SISY/ATI Machine et deep learning pour l'image

1. Introduction et apprentissage non supervisé

Thomas Oberlin

ISAE-SUPAERO, Département d'ingénierie des systèmes complexes (DISC)

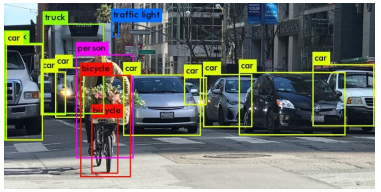
thomas.oberlin@isae-supaero.fr

Machine / deep learning pour l'image

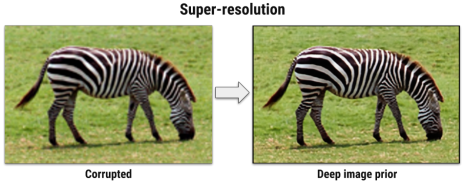
Objectifs du cours

- découvrir les principales problématiques de l'apprentissage automatique
- étudier et savoir manipuler les algorithmes classiques associés
- en particulier, les réeaux de neurones profonds avec PyTorch

Machine learning par et pour les images



Computer Vision (vision artificielle) : extraire du sens à partir d'images



Traitement d'images : les outils de ML/DL sont de plus en plus appliqués pour des tâches standard (débruitage)

Machine / deep learning pour l'image

Programme

1. Introduction et ML non-supervisé (TO), 3h
2. ML supervisé : techniques classiques (LG), 4h, BE noté
3. Introduction au deep learning (MA), 3h
4. Deep learning et CNNs (MA), 3h, BE noté
5. Segmentation et détection (TO), 3h
6. Réseaux génératifs et application en restauration (TO), 4h, BE noté

Évaluation

Trois notes de BEs, modalités à voir avec chaque intervenant

Intervenants

- Laurent Guillaume, Airbus Defense and Space
- Michelle Aubrun, Thalès AleniaSpace et IRT Saint-Exupéry
- Thomas Oberlin, ISAE/DISC

Plan de la séance

Intelligence artificielle

Définition (Larousse)

L'intelligence artificielle (IA) est l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine

Repères historique

- Années 50 : Alan Turing
- Années 60 : Marvin Minsky, John McCarth
- Années 80 et 90 : fondements de l'apprentissage
- 1997 : deep blue bat Kasparov

Un champ multidisciplinaire

- Informatique : logique, complexité, calcul distribué
- Mathématiques : statistiques, optimisation, systèmes dynamiques
- Biologie, en particulier computationnelle
- Neurosciences : s'inspirer des processus cognitifs chez l'humain et l'animal

Machine learning

Définition (Wikipédia)

Apprentissage automatique/machine/artificiel/statistique : étude des algorithmes capables d'améliorer leurs performances à partir de données ou de l'expérience.

Trois types d'apprentissage

- Supervisé : apprendre à partir de données labellisées (classification)
- Non supervisé : apprendre à partir de données sans labels (clustering)
- Par renforcement : apprendre à partir de règles et de l'expérience (AlphaGo)

Repères historiques

- Supervisé : réseaux de neurones (1960), SVM (Vapnik, 1970), deep learning (1990, puis 2010)
- Non supervisé : k-means (1960), GMMs (1970), deep learning (2015)
- Par renforcement : Richard Sutton et Chris Watkins (1990)
- Succès récent (2010) : explosion des données + calcul bon marché (GPUs, cloud)

Deep learning

Définitions

Algorithmes de ML composés de multiples couches, pour extraire l'information des données brutes vers l'abstraction et la sémantique.

Réseaux de neurones profonds (DNNs)

- Apprentissage de DNNs par backpropagation (Yann LeCun, 1989)
- Apprentissage en parallèle sur GPUs (2009)
- Début 2010's : révolution du deep learning; librairies PyTorch (Facebook) et TensorFlow (Google).
- 2019 : Prix Turing pour Yoshua Bengio, Geoffrey Hinton et Yann LeCun

Applications phares

- Computer Vision
- Traitement automatique des langues
- Systèmes de recommandation
- Drug discovery

Computer Vision

Définition

Vision artificielle/par ordinateur : permettre à une machine d'analyser, traiter et comprendre une ou plusieurs images prises par un système d'acquisition.

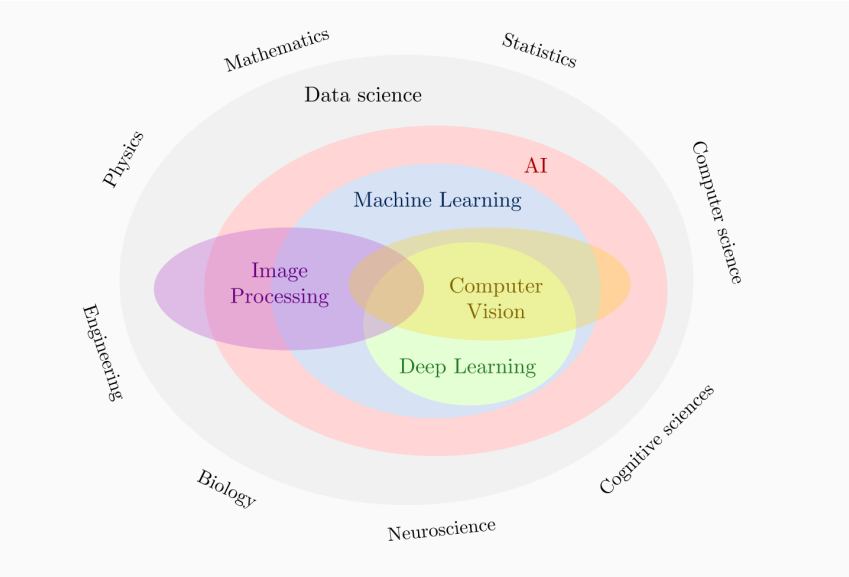
Historique

- 1970 : extraction de features, calibration
- 1980 : contours, scale-space
- 1990 : computer graphics (3D, rendering)
- 2000 et 2010 : ML puis DL . Rupture avec le papier NIPS 2012 d'Alex Krizhevsky, Ilya Sutskever et Geoffrey Hinton

Tâches et applications

- Navigation (robotique, véhicules autonomes)
- Détection d'événements (surveillance)
- Reconnaissance d'objets
- Analyse du mouvement

Hiérarchie des disciplines



[Charles Deledalle]

Recherche en ML/DL/CV

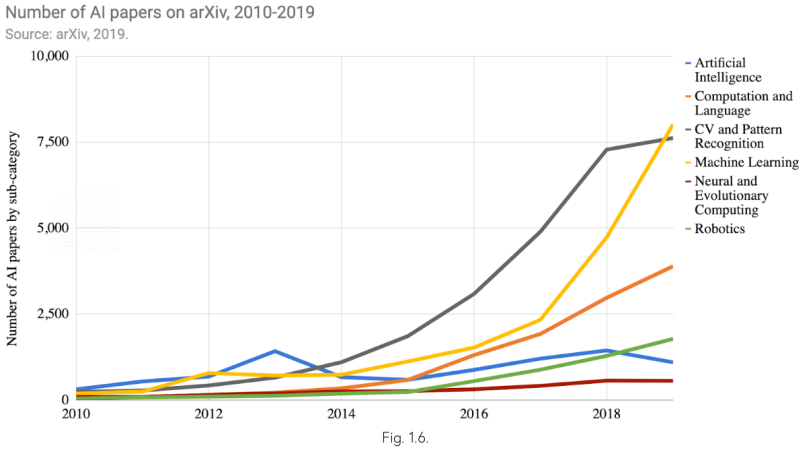


Fig. 1.6.

[Artificial Intelligence Index Report 2019]

Recherche en ML/DL/CV

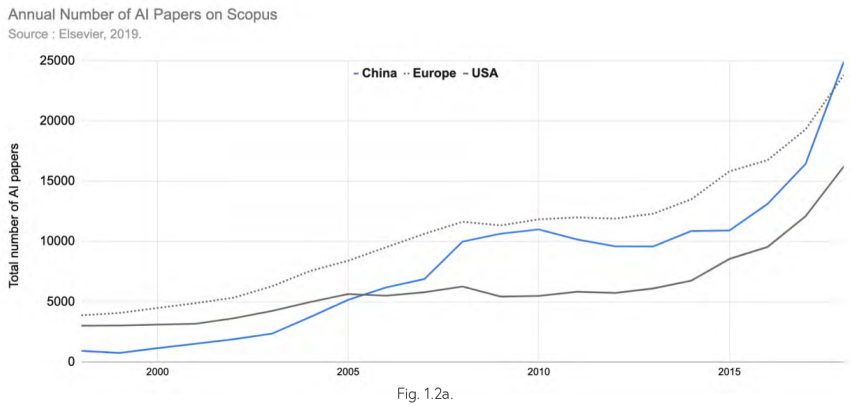


Fig. 1.2a.

[Artificial Intelligence Index Report 2019]

Recherche en ML/DL/CV

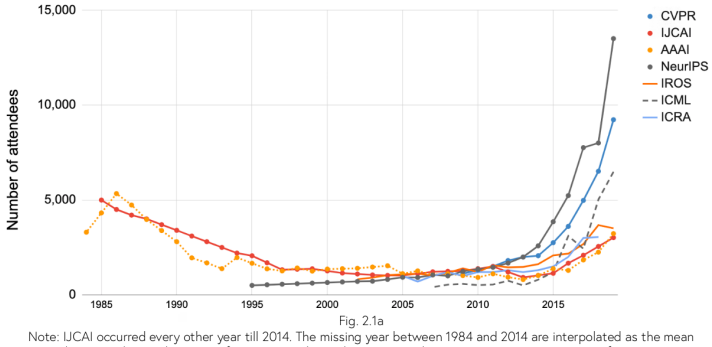


Fig. 2.1a

Note: IJCAI occurred every other year till 2014. The missing year between 1984 and 2014 are interpolated as the mean between the two known conference attendance dates to provide a comparative view across conferences.

[Artificial Intelligence Index Report 2019]

- NeurIPS: Neural Information Processing Systems
- CVPR: Computer Vision and Pattern Recognition
- ICML: International Conference on Machine Learning

Plan de la séance

Position du problème

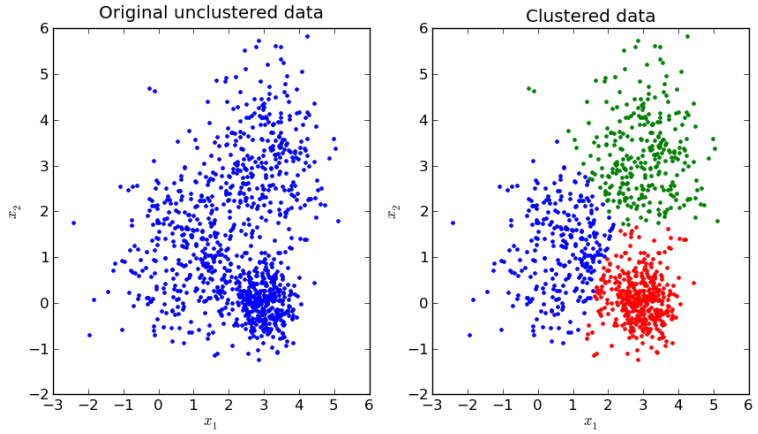
- On dispose de N données $x_n \in \mathbb{R}^P$
- On cherche à partitionner ces points en K groupes ou clusters
- On note les clusters $C_k \subset \{1, \cdots, N\}$, et $|C_k|$ leur cardinal
- On note μ_k le centre de gravité du cluster k :

$$\mu_k = \frac{1}{|C_k|} \sum_{n \in C_k} x_n.$$

Un problème difficile

- Nombre de partitions possibles $\approx K^n/K!$
- Exemple pour $n = 100$ et $K = 5$: 10^{68} partitions possibles!
- En règle générale, le problème du clustering est NP-hard

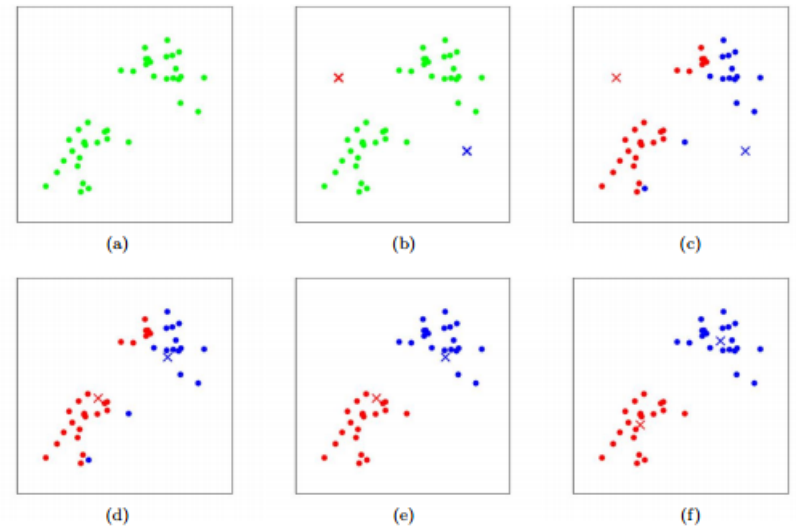
Illustration



Algorithme K-means

1. Initialiser les barycentres μ_k, possiblement aléatoirement
2. Associer chaque point x_n au barycentre le plus proche : $C_k = \{n \,/\, \|x_n - \mu_k\| \leq \|x_m - \mu_j\|\,, \forall m \neq n\}$
3. Mettre à jour les barycentres $\mu_k = \frac{1}{|C_k|} \sum_{n \in C_k} x_n$
4. Retour à l'étape 2 jusqu'à convergence

Algorithme K-means



Analyse des K-means

L'algorithme K-means converge vers un point stationnaire de la fonction coût

$$J(C, \mu) = \sum_{k=1}^{K} \sum_{n \in C_k} \|x_n - \mu_k\|_2^2 .$$

Remarques

- J est la somme des distances intra-classes (au carrées)
- J est non-convexe et admet de nombreux minima locaux et points stationnaires
- Le résultat dépend fortement de l'inititalisation
- Le résultat dépend fortement du nombre de clusters K, choisi par l'utilisateur.
- Algorithme de minimisation alternée
- Complexité en $O(npK)$ par itération, convergence assez rapide en pratique

Autres approches

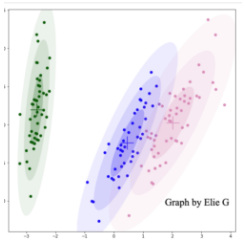
Variantes de K-means

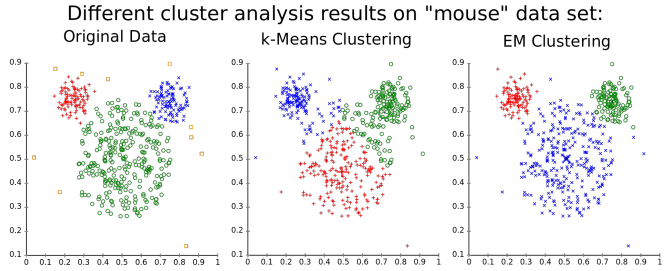
- K-medians : moyenne remplacée par médiane, et donc ℓ_2 par ℓ_1 dans J
- K-medoids
- Fuzzy C-means : clustering "doux", ie avec une probabilité d'appartenance aux K clusters
- K-means++ : meilleure initialisation
- ...

- Clustering hiérarchique (dendrogrammes)
- Clustering spectral : analyse spectrale (valeur propres) de la matrice de similarités
- ACP à noyaux (voir 3.)
- Mélange de gaussiennes

Mélange de Gaussiennes

- On suppose $x_n \sim \mathcal{N}(\mu_k, \Sigma_k)$
- On estime les paramètres μ_k, Σ_k par un algorithme expectation-maximization (EM)





Application en imagerie

Segmentation d'images couleurs, multi- ou hyperspectrales

- Echantillons x_n = les pixels d'une image
- Clustering des pixels = segmentation de l'image en zones de teinte homogène
- Méthodes simples et rapides, mais ne prennent pas compte l'information spatiale

Quantification adaptative

- Echantillons x_n = les pixels d'une image (ou de plusieurs images)
- Clustering avec un grand nombre de clusters $\longrightarrow$ on repère les teintes principales
- On peut ensuite faire de la quantification non uniforme : on remplace chaque pixel par la valeur du barycentre correspondant
- Utilisé par exemple en compression

Plan de la séance

Réduction de dimension

Motivation

Cas des données en grande dimension ($P \gg 1$)

- Difficile de les visualiser
- Grande complexité algorithmique (temps et mémoire)
- Beaucoup de variables sont souvent peu informatives (exemple : variables fortement corrélées)

Obectifs

- Réduire la dimension des données
- Sélectionner les variables pertinentes
- Décorréler les données
- Permettre de les visualiser
- Pré-processing (débruitage par exemple)

Analyse en composantes principales

Notation et définition

- N observations $x_n \in \mathbb{R}^P \longrightarrow$ matrice $X \in \mathbb{R}^{N \times P}$
- L'analyse en composantes principales de X consiste en la projection des données formées par les lignes de X sur un sous-espace affine de dimension $K \leq P$ qui maximise la dispersion du nuage projeté.

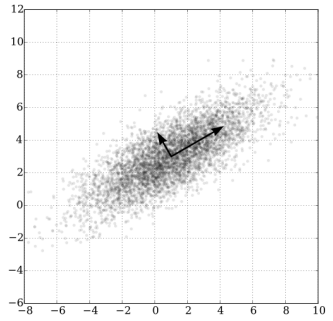


Illustration en dimension 2 ($p = 2$) [Wikipedia]

Analyse en composante principale

Formalisation

- On cherche le vecteur unitaire $u \in \mathbb{R}^P$ tel que, projetées sur le sous-espace de dimension 1 engendré par u, nos données gardent la dispersion la plus grande possible.
- On commence par retirer la moyenne $x_n \leftarrow x_n - \bar{x} \forall n$ avec $\bar{x} = \frac{1}{N} \sum_n x_n$
- La solution u est celle qui maximise la variance résiduelle $\varphi(u) = \|Xu\|_2^2 = u^T (X^T X) u$
- En décomposant $X^T X$ en valeurs propres, on voit facilement que u est le vecteur propre associé à la plus grande valeur propre λ_1.
- On peut itérer : les K composantes principales sont les K vecteurs propres associés aux K plus grandes valeurs propres.

Analyse en composantes principales

- Composantes principales : $U \in \mathbb{R}^{P \times K}$ formé des vecteurs propres associés aux K plus grandes valeurs propres
- Données dans la nouvelle base : $Y = XU$
- Re-projection dans l'espace d'origine : $\tilde{X} = YU^T \quad (+\bar{x}$ si besoin)

Analyse en composantes principales

Meilleure approximation dans un sous-espace

- $\tilde{X} = YU^T = XUU^T$
- $\tilde{X}$ est la meilleure approximation des données X dans l'espace de dimension réduite K (théorème d'Eckart–Young–Mirsky)

Valeurs propres

- On note $\lambda_1 > \cdots > \lambda_K$ les K plus grandes valeurs propres de $X^T X$
- Variance capturée $\|XU\|_F^2 = \sum_{k=1}^K \lambda_k$
- Erreur d'approximation $\left\|X - \tilde{X}\right\|_F^2 \propto \left\|I - UU^T\right\|_2^2 = \sum_{k=K+1}^P \lambda_k$
- Lien avec la SVD $X = PDQ^*$, alors $d_{k,k}^2 = \lambda_k$
- Choix de K en pratique : L-curve

Analyse en composantes principales

Algorithme

- Entrées : données $X \in \mathbb{R}^{N \times P}$, nombre de composantes K
- Sortie : composantes principales $U \in \mathbb{R}^{K \times p}$

1. Centrage des données : $x_n \leftarrow x_n - \bar{x} \quad \forall n$ avec $\bar{x} = \frac{1}{N} \sum_n x_n$
2. [Optionnel] Réduction des données $x_n \leftarrow x_n ./ \sigma_x$ avec $\sigma_x = \frac{1}{N} \sum_n (x_n - \bar{x})^{\cdot 2}$
3. Calcul de la matrice de covariance empirique $Y = X^T X$
4. Diagonalisation partielle pour trouver les K premiers vecteurs propres formant U

Complexité

1. $O(np)$
2. $O(np)$
3. $O(np^2)$
4. $O(Kp^2)$

Extensions

Robust PCA, kernel PCA, etc

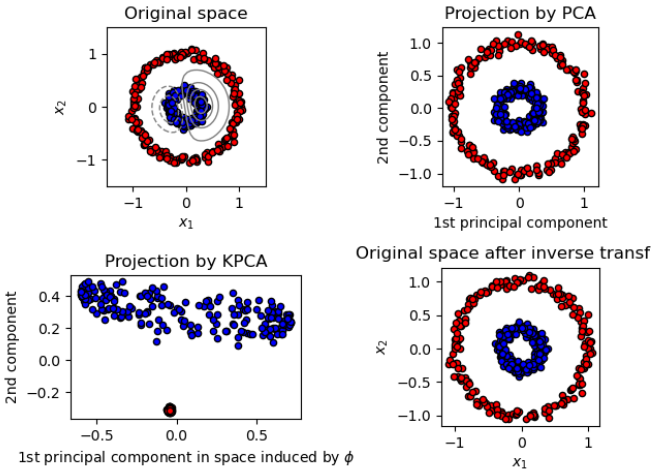


Illustration de l'ACP à noyau [scikit-learn]

Application en imagerie

Reconnaissance de visages (eigenface)

- Echantillons x_n = des images de visages vectorisées
- ACP puis clustering
- Premiers succès en reconnaissance de formes ($\approx$ 1990)
- Peu efficaces, notamment car pas invariant par translation : on utilise plutôt des features (attributs) plutôt que les images brutes

Compression ou débruitage

- Echantillons x_n = des patches d'une ou plusieurs images (avec possible recouvrement)
- ACP puis seuillage des valeurs propres
- Idée : redondance spatiale dans les images

Plan de la séance

L'ACP comme factorisation matricielle

L'ACP (la SVD) réalise la meilleure approximation de rang K au sens de la distance euclidienne

$$X \approx YU^T$$

Avantages et inconvénients

+ Méthode simple, bien posée mathématiquement, facile à calculer
- Composantes principales (U) et coefficients de représentation (Y) pas interprétables

Décomposition en matrices non-négatives (NMF)

Données $X \in \mathbb{R}^{N \times P}$.

ACP (rang K)

Coefficients $Y \in \mathbb{R}^{N \times K}$ et composantes principales $U \in \mathbb{R}^{K \times P}$

$$\min_{Y,U} \|X - YU\|_F^2 \text{ s. t. } U^T U = I$$

NMF [Lee and Seung, Nature, 1999]

Coefficients $H \in \mathbb{R}^{N \times K}$ et dictionnaire $W \in \mathbb{R}^{K \times P}$

$$\min_{W,H} \|X - HW\|_F^2 \text{ s. t. } H, W \geq 0$$

Intérêt de la non-négativité

- Composantes interprétables
- Représentation additive (par parties), car coefficients positifs ou nuls
- Mais : plus difficile à calculer (plus lent, et problème non convexe)

Autres méthodes de séparation de sources

- Variantes de la NMF avec d'autres divergences (KL, β-divergences)
- Apprentissage de dictionnaire avec contraintes de parcimonie
- Analyse en composantes indépendantes (ICA) : basée sur une hypothèse d'indépendance statistique entre les composantes (les "sources")
- Démélange linéaire et non-linéaire
- Factorisations bayésiennes : NMF, LDA