons, sous leur représentation  $X$ ,  $\{x_1, \dots, x_n\}$
- Pas de variable de sortie !

## Objectifs

- Trouver des groupes d'objets "semblables"
  - Organiser les données d'une manière "logique"
  - Trouver les "similarités" des objets
  - Trouver des "représentations" des objets
- ⇒ on ne sait pas bien ce que l'on cherche
- ⇒ tout un art !

# Apprentissage par renforcement

Apprentissage continu en fonction du retour d'expérience

## Données du problème

- Un état décrit l'environnement courant
- Un ensemble d'actions sont possibles
- Une politique permet de choisir en fonction de l'état l'action à effectuer
- A l'issue de chaque action, une récompense est observée

## Objectifs

- S'améliorer ! (améliorer la politique de choix de l'action)
- Éviter les situations d'échecs
- Comprendre la dynamique du problème

# Ce cours

## Centré sur le ML jusqu'aux années 2010 !

- Problématiques générales (biais, variance, évaluation, sur-apprentissage, représentation des données)
- Algorithmes supervisés (k-nn, bayésien, perceptron, réseaux de neurones, svm, ...)
- Algorithmes non supervisés (hiérarchique, k-means, ...)
- Une prise en main du Deep (approfondissement en M2)

## Objectifs

- Comprendre les différentes techniques en profondeur, principalement algorithmiquement (et un peu théoriquement)
- Comprendre les notions fondamentales de l'apprentissage
- Savoir choisir et évaluer une approche

# Plan du cours

- 1 Intro, Estimation de densité (MLL)
- 2 Modèles linéaires, Descente de gradient (MLL)
- 3 Réseaux de neurones 1
- 4 Réseaux de neurones 2
- 5 SVMs et noyaux (MLL)
- 6 Ensemble Learning, Boosting, Théorie de l'apprentissage (MLL)
- 7 Apprentissage non supervisé, Recommandation (MLL)
- 8 Séries temporelles
- 9 Processus Gaussien
- 10 Apprentissage par renforcement

# Apprentissage bayésien

## Estimation de densité

Nicolas Baskiotis

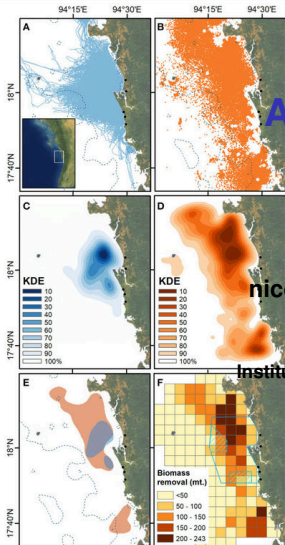
[nicolas.baskiotis@sorbonne-universite.fr](mailto:nicolas.baskiotis@sorbonne-universite.fr)

équipe MLIA,

Institut des Systèmes Intelligents et de Robotique (ISIR)

Sorbonne Université

S2 (2023-2024)



# Plan

- 1 Organisation de l'UE
- 2 Introduction
- 3 Les problématiques générales
- 4 Classification bayésienne**
- 5 Estimation de densité par histogramme
- 6 Estimation de densité par noyaux
- 7 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) (3 dimensions,  $\mathcal{X} = \mathbb{R}^3$ )
- Une fonction de prédiction :  $f(\mathbf{x}) = y_+$  si  $x_1 \geq 2000$  sinon  $y_-$

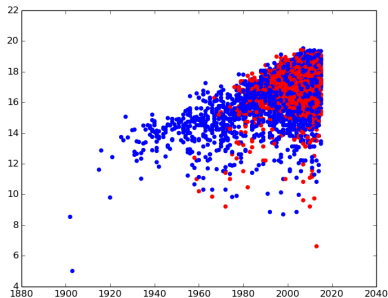
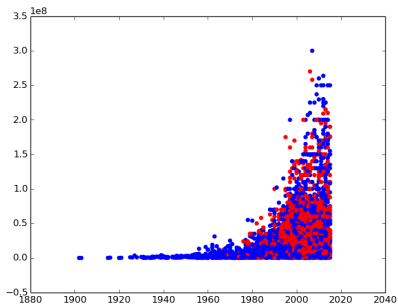
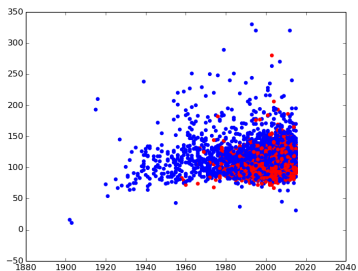
## Tweets de Trump

- 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

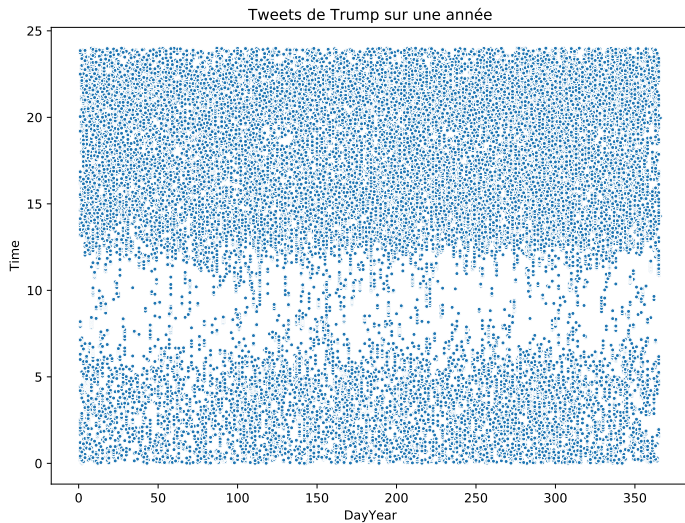
Sur la base imdb ... :

- Année vs Durée
- Année vs Budget

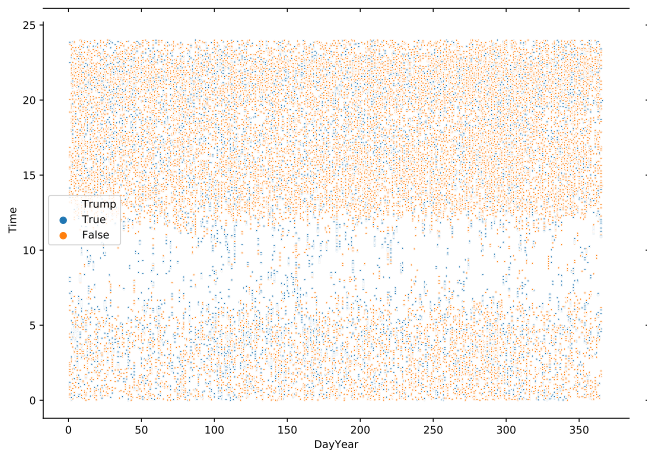




# Classification binaire : tweets



# 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

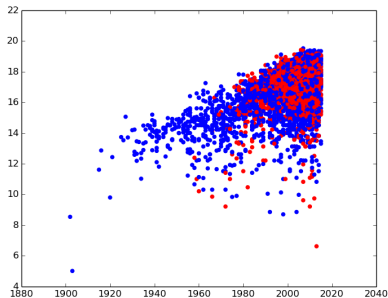
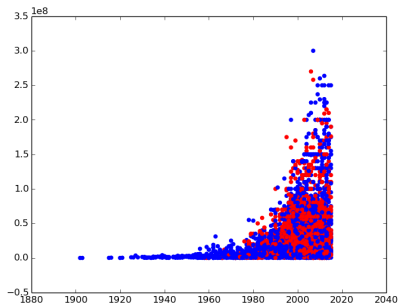
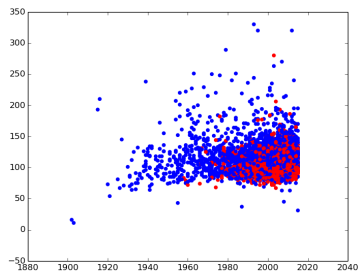
- Toujours la même décision
- On ne tient pas compte de la description  $\mathbf{x} \in \mathcal{X}$ .
- Évaluation du risque :  $R = \min(P(y_+), P(y_-))$

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  $p(\mathbf{x}|y)$  :

# Dans un monde idéal (bayésien)

## Si on dispose ...

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

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

# Dans un monde idéal (bayésien)

## Si on dispose ...

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

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

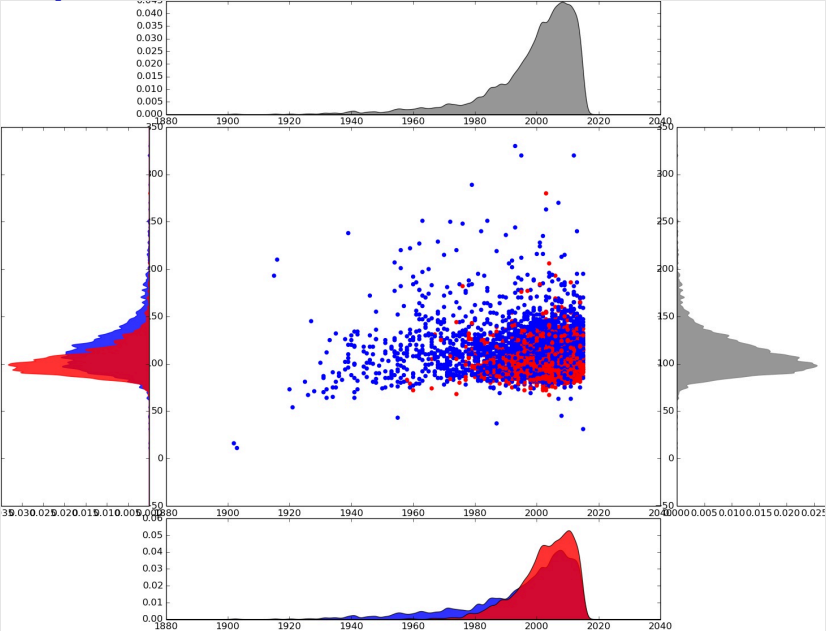
## Alors

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

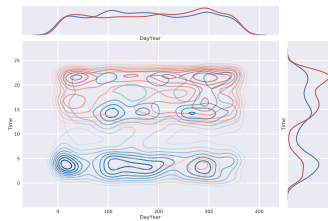
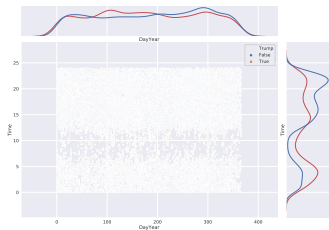
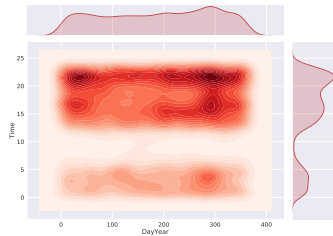
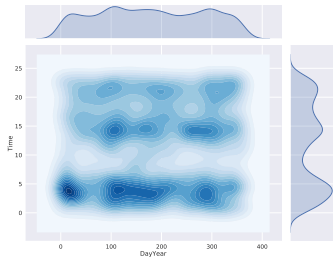
- On appelle  $p(\mathbf{x}|y)$  la vraisemblance de  $\mathbf{x}$  par rapport à  $y$ .
  - décision bayésienne : choisir  $y_+$  si  $p(y_+|\mathbf{x}) > p(y_-|\mathbf{x})$ , le contraire sinon
- $\Rightarrow f(\mathbf{x}) = \operatorname{argmax}_y p(y|\mathbf{x}) = \operatorname{argmax}_y \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})}$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

⇒ 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 :  $R(y_i|\mathbf{x}) = \sum_j l(y_i, y_j)P(y_j|\mathbf{x}) = 1 - P(y_i|\mathbf{x})$
- $R(f) = \int_{\mathbf{x}} R(f(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x}$
- 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 ?

# 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 ?
- ⇒ 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, ...

## Ou s'en affranchir (en partie)

- C'est la suite de ce cours !

# Plan

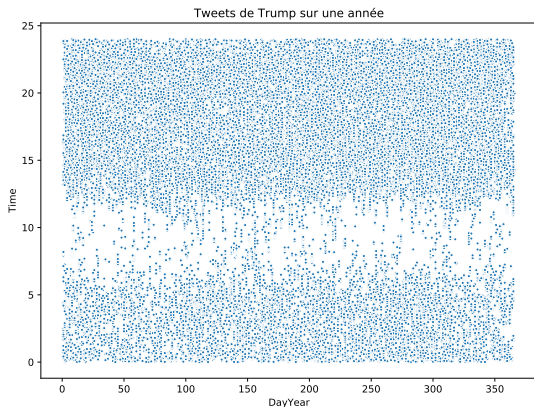
- 1 Organisation de l'UE
- 2 Introduction
- 3 Les problématiques générales
- 4 Classification bayésienne
- 5 Estimation de densité par histogramme**
- 6 Estimation de densité par noyaux
- 7 Estimation de densité et classification

# Le problème

## Données

Un échantillon est décrit selon des caractéristiques (dimensions) continues/catégorielles/ordinales

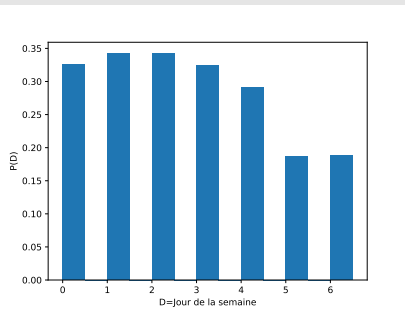
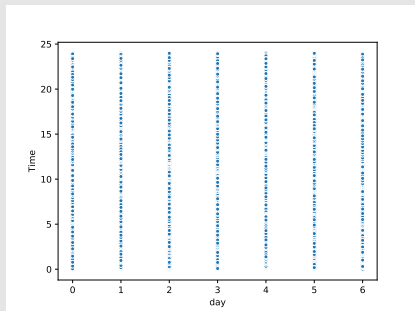
⇒ Quelle est la loi sous-jacente de la distribution des données ?





# 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 :

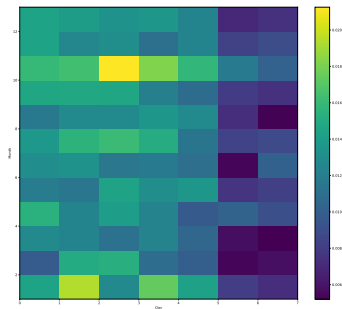
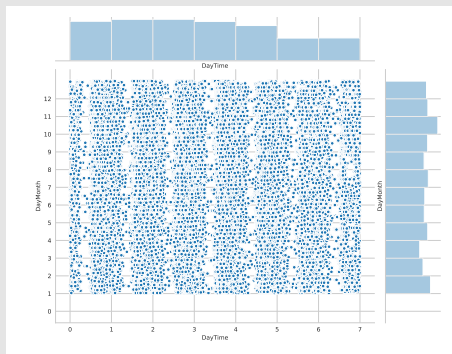
$$\bullet P(D = k) = \frac{|\{\mathbf{x}_i | x_i^d = k\}|}{N} = \frac{\sum_{i=1}^N \mathbf{1}_{x_i^d = k}}{N}$$

- 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 :

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

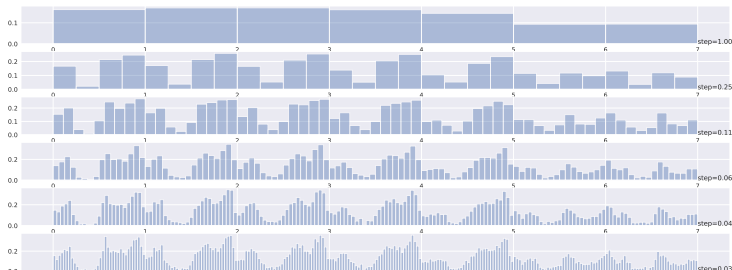


# Lorsque les variables sont continues

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 $H$

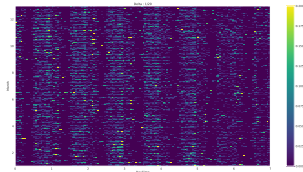
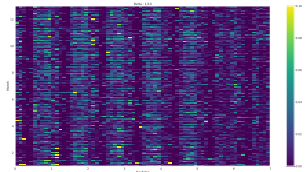
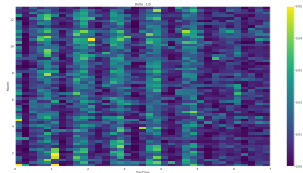
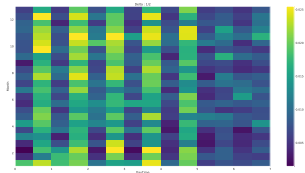
- Discrétisation des valeurs de la v.a.
- Choix d'un pas de discrétisation : 0.5 jours par exemple  $\Delta = 0.5 \Rightarrow 7/\Delta$  créneaux  $T_j$ .
- L'estimation discrète est alors :  $P(H \in T_j) = \frac{|\{x_i | x_i^t \in T_j\}|}{N}$
- Densité de probabilité associée :  $p(h \in T_j) = \frac{P(H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



# Limites de la méthode des histogrammes

En grande dimension  $d$  :

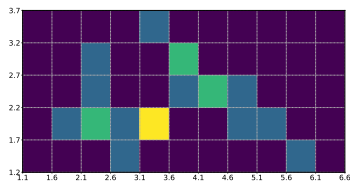
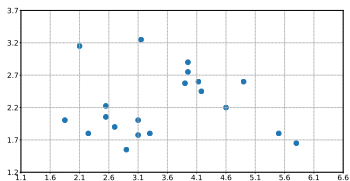
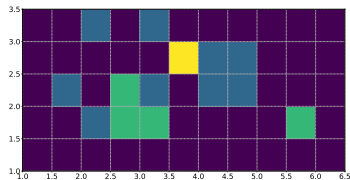
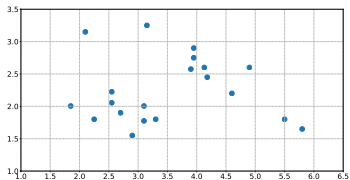
- la taille du modèle (le nombre de cases) augmente exponentiellement en  $(\frac{D}{\Delta})^d$
- Beaucoup de cases sans aucun échantillon
- Celles qui en ont en ont un nombre faible  $\rightarrow$  peu représentatives



# Limites de la méthode des histogrammes

## Effet de bords

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



Exemple données artificielles

# Limites de la méthode des histogrammes

## Effet de bords

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



# Plan

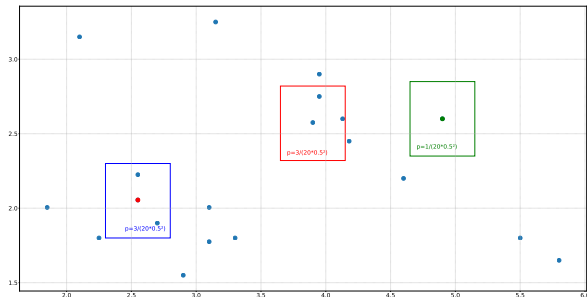
- 1 Organisation de l'UE
- 2 Introduction
- 3 Les problématiques générales
- 4 Classification bayésienne
- 5 Estimation de densité par histogramme
- 6 Estimation de densité par noyaux**
- 7 Estimation de densité et classification

# Estimation non paramétrique par noyaux

## Intuition : histogramme centré au point d'intérêt

- Plutôt que de décider d'une discrétisation 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 :  $P_{\mathcal{R}} = \int_{\mathcal{R}} p(\mathbf{x}_0) d\mathbf{x} = r^d p(\mathbf{x}_0)$

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





# Estimation non paramétrique par noyaux

## Intuition : histogramme centré au point d'intérêt

- Plutôt que de décider d'une discrétisation 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

$$\bullet P(X = k) = C_N^k P_{\mathcal{R}}^k (1 - P_{\mathcal{R}})^{N-k}, \mathbb{E}[X] = NP_{\mathcal{R}}, \text{ donc } P_{\mathcal{R}} = \frac{\mathbb{E}[X]}{N}$$

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

# 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
  - $\phi$  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 :

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

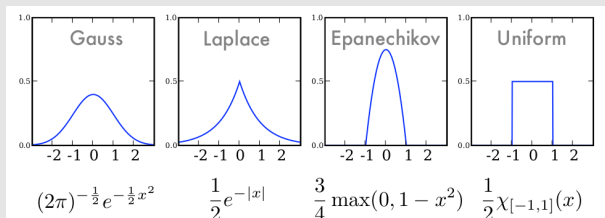
- On note  $\delta(\mathbf{x}) = \frac{1}{V} \phi\left(\frac{\mathbf{x}}{r}\right)$ , alors

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

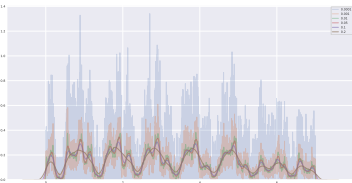
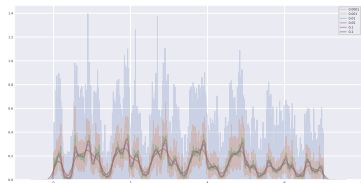
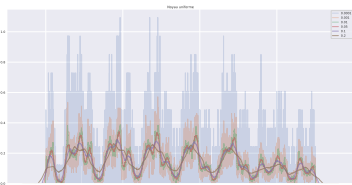
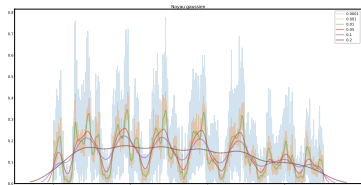
# Discussion

## Pourquoi se limiter à des hypercubes ?

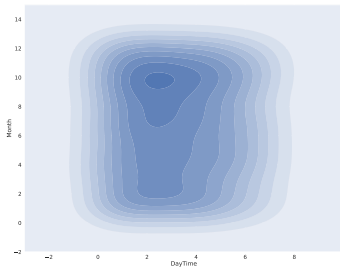
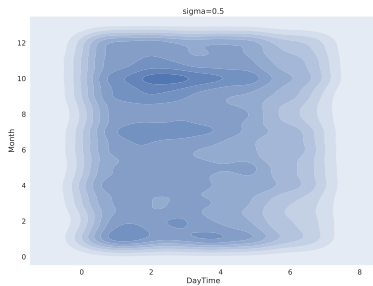
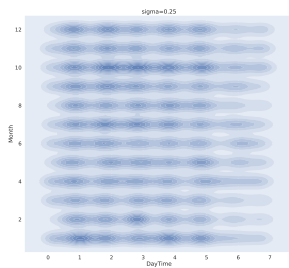
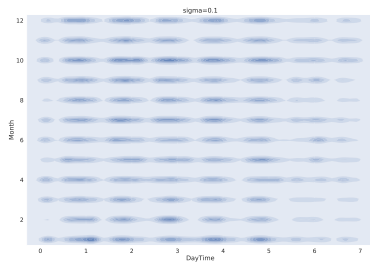
- $\phi$  peut être plus générale (noyaux)
- permet de pondérer différemment selon la distance au point d'estimation
- conditions nécessaires :
  - ▶  $\phi(x) \geq 0$
  - ▶  $\int \phi(x) dx = 1$
  - ▶  $\phi(x)$  symétrique
  - ▶  $\phi(x)$  maximale en 0 et double monotonie.



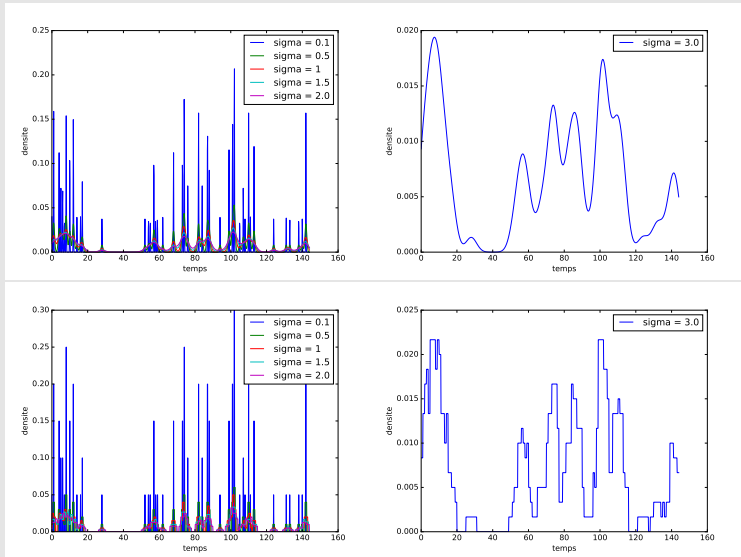
# Discussion : exemples Tweets



# Discussion : exemples Tweets



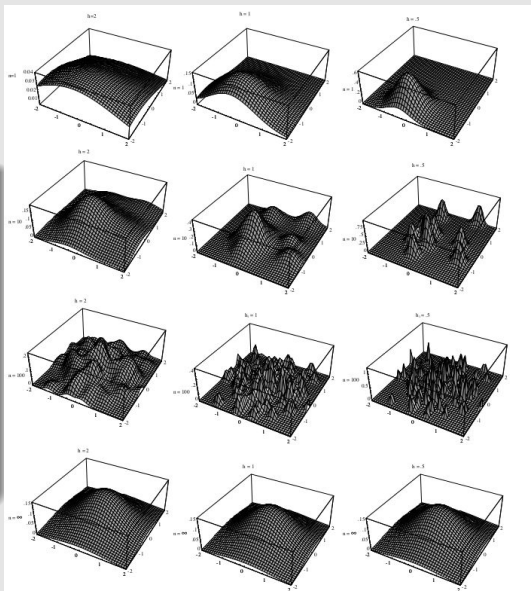
# Discussion : exemples velib



# Discussion

## Effet de $r$

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

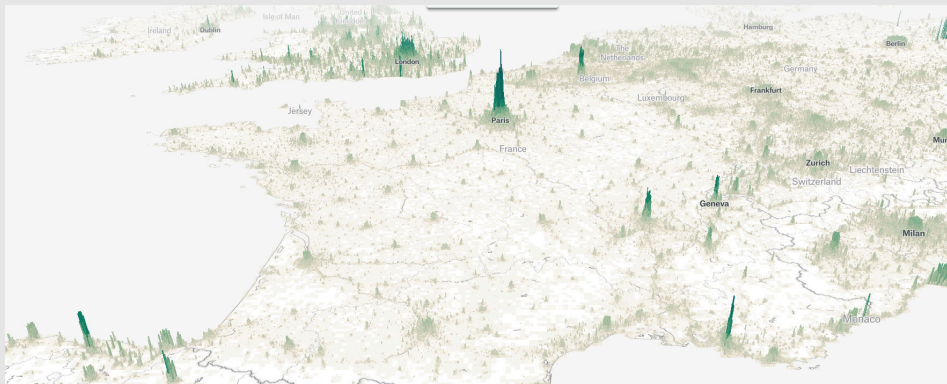


Duda et. al. 01

# 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 Organisation de l'UE
- 2 Introduction
- 3 Les problématiques générales
- 4 Classification bayésienne
- 5 Estimation de densité par histogramme
- 6 Estimation de densité par noyaux
- 7 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  $\delta$  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 ?