

Apprentissage non supervisé

Cours 8
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 Introduction

2 Formalisation

3 Clustering agglomératif

4 K-means

5 Interlude : Gaussiennes multivariées

6 Spectral clustering

7 Réduction de dimensions

Que faire sans label de disponible ...

Pourquoi et quand ?

- pas le temps ni l'argent
- pas de spécialiste pour étiquetter
- impossible à étiquetter
- évolution dynamique des structures
- trop de catégories sans beaucoup de sens
- l'important est la structuration des données, les motifs
- on ne sait pas ce qu'on cherche
- ...

Principe

- Regrouper tout ce qui se ressemble,
- Eloigner tout ce qui est franchement différent.
- Un *cluster* : un regroupement de donnée.

Simple, mais ...

Données



Clustering

Principe

- Regrouper tout ce qui se ressemble,
- Eloigner tout ce qui est franchement différent.
- Un *cluster* : un regroupement de donnée.

Simple, mais ...

Données



Clustering



Principe

- Regrouper tout ce qui se ressemble,
- Eloigner tout ce qui est franchement différent.
- Un *cluster* : un regroupement de donnée.

Simple, mais ...

Données



Clustering



Principe

- Regrouper tout ce qui se ressemble,
- Eloigner tout ce qui est franchement différent.
- Un *cluster* : un regroupement de donnée.

Simple, mais ...

Données



Clustering



Principe

- Regrouper tout ce qui se ressemble,
- Eloigner tout ce qui est franchement différent.
- Un *cluster* : un regroupement de donnée.

Simple, mais ...

Données

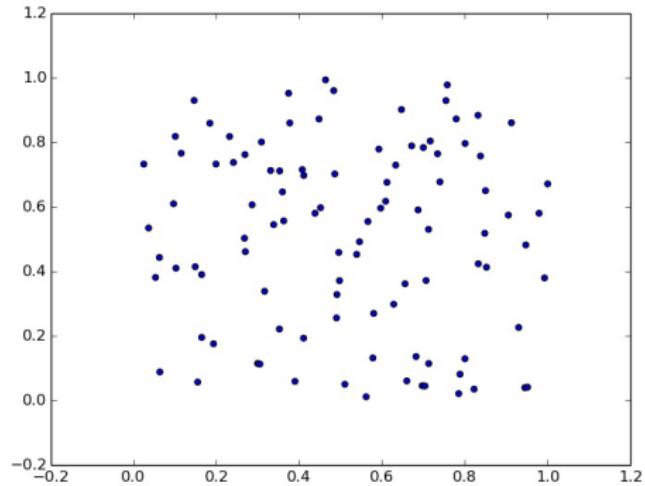


Clustering



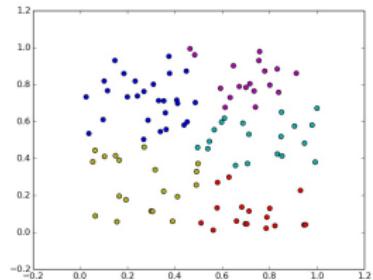
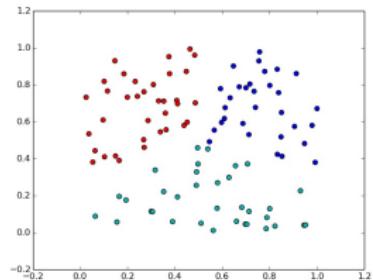
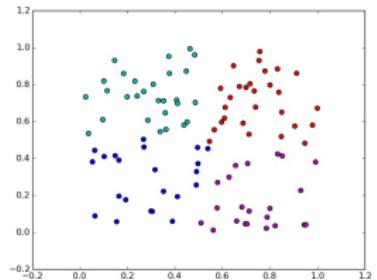
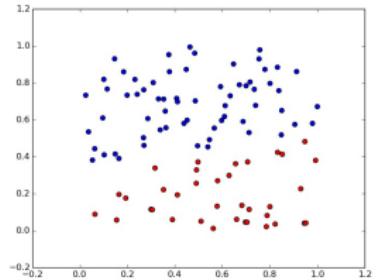
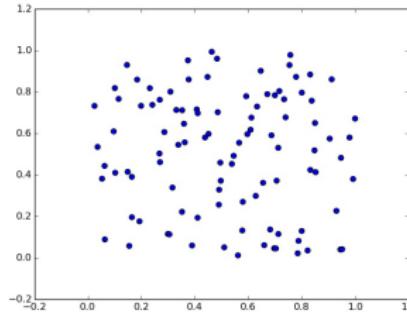
- L'apprentissage non supervisé : très subjectif !
- Pas de but global bien défini, l'objectif est induit par la formulation du problème.

Exemple



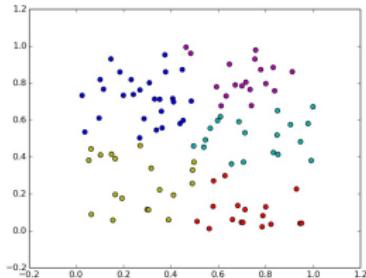
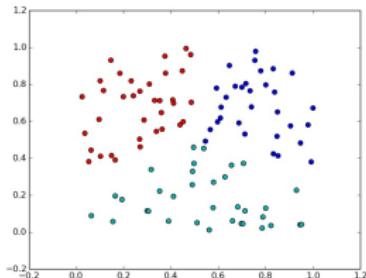
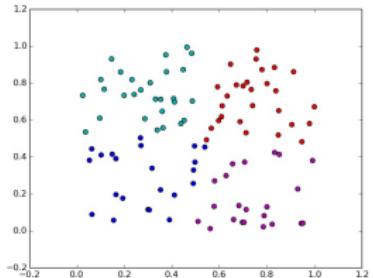
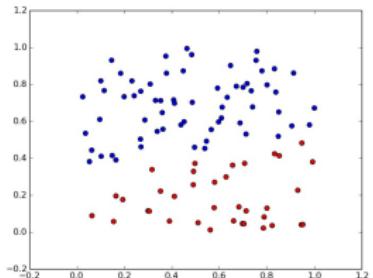
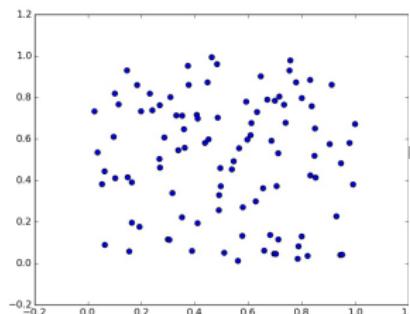
Exemple

Quel est le bon clustering ?



Exemple

Quel est le bon clustering ?



aucun ! (distribution uniforme)

Plan

1 Introduction

2 Formalisation

3 Clustering agglomératif

4 K-means

5 Interlude : Gaussiennes multivariées

6 Spectral clustering

7 Réduction de dimensions

Apprentissage non supervisé

- Ensemble très varié de techniques qui visent à trouver des sous-ensembles cohérents des données
 - Tout ce qui ressemble s'assemble ⇒ définir une *similarité* entre exemples
 - Différentes approches :
 - ▶ Connectivité
 - ▶ Centroïde
 - ▶ Distribution latente
 - ▶ Graphes
 - ▶ modèles bayésiens, graphical models (HMM, CRF, ...)
 - ▶ apprentissage génératif
 - ▶ ...
 - Clustering :
 - ▶ *hard* (un exemple n'appartient qu'à un groupe)
 - ▶ *soft* (probabilité d'appartenance)
- ⇒ Domaine-spécifique, pas de règle générale, tout dépend de la tâche !

Formalisation

Objectif

- Soit $D = x^1, \dots, x^N \in \mathcal{X}$ un ensemble de données
- $\pi = (D_1, D_2, \dots, D_k)$ partition de $D : D = \bigcup D_i$ et (souvent) $D_i \cap_{i \neq j} D_j = \emptyset$
- un critère ϕ dénotant la similarité sur D
- Clustering : trouver $\pi^* = \operatorname{argmin}_{\pi} f(\pi)$ où f est formulée en accord avec ϕ .

Similarité

Plusieurs formes en fonction du problème :

- inverse d'une distance sur l'espace de description \mathcal{X}
- si pas de description dans \mathbb{R}^n , une similarité sur \mathcal{X} peut être définie
- décomposition sur des facteurs latents (processus génératif des exemples)

Plan

- 1 Introduction
- 2 Formalisation
- 3 Clustering agglomératif
- 4 K-means
- 5 Interlude : Gaussiennes multivariées
- 6 Spectral clustering
- 7 Réduction de dimensions

Principe

Algorithme glouton

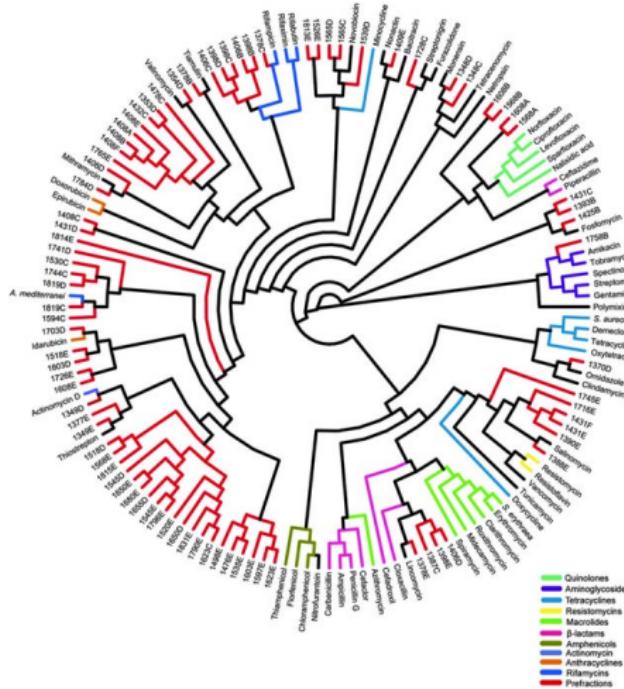
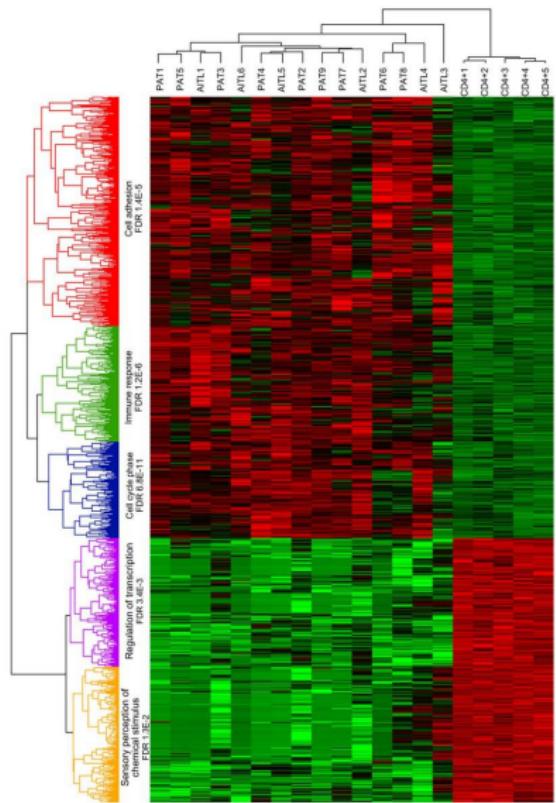
- Fusionner les instances les plus similaires dans un même cluster
- Construire incrémentalement des clusters plus larges en fusionnant les 2 clusters les plus proches
- S'arrêter lorsqu'il n'y a plus qu'un cluster.
⇒ Construit un arbre de partitionnement, *un dendrogramme*.

Distance entre clusters ?

Plusieurs choix, mais souvent pas une vraie distance :

- $d(c_1, c_2) = \min d(x, x')$, $x \in c_1, x' \in c_2$
- $d(c_1, c_2) = \max d(x, x')$, $x \in c_1, x' \in c_2$
- $d(c_1, c_2) = \mathbb{E}(d(x, x'))$, $x \in c_1, x' \in c_2$

Exemples



Plan

- 1 Introduction
- 2 Formalisation
- 3 Clustering agglomératif
- 4 K-means
- 5 Interlude : Gaussiennes multivariées
- 6 Spectral clustering
- 7 Réduction de dimensions

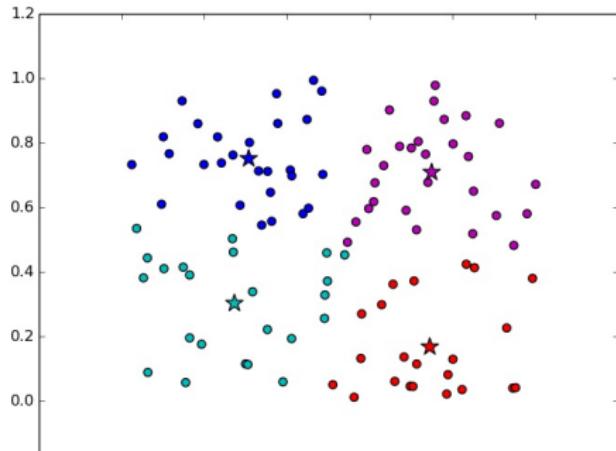
Algorithme des k -means

Objectif

Partitionner les données de manière à minimiser la distance intra-cluster :

$$\operatorname{argmin}_{\pi=(D_1, \dots, D_K)} \sum_{i=1}^K \sum_{x_j \in D_i} \|x_j - \mu_i\|^2$$

avec μ_i le centre du cluster i , i.e. $\mu_i = \frac{1}{|D_i|} \sum_{x_j \in D_i} x_j$



Algorithme des k -means

Objectif

Partitionner les données de manière à minimiser la distance intra-cluster :

$$\operatorname{argmin}_{\pi=(D_1, \dots, D_K)} \sum_{i=1}^K \sum_{x_j \in D_i} \|x_j - \mu_i\|^2$$

avec μ_i le centre du cluster i , i.e. $\mu_i = \frac{1}{|D_i|} \sum_{x_j \in D_i} x_j$

Remarques

- Problème NP -difficile
- Chaque centre est un représentant du cluster : *prototype*
- Très proche d'une notion de compression
- Provient des méthodes de quantification (traitement du signal).

Résolution approchée

Algorithme en deux étapes

- ➊ Affecter chaque point à un cluster en fonction des centres μ_i
- ➋ Ré-estimer le centre μ_i de chaque cluster en fonction de la nouvelle répartition

Et boucler jusqu'à convergence.

Complexité en $O((K + 1)n)$

Initialisation

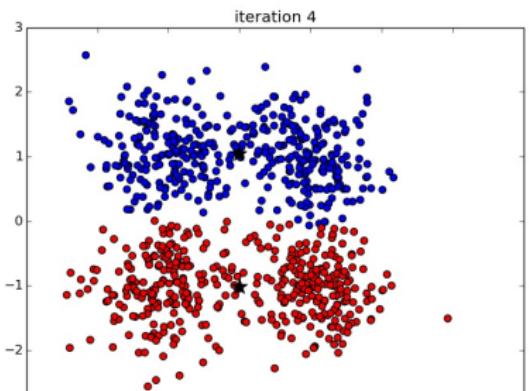
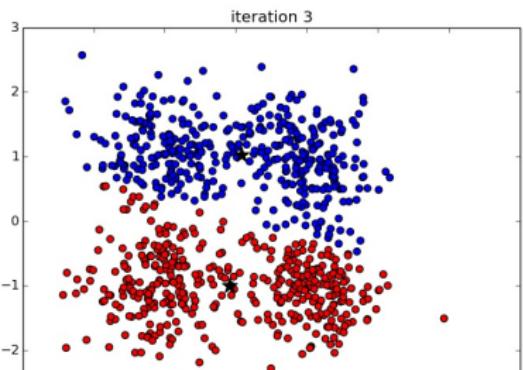
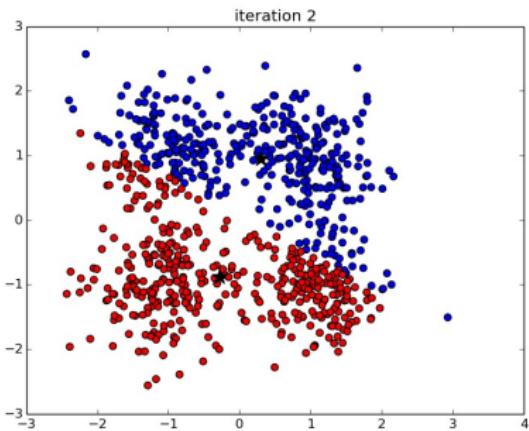
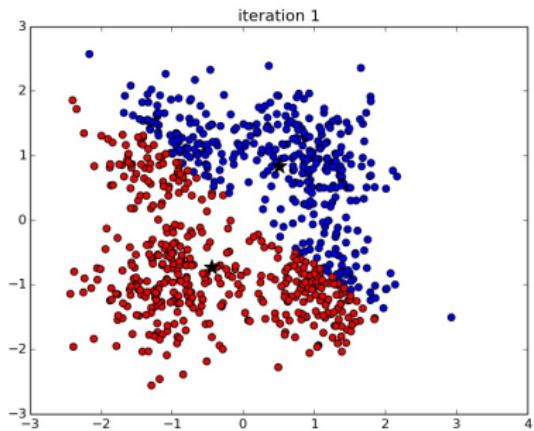
- Algorithme très sensible à l'initialisation
- Coincé dans un minimum local très souvent \Rightarrow multiples tentatives et prendre la meilleure

Détails

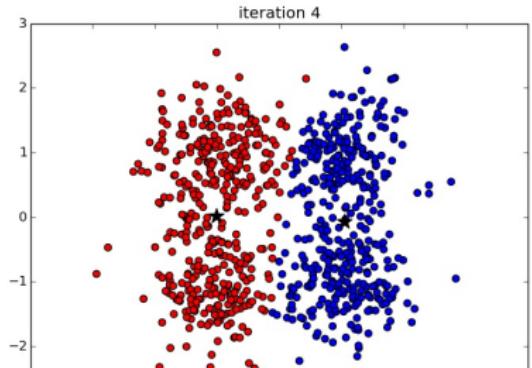
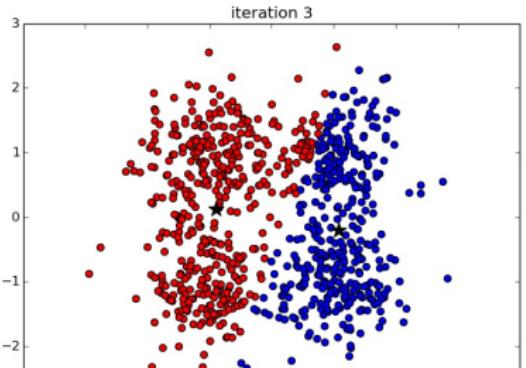
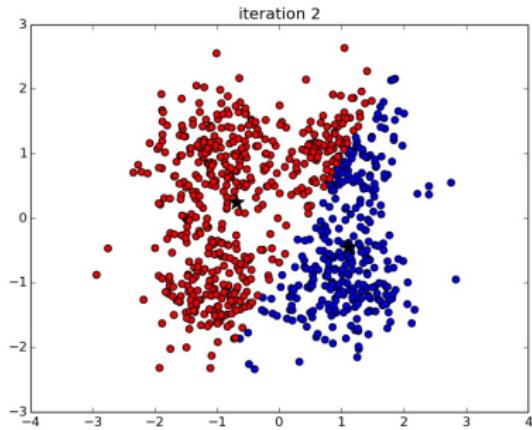
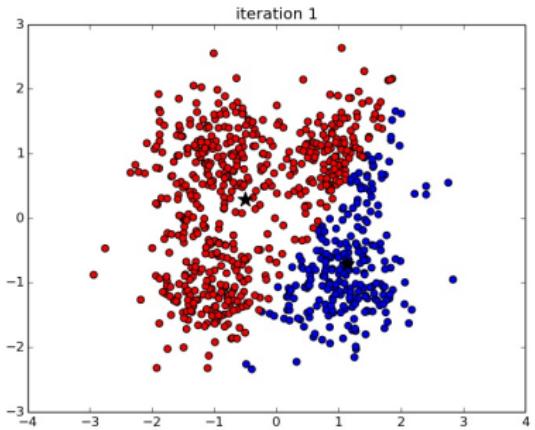
- Initialisation aléatoire des k centres : $\mu^0 = (\mu_1^0, \dots, \mu_K^0)$
- Affectation des points x_j : classe $C^t(x_j) = \operatorname{argmin}_i \|\mu_i^t - x_j\|^2$
- Estimation des centres : $\mu_i^{t+1} = \sum_{j:C^t(x_j)=i} \|\mu - x_j\|^2$
- On optimise $C^t = (C^t(x_j))$ et $\mu = (\mu_i)$
- Fonction de coût : $F(\mu, C) = \sum_j \|\mu_{C(x_j)} - x_j\|^2 = \sum_{i=1}^K \sum_{j:C(x_j)=i} \|\mu_i - x_j\|^2$
- Première étape : on fixe μ , on optimise $C \Rightarrow$ Calcul de l'espérance
- Seconde étape : on fixe C , on optimise $\mu \Rightarrow$ Calcul du maximum de vraisemblance

Algorithme dit *Expectation/Maximization, (EM)*

Exemples



Exemples



Utilisation pour la segmentation et la compression

Par moyennage des couleurs : nombre de couleurs = nombre de clusters



Autre approche, 2 couleurs par cellule :

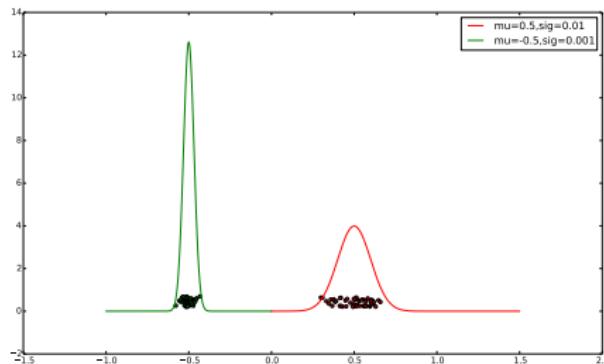


Plan

- 1 Introduction
- 2 Formalisation
- 3 Clustering agglomératif
- 4 K-means
- 5 Interlude : Gaussiennes multivariées
- 6 Spectral clustering
- 7 Réduction de dimensions

Rappel : distribution gaussienne

- En 1d : $p(x) = \mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{(-\frac{1}{2\sigma^2}(x-\mu)^2)}$



Remarque : à quoi sert la constante : $\frac{1}{(2\pi\sigma^2)^{1/2}}$?

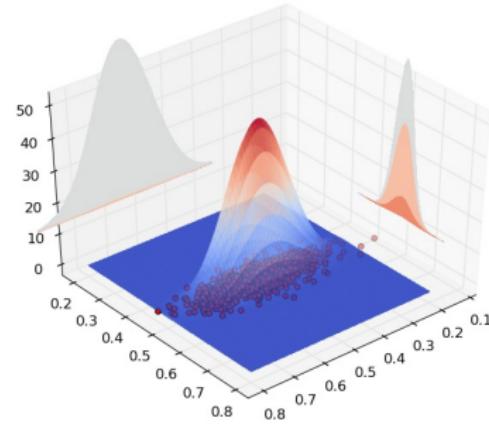
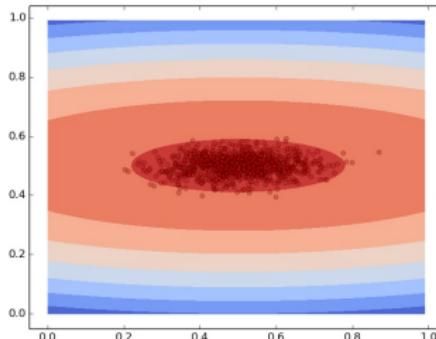
- Multivariée en d dimensions:

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)' \Sigma^{-1} (x - \mu)\right)$$

- $\mu = (\mu_1, \mu_2, \dots, \mu_d)$, mais Σ ?

Gaussienne 2D : cas simple

- En 2d : on suppose que $x_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ et $x_2 \sim \mathcal{N}(\mu_2, \sigma_2)$
- hypothèse Naive Bayes, x_1 indépendant de x_2
- $p(x) = p(x|\mathcal{N}_1)p(x|\mathcal{N}_2) = \frac{1}{(2\pi\sigma_1^2)^{1/2}} e^{-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2} \frac{1}{(2\pi\sigma_2^2)^{1/2}} e^{-\frac{1}{2\sigma_2^2}(x_2-\mu_2)^2}$
- $p(x) = \frac{1}{(2\pi)^{2/2}(\sigma_1^2\sigma_2^2)^{1/2}} e^{-\frac{1}{2}(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2})} = \frac{1}{2\pi\Sigma^{1/2}} e^{-\frac{1}{2}(x-\mu)' \Sigma^{-1} (x-\mu)}$
avec $\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$



Gaussienne 2D : cas générique

Transformation affine

- Supposons $x_1, x_2 \sim \mathcal{N}(0, 1)$ et $X = (x_1, x_2)$;
- Soit T une transformation affine inversible $T = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix}$
- Soit $Y = TX + \mu$, $y_1 = t_{11}x_1 + t_{12}x_2 + \mu_1$, $y_2 = t_{21}x_1 + t_{22}x_2 + \mu_2$
- alors $\mathbb{E}(Y) = \begin{pmatrix} \mathbb{E}(t_{11}x_1 + t_{12}x_2 + \mu_1) \\ \mathbb{E}(t_{21}x_1 + t_{22}x_2 + \mu_2) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$
- Variance d'un vecteur aléatoire ?

Covariance

- Covariance de deux variables aléatoires :
$$\text{Cov}(x, y) = \mathbb{E}((x - \mathbb{E}(x))(y - \mathbb{E}(y))) = \mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y)$$
- Matrice de covariance d'un vecteur aléatoire X , $\text{Cov}(X)$:
$$\begin{pmatrix} \text{Cov}(x_1, x_1) & \cdots & \text{Cov}(x_1, x_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(x_n, x_1) & \cdots & \text{Cov}(x_n, x_n) \end{pmatrix} = \mathbb{E}((X - \mu)(X - \mu)') = \mathbb{E}(XX') - \mu\mu'$$

Gaussienne 2D : cas générique

Covariance de $Y = TX + \mu$: $\text{Cov}(Y) = TT'$

$$\bullet \text{Cov}(Y) = \begin{pmatrix} \mathbb{E}((t_{11}x_1 + t_{12}x_2)^2) & \mathbb{E}((t_{11}x_1 + t_{12}x_2)(t_{21}x_1 + t_{22}x_2)) \\ \mathbb{E}((t_{11}x_1 + t_{12}x_2)(t_{21}x_1 + t_{22}x_2)) & \mathbb{E}((t_{21}x_1 + t_{22}x_2)^2) \end{pmatrix}$$
$$= \begin{pmatrix} t_{11}^2 + t_{12}^2 & t_{11}t_{21} + t_{12}t_{22} \\ t_{11}t_{21} + t_{12}t_{22} & t_{21}^2 + t_{22}^2 \end{pmatrix}$$

On note $\Sigma = \text{Cov}(Y)$

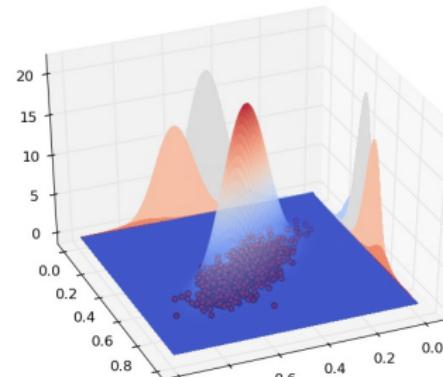
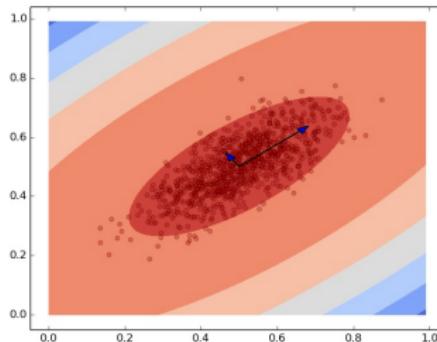
Changement de variable

- $p(x) = \frac{1}{2\pi\Sigma_{\mathcal{N}(0,1)}^{1/2}} e^{-\frac{1}{2}(x - \mu_{\mathcal{N}(0,1)})' \Sigma_{\mathcal{N}(0,1)}^{-1} (x - \mu_{\mathcal{N}(0,1)})}$, avec $\mu_{\mathcal{N}(0,1)} = 0, \Sigma_{\mathcal{N}(0,1)} = I$
- Si $Y = TX + \mu$, alors $p(Y) = \frac{1}{|det(T)|} p(T^{-1}(Y - \mu))$
- $p(Y) = \frac{1}{2\pi|\Sigma|^{-1/2}} e^{-\frac{1}{2}((T^{-1}(Y-\mu))' IT^{-1}(Y-\mu))} = \frac{1}{|\Sigma|^{-1/2}2\pi} e^{-\frac{1}{2}(Y-\mu)' T'^{-1} T^{-1} (Y-\mu)}$
- $p(Y) = \frac{1}{2\pi|\Sigma|^{-1/2}} e^{\frac{1}{2}(Y-\mu)' \Sigma^{-1} (Y-\mu)}$

Gaussienne 2D : interprétation géométrique

Transformation affine inversible

- T peut être décomposé en $T = UD$, D diagonale (valeurs propres) et U orthogonale (vecteurs propres, matrice de rotation et reflexion)
 - $\Sigma = UD(UD)' = UDD'U' = UD^2U'$
 - $Det(\Sigma) = Det(UD^2U') = Det(D^2) = \sum_i \sigma_i^2$, σ_i valeurs propres de T
- ⇒ Loi normale multivariée : D représente la variance sur chaque composante normale indépendante des autres,
 U représente la rotation/reflexion par rapport aux axes.

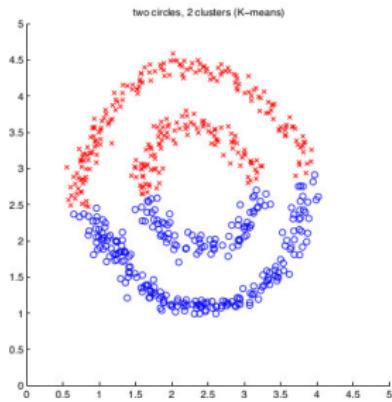


Plan

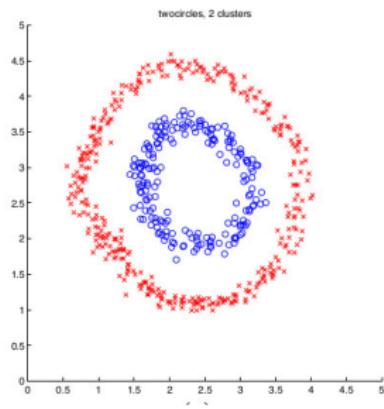
- 1 Introduction
- 2 Formalisation
- 3 Clustering agglomératif
- 4 K-means
- 5 Interlude : Gaussiennes multivariées
- 6 Spectral clustering
- 7 Réduction de dimensions

Problématique

K-means



Spectral clustering



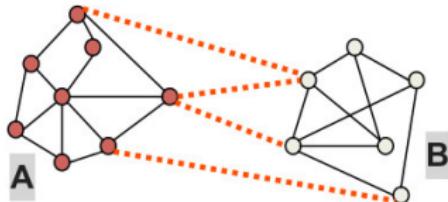
Limites des approches vues

- K-means (et en général clustering de métrique) ne trouvent que des clusters sphériques
- Comment encoder une structuration des données ? des relations de voisinages ?
- Une solution parmi d'autres : spectral clustering \Rightarrow projeter les données sur un graphe de relation

Graphe de données

Notations graphe

- Données : les nœuds $V = \{x_i\}$ du graphe
- Les liens/arêtes pondérés : $E = \{w_{ij} = s(x_i, x_j)\}$ similarité entre données
- Restriction : graphe connexe

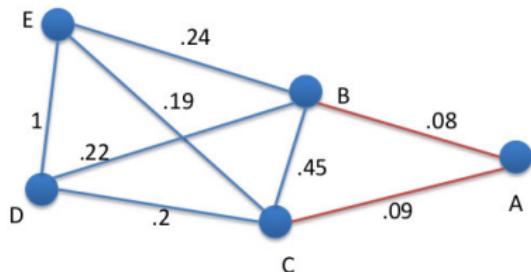


Création du graphe

- Difficile de travailler sur un graphe entièrement connecté :
 - ▶ seuil sur la mesure de similarité
 - ▶ k -nn avec k fixé
- ou utilisation de noyaux pour pondérer les arêtes : $w_{ij} = e^{-\|x_i - x_j\|^2 / \sigma^2}$

Objectif

- Toujours le même :
 - ▶ données d'un même cluster très similaires
 - ▶ données de différent cluster dissimilaires
- En termes de graphe :
 - ▶ Notion de coupe : $cut(C_1, C_2) = \sum_{i \in C_1, j \in C_2} w_{ij}$, $C_1 \cap C_2 = \emptyset$
 - ▶ Coupe normalisé : $NormCut(C_1, C_2) = \frac{Cut(C_1, C_2)}{Vol(C_1)} + \frac{Cut(C_1, C_2)}{Vol(C_2)}$,
 $Vol(C) = \sum_{i,j \in C} w_{ij}$
- Problème NP-difficile...

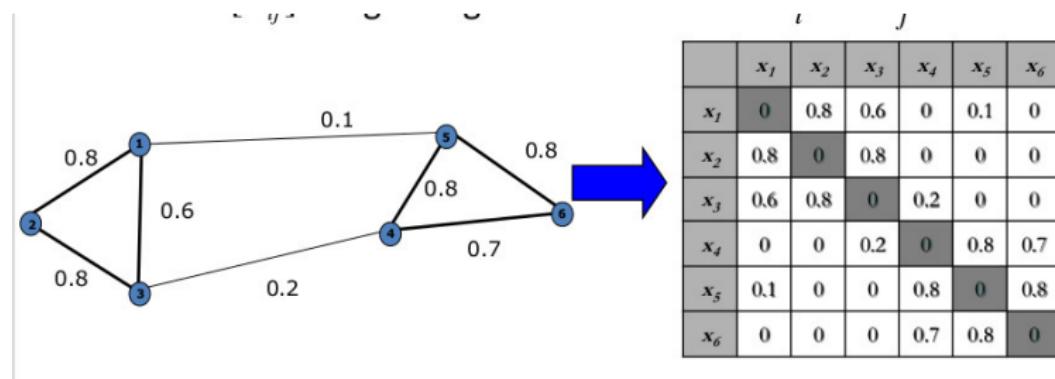


$$Cut(BCDE, A) = 0.17$$

$$NormCut(BCDE, A) = 1.067, NormCut(ABC, DE) = 1.038$$

Representation matricielle

- Matrice de similarité/d'adjacence: $N \times N$,
- $W : \{w_{i,j}\}$
- Matrice symétrique
- D matrice des degrés : $d_{i,i} = \sum_j w_{ij}$ pour normaliser la matrice d'adjacence



Représentation matricielle

- Matrice considérée : matrice laplacienne : $L = D - W$
- Propriétés :
 - ▶ Valeurs propres positives
 - ▶ Vecteurs propres orthogonaux
 - ▶ Ce sont des indicateurs de la connectivité du graphe
- Interprétation : u vecteur binaire de taille n ,
 - ▶ Lu : poids des connections sortantes des noeuds exprimés par u .
 - ▶ $u'L$: poids des connections entrantes des noeuds exprimés par u
- Pour deux partitions C_1, C_2 :
 - ▶ soit f un vecteur dans $\{-1, 1\}$ de taille n , tel que $f_i = 1$ si $i \in C_1$, -1 si $i \in C_2$.
 - ▶ $f'Lf = \sum_{i,j} w_{ij}(f_i - f_j)^2$:
$$\begin{aligned} f'Lf &= f'(D - W)f = f'Df - f'Wf = \sum_i d_i f_i^2 - \sum_{i,j} f_i f_j w_{ij} \\ &= \frac{1}{2} (\sum_i (\sum_j w_{ij}) f_i^2 - 2 \sum_{i,j} f_i f_j w_{ij} + \sum_j (\sum_i w_{ij}) f_j^2) = \frac{1}{2} \sum_{i,j} w_{i,j} (f_i - f_j)^2 \end{aligned}$$
- Objectif : trouver f qui minimise $f'Lf$ tel que $f'Df = 1$
- Relaxation : $\min_f f'Lf$ tel que $f'Df = 1 \Rightarrow Lf = \lambda Df$: deuxième vecteur propre.

Conclusions générales

Questions à se poser

- Qu'est-ce qu'un cluster ?
- Comment définir la similarité ?
- Quels features, quelle normalisation ?
- Combien de clusters ?
- Quelle méthode de clustering ?
- Les résultats ont-ils un sens ? clusters valides ?

Plan

- 1 Introduction
- 2 Formalisation
- 3 Clustering agglomératif
- 4 K-means
- 5 Interlude : Gaussiennes multivariées
- 6 Spectral clustering
- 7 Réduction de dimensions

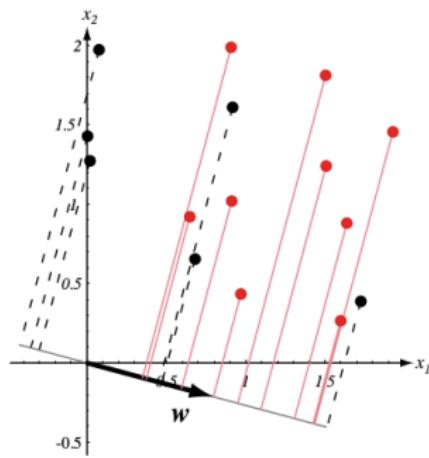
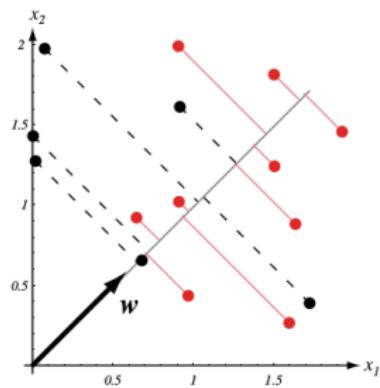
Problématique

- Ajouter des dimensions \Rightarrow augmente l'expressivité
- Trop de dimension : peu de variance sur une même dimension, problème sur-appris, sujet au bruit, . . .
- En texte, image (et ailleurs) : trop de dimensions \Rightarrow dimensions peu informatives, très corrélées
- Découverte d'un espace latent (caché) des données
 \Rightarrow Réduire les dimensions mais sans perte d'informations !
- Objectif : chercher une projection $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ avec $d' \ll d$
- Applications
 - ▶ apprentissage
 - ▶ visualisation
 - ▶ réduction du bruit

Analyse en Composante Principale

Exprimer les données avec le moins de dimensions possibles

- Trouver une nouvelle base vectorielle de faible dimension
- Chaque nouvelle dimension est une combinaison linéaire des anciennes dimensions
- Chaque nouvel axe doit exprimer le plus d'informations possibles des données \Rightarrow maximiser la variance sur cet axe



Analyse en composante principale

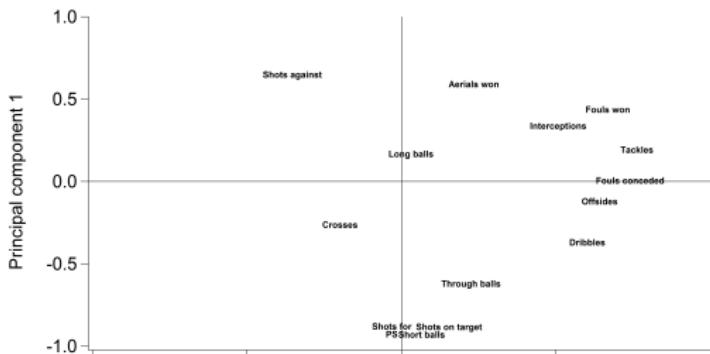
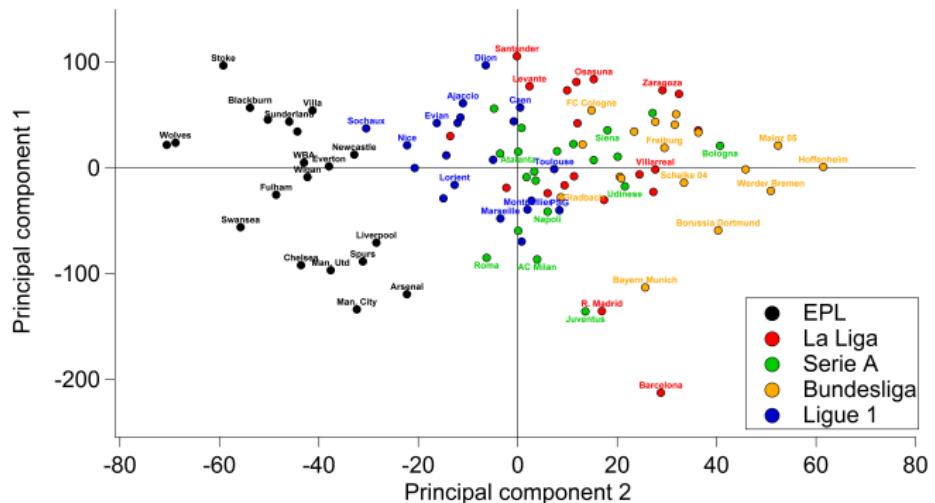
Réduction sur une dimension

- on cherche à projeter sur une ligne $a\mathbf{e} + b$, où \mathbf{e} vecteur unitaire en dimension d
 - Chaque point $x_i : a_i\mathbf{e} + b$
 - $J(X, e) = \sum_i \|a_i\mathbf{e} + b - x_i\|^2 = \sum_i a_i^2 - 2 \sum_i a_i \mathbf{e}'(x_i - b) + \sum_i \|x_i - b\|^2$
 - $\frac{\partial J}{\partial a_i} = 0 \Leftrightarrow a_i = \mathbf{e}'(x_i - b)$
- $\Rightarrow J(X, e) = \sum_i a_i^2 - 2 \sum_i a_i^2 + \sum_i \|x_i - b\|^2 = - \sum_i (\mathbf{e}'(x_i - b))^2 + \sum_i \|x_i - b\|^2$
- avec $S = \sum_i (x_i - b)(x_i - b)'$, $J(X, e) = -e' Se + \sum_i \|x_i - b\|^2$
 - Solution (lagrangien) : $Se = \lambda \mathbf{e}$

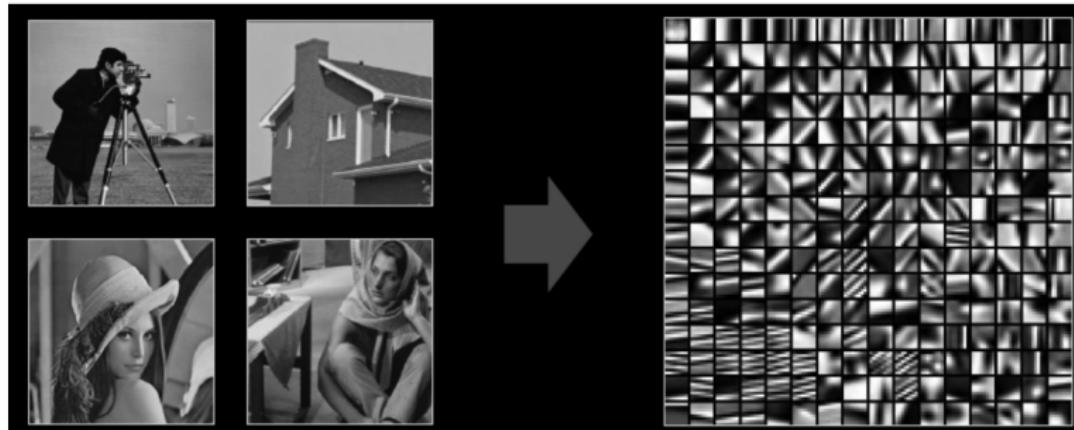
Réduction en plusieurs dimensions

- Même démarche : ensemble des vecteurs propres
- Équivalent à $X - \mu = WSV'$, S diagonale, V vecteurs propres, W coordonnées des points dans le nouveau repère
- Ou à la diagonalisation de la matrice de covariance $(X - \mu)(X - \mu)'$

Exemple



Apprentissage de dictionnaire



- Une image : constituée d'un petit ensemble de primitives.
- Problème de la PCA : base orthogonale ! pas de redondance
- Peut-on apprendre un dictionnaire de primitives pour représenter un jeu de données ?

Compress sensing

- Objectif : trouver D tel que $x \approx Dx'$
- Contrainte de sparsité : $\|x'\|_0$ très faible, peu d'*atomes* sont nécessaires à reconstruire x
 - ▶ simplicité : quelques atomes suffisent à expliquer x
 - ▶ signification : la représentation explique x
 - ▶ parcimonie : x est décrit que parce qui le représente
- Problème d'optimisation : $\operatorname{argmin}_D \|Dx' - x\| + \lambda \|x'\|_0$ (différentes normes, différentes variantes)
- Approche dérivée de la physique (analyse de wavelet, fourier, ...)

Applications



Débruitage



[Mairal et al 2009]

Applications



Since 1699, when French explorers landed at the great bend of the Mississippi River and celebrated the first Mardi Gras in North America, New Orleans has brewed a fascinating mélange of cultures. It was French, then Spanish, then French again, then sold to the United States. Through all these years, and even into the 1900s, others arrived from everywhere: Acadians (Cajuns), Africans, indige-



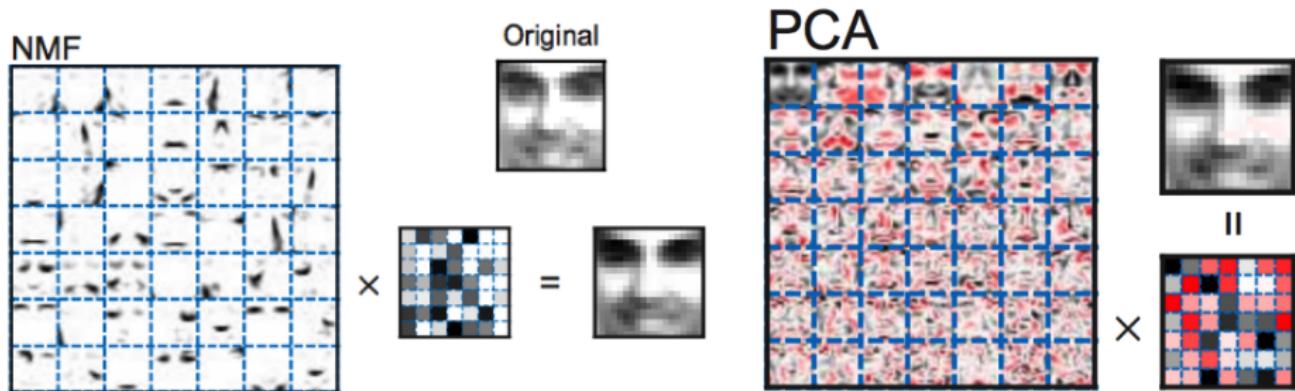
Restored

Inpainting

[Mairal, Elad, Sapiro 2008]

Factorisation matricielle non négative

- Décomposition sur un dictionnaire additif uniquement
- $x \approx Dx'$, avec $x' > 0$
- Intérêt : plus interprétable, plus réaliste sur un ensemble de problèmes

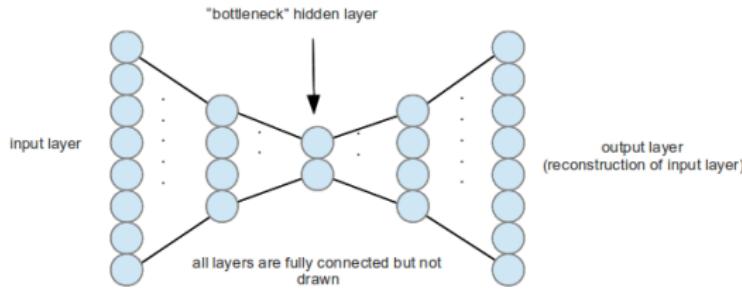


Multiples applications : séparation de sources, en topic discovery, ...

Auto-encoders

Principe

- Apprendre un réseau de neurones qui reconstruit au mieux l'entrée
- Encodage sur une couche cachée \Rightarrow réduction de dimension



Applications

- Visualisation
- Débruitage
- Réduction de dimension, espace latent de représentation