

Machine Learning

Cours 1

Master 1 DAC

Nicolas Baskiotis

nicolas.baskiotis@lip6.fr
<http://webia.lip6.fr/~baskiotis>

équipe MLIA, Laboratoire d'Informatique de Paris 6 (LIP6)
Sorbonne Université

S2 (2019-2020)

Plan

- 1 Organisation de l'UE
- 2 Introduction
- 3 Les problématiques générales
- 4 Premier exemple : classification de films et arbres de décision

Informations administratives

Crénaux

- Cours : Mercredi 16h-18h
- TD/TME :
 - ▶ groupe 1 : Mardi 8h30-12h45 (Nicolas Baskiotis)
 - ▶ groupe 2 : Jeudi 8h30-12h45 (Sylvain Lamprier)
 - ▶ groupe MLL : Mercredi 8h30-12h45 (Benjamin Piwowarski)

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
 - {nicolas.baskiotis, sylvain.lamprier,
benjamin.piwowarski}@lip6.fr
 - Mattermost : channel ML

Évaluation

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

Pré-requis

- Mathématiques :
 - ▶ Probabilités et Statistiques
→ Cours MAPSI de M1-S1, Cours 3I005 de L3
 - ▶ Notions d'algèbre linéaire : matrice, vecteur, norme, produit scalaire, ...
 - ▶ Notions d'analyse : dérivée, dérivée partielle, gradient, intégrale, ...
- Du formalisme logique :
 - ▶ savoir lire une formule mathématique :
 $\forall x \in \mathcal{E} \cup \mathcal{F}, \exists Y \subset \mathcal{Y}, p(x) \in Y \Rightarrow \neg q(Y)$
- Programmation (pour les TMEs et projet) : Python
 - ▶ savoir faire des scripts (MAPSI)
 - ▶ modules : `matplotlib`, `numpy` puis `scikit-learn`

Plan

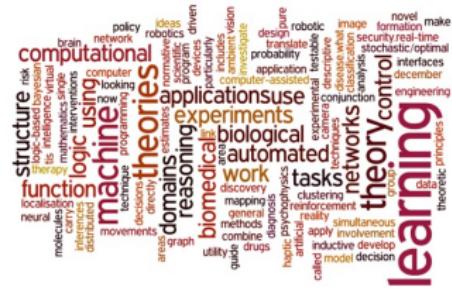
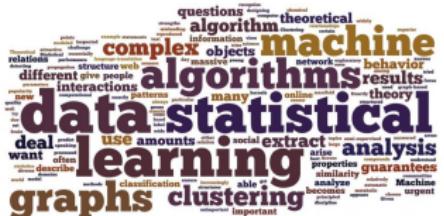
1 Organisation de l'UE

2 Introduction

3 Les problématiques générales

4 Premier exemple : classification de films et arbres de décision

Apprentissage artificielle ? (Machine Learning)



A votre avis, ça regroupe quoi ?

En texte

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

gmail.com

Principale	Réseaux sociaux	Promotions		
<input type="star"/>	CollierPrenom Annonce Spécial St Valentin - 3 Jours Seulement - 15% de Réduction !		X	
<input type="star"/>	SoftLayer.com Annonce Get a Secure Cloud - We've secured the public cloud with private servers, private networks, and full private clouds.		X	
<input type="checkbox"/>	<input type="star"/>	<input checked="" type="square"/>	Booking.com Last-minute deals for Montréal and London. Get them before they're gone!	28/12/2014
<input type="checkbox"/>	<input type="star"/>	<input checked="" type="square"/>	Voyages-sncf.com DERNIERE MINUTE NOUVEL AN : profitez des meilleures prix !	26/12/2014
<input type="checkbox"/>	<input type="star"/>	<input checked="" type="square"/>	Impossible Year's End Clearance - Up to 20% off Film and Accessories	26/12/2014
<input type="checkbox"/>	<input type="star"/>	<input checked="" type="square"/>	Booking.com Nicolas - you qualify for at least 20% off places to stay	26/12/2014
<input type="checkbox"/>	<input type="star"/>	<input checked="" type="square"/>	Communauté d'entraide Gr. Nicolas, des questions sur vos produits ?	25/12/2014
<input type="checkbox"/>	<input type="star"/>	<input type="square"/>	Dernières offres Tous les derniers avis et recommandations et de sites	25/12/2014

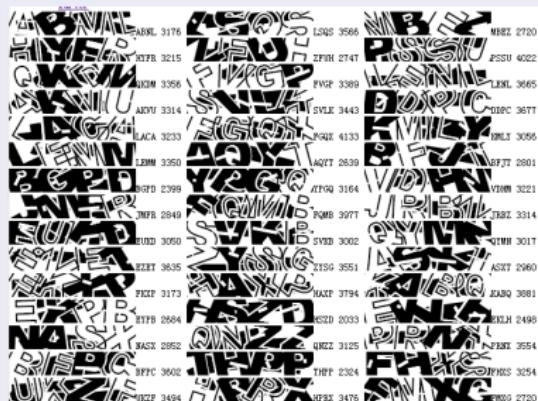
En texte toujours

Reconnaissance de chiffres

82944649709295159123
23591762822507497832
11836103100112730465
26471899307102035465

Ou de captcha

[Yann et al. 08], Newcastle University



Characters under typical distortions



Recognition rate

~100%

96+%

100%

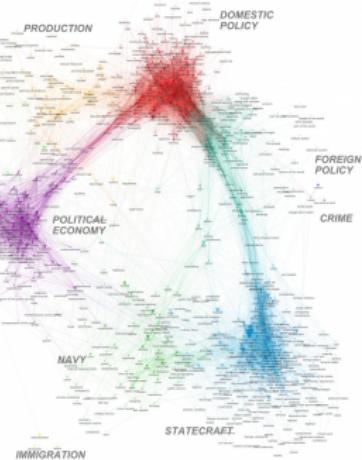
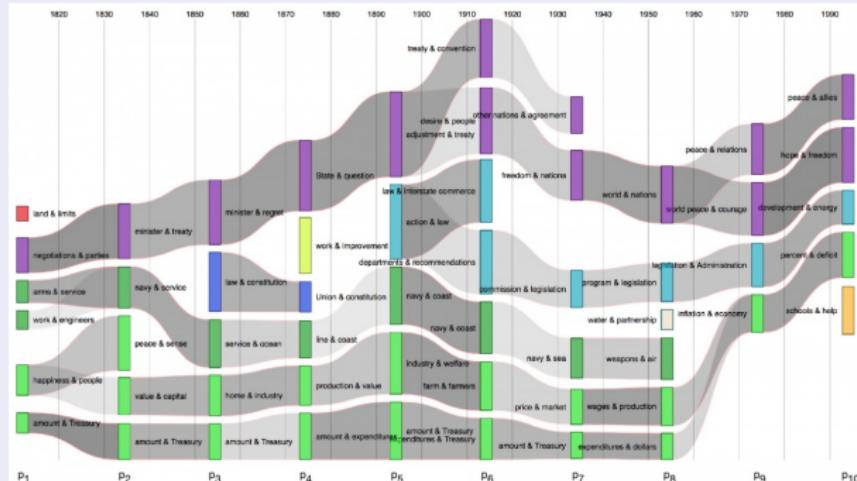
98%

~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 : 0c0cd0 (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

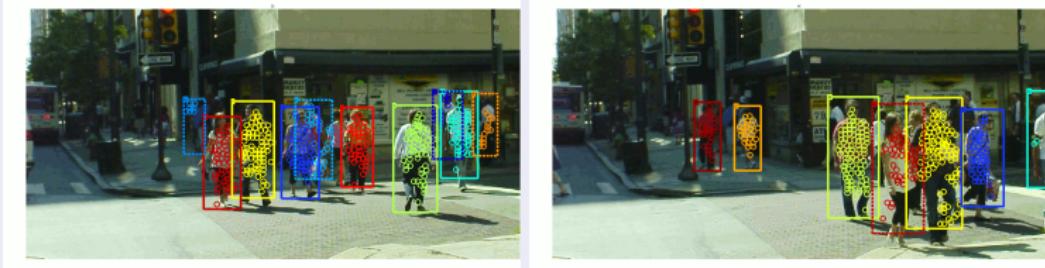
Détection d'objets

teradeep.com, Purdue University



Tracking

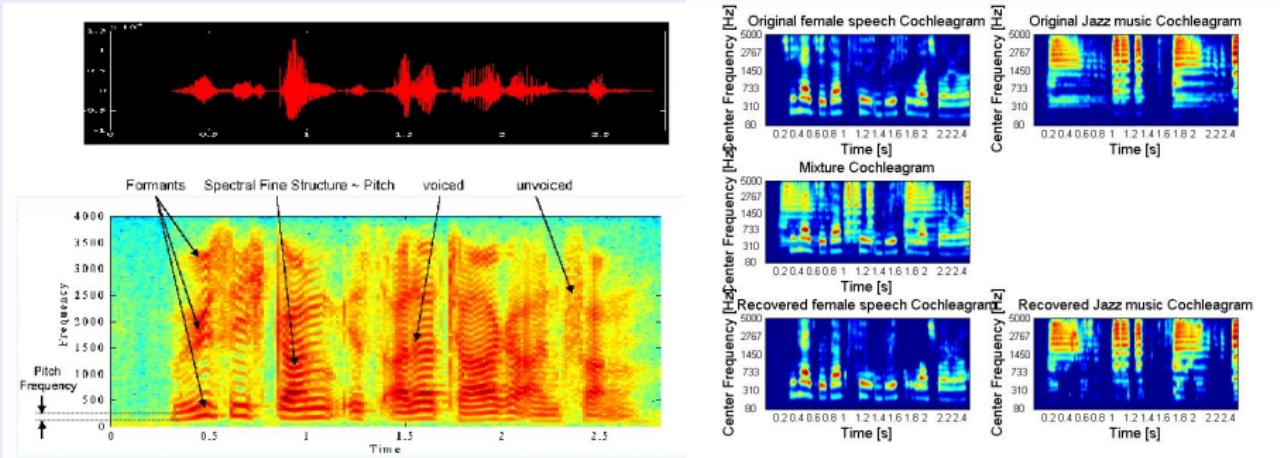
[Fragkiadaki et al. 12], Pennsylvania University



Et l'audio ...

Reconnaissance de la parole, séparation de sources

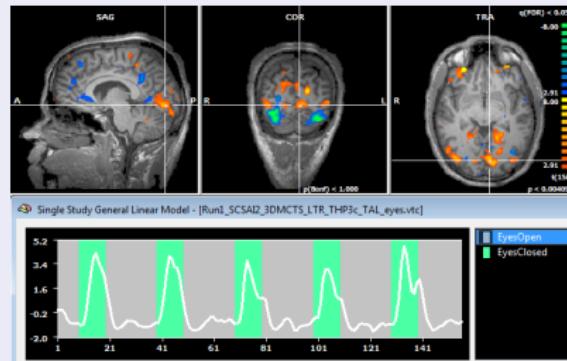
<http://markus-hauenstein.de>



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

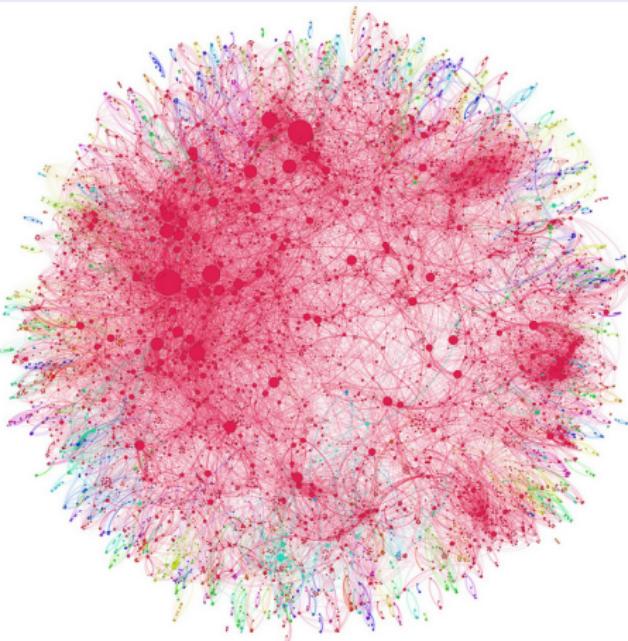
Largest Diffusion Network



Legend
Retweet
@Mention
Click to Enlarge

Meme Activity

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



Matchmaking

de profils, sites de rencontre



Experts, CV - Emplois, Jeux



LinkedIn
viadeo



Systèmes de recommandation

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

The screenshot shows a user interface for a recommendation engine. At the top, there's a navigation bar with links for 'Recommendation Engine', 'Home', 'Item store', 'My Lists', and a user email 'mehdi.circulaire@univ-tours.fr'. On the right, there are 'Logout' and 'Search All' buttons.

Similar Artists

A list of artists with small thumbnail images next to them. A red arrow points to 'Radiohead' in the list:

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

Movies

A grid of movie posters with titles and details below them. Each poster has a 'Select' button to its right:

Movie Title	Director	Year	Rating
Umbrellas of Cherbourg	Jacques Demy	1964	Unrated
Brokedown Palace	Jonathan Kaplan	1998	PG-13
West Beirut	Ziad Doueiri	2005	PG-13
Suspect	Peter Yates	Thriller 1987 R	
Heights	Jeremy Kagan	Drama 2004 R	
Babylon A.D.	Mathieu Kassovitz	Action 2008 PG-13	

amazon.com

Recommended for You

Amazon.com has new recommendations for you based on items you purchased or told us you own.

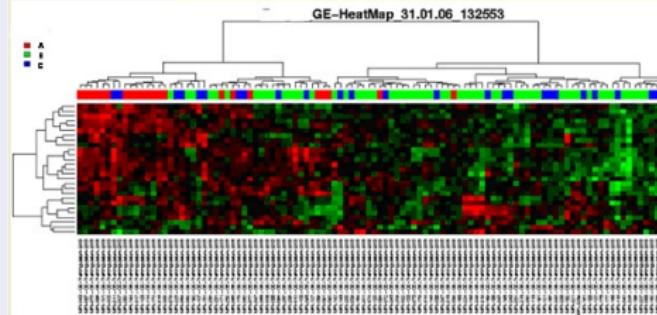


Et dans d'autres sciences

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.



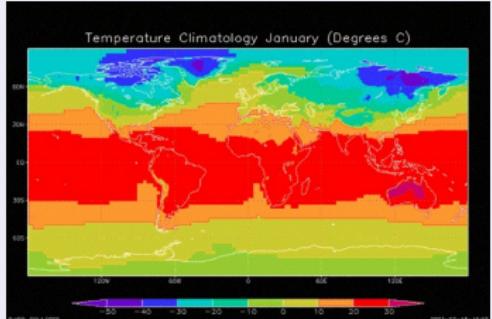
Économie



Astronomie



Climatologie (complémentation données)

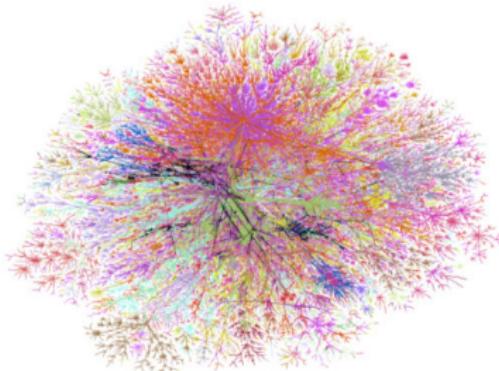


Dans les jeux et la robotique



L'apprentissage aujourd'hui : Big Data

- Webpages (content, graph)
- Clicks (ad, page, social)
- Users (OpenID, FB Connect)
- e-mails (Hotmail, YM!Mail, Gmail)
- Photos, Movies (Flickr, YouTube, Vimeo ...)
- Cookies / tracking info (see Ghostery)
- Installed apps (Android market etc.)
- Location (Latitude, Loopt, Foursquare)
- User generated content (Wikipedia & co)
- Ads (display, text, DoubleClick, Yahoo)
- Comments (Disqus, Facebook)
- Reviews (Yelp, Y!Local)
- Third party features (e.g. Experian)
- Social connections (LinkedIn, Facebook)
- Purchase decisions (Netflix, Amazon)
- Instant Messages (YM!, Skype, Gtalk)
- Search terms (Google, Bing)
- Timestamp (everything)
- News articles (BBC, NYTimes, Y!News)
- Blog posts (Tumblr, Wordpress)
- Microblogs (Twitter, Jaiku, Meme)



>10B useful webpages

Carnegie Mellon University

extrait du cours d'A. Smola

Entreprises concernées :

Yahoo, Google, Amazon, Netflix, Microsoft, Apple, Xerox, Samsung, Critéo, Facebook, Twitter, Flickr, Instagram, Reddit, Valve, Steam, Deezer, Dailymotion, Youtube, SNCF, AXA, EDF, GDF-Suez, Veolia, Safran, Thalès, les médias, ...

Plan

- 1 Organisation de l'UE
- 2 Introduction
- 3 Les problématiques générales
- 4 Premier exemple : classification de films et arbres de décision

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

Les grandes familles

Apprentissage supervisé

- Classification
- Régression
- Forecasting
- Compléction 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

Méthodes jusqu'aux années 2010 !

- Problématiques générales (biais, variance, évaluation, sur-apprentissage, représentation des données)
- Algorithmes supervisées (k-nn, bayésien, perceptron, réseaux de neurones, svm, ...)
- Algorithmes non supervisées (hiérarchique, k-means, ...)
- Pas de méthodes Deep ⇒ en M2 !

Objectifs

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

L'apprentissage statistique est au cœur de la formation d'un *data scientist*

Data scientist map

Chandrasekaran, 2013]



Plan

- 1 Organisation de l'UE
- 2 Introduction
- 3 Les problématiques générales
- 4 Premier exemple : classification de films et arbres de décision

Formalisation de l'apprentissage supervisé

On dispose :

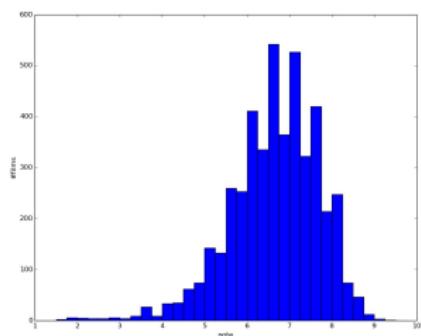
- d'un espace de représentation \mathcal{X} , usuellement \mathbb{R}^n
 n est la dimension de l'espace de représentation,
chaque dimension = un attribut
- d'un ensemble d'exemples X décrit dans cette espace :
 $x \in X, x = (x_1, x_2, x_3, \dots, x_n)$
- d'un ensemble d'étiquettes/labels Y décrivant les classes d'intérêt
quand Y contient deux classes \rightarrow classification binaire, usuellement
 $Y = \{0, 1\}$ ou $Y = \{-1, 1\}$
- pour chaque exemple x^i de X , son étiquette y^i
 \Rightarrow ensemble d'apprentissage $E = \{(x^i, y^i)\}$

On veut :

Trouver une fonction $f : \mathcal{X} \rightarrow Y$ telle que la prédiction sur de futurs exemples soit la plus précise possible.

Première étape : choix de la représentation

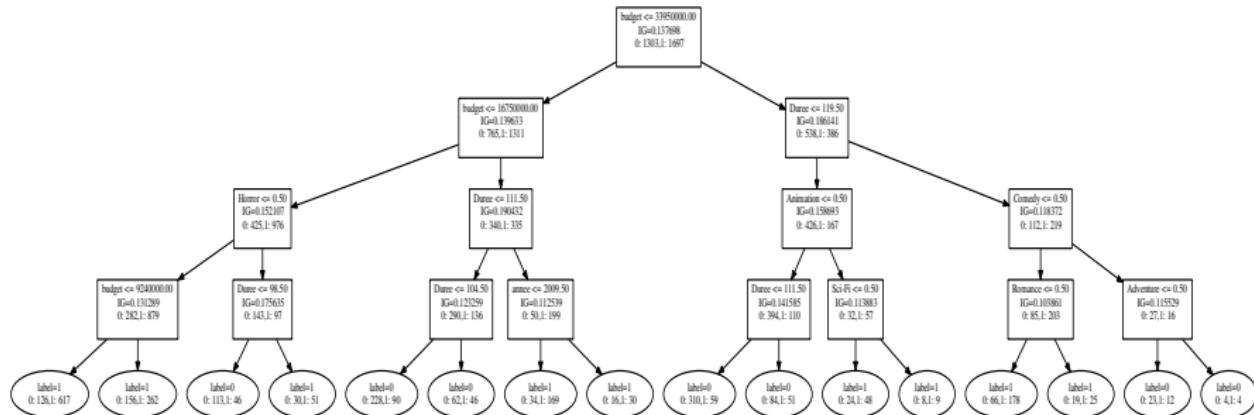
Collection de films, des notes à chaque film



Questions

- Comment représenter un film ?
- Comment classifier ?

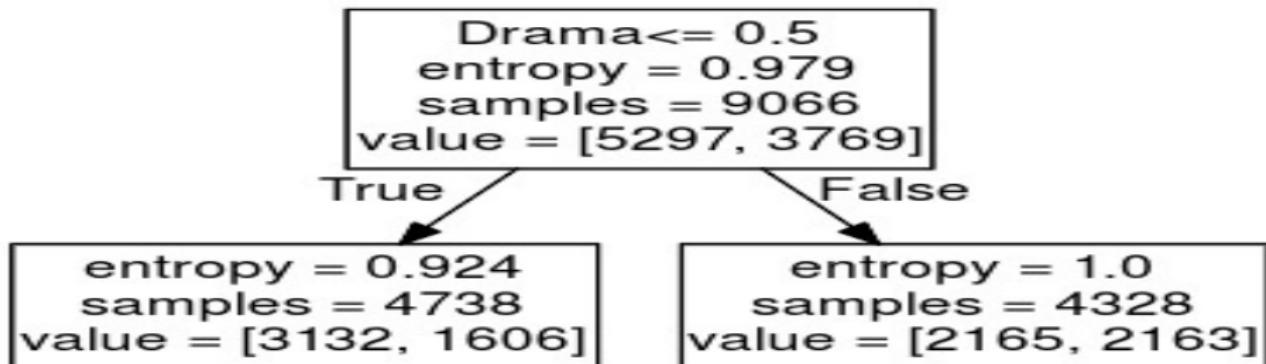
Arbres de décision



Principe

- Chaque nœud interne : un test sur une des dimensions de \mathcal{X}
⇒ Est-ce que le film appartient au genre *Comédie*
- Chaque branche : un résultat du test
- Chaque feuille : un label de Y
⇒ classification en parcourant un chemin de la racine à une feuille.

Apprentissage d'un arbre de décision

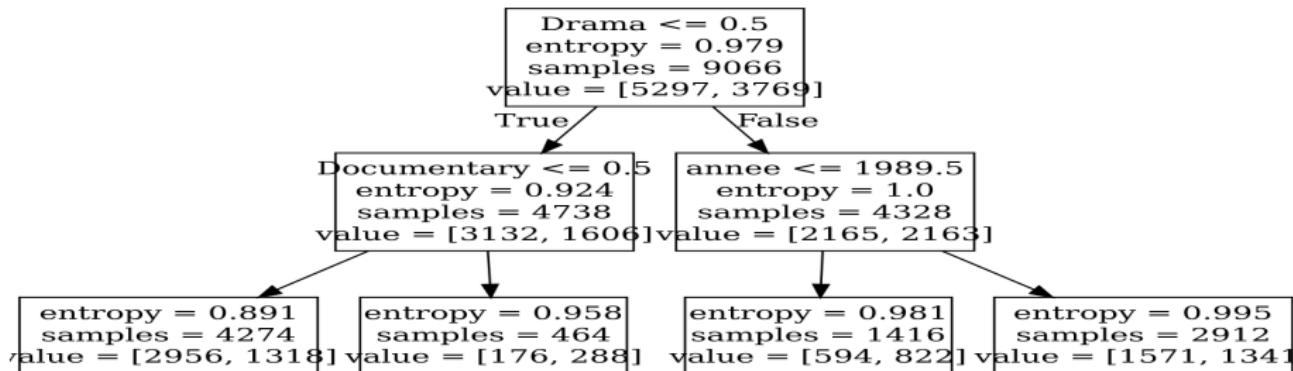


Algorithme glouton, top-down

Initialisation à la racine, considérer tous les exemples

- Si le nœud n'est pas pur, alors
 - ▶ Trouver x_i le "meilleur" attribut pour ce nœud et le cas test associé
 - ▶ Pour chaque test, créer un fils au nœud courant
 - ▶ Faire "tomber" les exemples du nœud courant à leur fils correspondant
- sinon transformer le nœud en feuille.

Apprentissage d'un arbre de décision

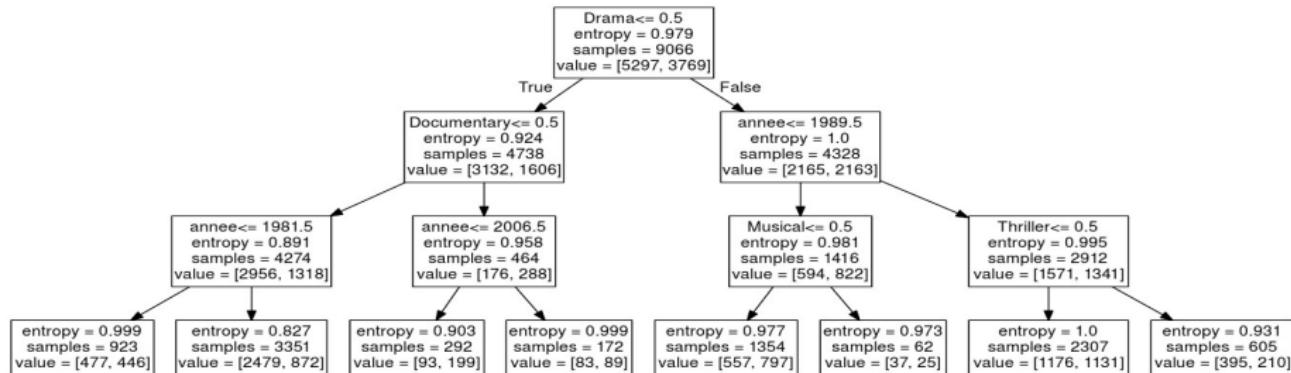


Algorithme glouton, top-down

Initialisation à la racine, considérer tous les exemples

- Si le nœud n'est pas pur, alors
 - ▶ Trouver x_i le "meilleur" attribut pour ce nœud et le cas test associé
 - ▶ Pour chaque test, créer un fils au nœud courant
 - ▶ Faire "tomber" les exemples du nœud courant à leur fils correspondant
- sinon transformer le nœud en feuille.

Apprentissage d'un arbre de décision

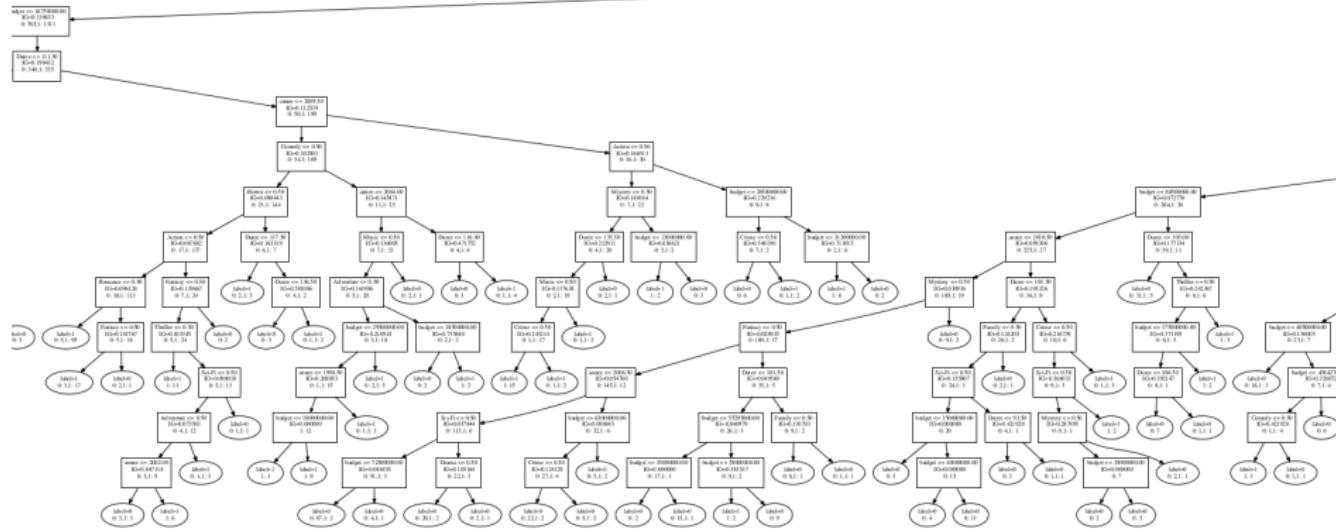


Algorithme glouton, top-down

Initialisation à la racine, considérer tous les exemples

- Si le nœud n'est pas pur, alors
 - ▶ Trouver x_i le "meilleur" attribut pour ce nœud et le cas test associé
 - ▶ Pour chaque test, créer un fils au nœud courant
 - ▶ Faire "tomber" les exemples du nœud courant à leur fils correspondant
- sinon transformer le nœud en feuille.

Apprentissage d'un arbre de décision

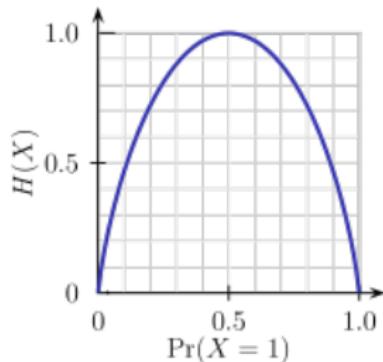


Algorithme glouton, top-down

Initialisation à la racine, considérer tous les exemples

- Si le nœud n'est pas pur, alors
 - ▶ Trouver x_i le "meilleur" attribut pour ce nœud et le cas test associé
 - ▶ Pour chaque test, créer un fils au nœud courant
 - ▶ Faire "tomber" les exemples du nœud courant à leur fils correspondant

Sélectionner le meilleur attribut



Entropie d'une variable aléatoire

Soit X une variable aléatoire pouvant prendre n valeurs x_i :

$$H(X) = - \sum_{i=1}^n P(X = x_i) \log(P(X = x_i))$$

Plus l'entropie est grande, plus le désordre est grand.
Entropie nulle \rightarrow pas d'aléa.

Sélectionner le meilleur attribut

Entropie d'un échantillon : cas binaire

- X un ensemble de données, Y leur étiquette (positif/négatif)
- p_+ la proportion d'exemples positifs
- p_- la proportion d'exemples négatifs
- $H(Y) = -p_+ \log(p_+) - p_- \log(p_-)$

Entropie conditionnelle

- Entropie conditionnelle : $H(Y|X) = \sum_i P(X = x_i)H(Y|X = x_i)$
- Dans notre cas, en faisant un test T sur un des attributs, on obtient deux partitions d'exemples : Y_1 qui vérifie le test et Y_2 qui ne vérifie pas le test.
L'entropie conditionnelle au test T est :

$$H(Y|T) = \frac{|Y_1|}{|Y|}H(Y_1) + \frac{|Y_2|}{|Y|}H(Y_2)$$

⇒ Gain d'information : $I(T, Y) = H(Y) - H(Y|T)$ à maximiser (donc $H(Y|T)$ à minimiser)

Où s'arrêter ? Est ce un bon modèle ?

