Université de Batna 2 (Mostefa Ben Boulaid)
Faculté de Technologie
Département de Génie Industriel

Apprentissage Automatique

Cours 5 (Réduction de dimensionnalité)

Pr Hassen BOUZGOU

Différents problèmes de ML

	Apprentissage Supervisé	Apprentissage Non-supervisé
Discrète	Classification / Catégorisation	Clustering
Continue	Régression	Réduction de dimensionnalité

Intro

- Les sources de données sont variés et produisent, une quantité variable d'information à représenter. Tanque cette quantité augmente, les techniques traditionnelles de visualisation (nuage de points, histogramme, etc.) deviennent insuffisantes.

- Dans le sens commun, la notion de dimension renvoie à la taille ; les dimensions d'une pièce sont sa longueur, sa largeur et sa profondeur/son épaisseur/sa hauteur, ou bien son diamètre.

Origine des données

L'origine des données peut être divisée en deux catégories :

- **Valeurs réelles** (données prises par observation)
  - Mesures: température, pression, tension électrique…..
  - Images: satellitaires, médicales, surveillance….
  - Texte: e-mails, SMS, ….
- **Valeurs Simulées** (données prises par modélisation)
  - Architecture, design automobile
  - Modélisation économique
  - …. etc.

La quantité de données à représenter dépend principalement du nombre de **paramètres (caractéristiques)** et du **nombre de mesures**.

Données de haute dimension

L'analyse de données moderne doit faire face à d'énormes quantités de données. Les données sont en effet de plus en plus facilement acquises et stockées, en raison d'énormes progrès dans les capteurs et les moyens de recueillir des données d'un côté, et dans les dispositifs de stockage de l'autre côté. De nos jours, il n'y a pas d'hésitation dans de nombreux domaines à acquérir de très grandes quantités de données sans savoir à l'avance si elles seront analysées et comment.

Exemples des données HD

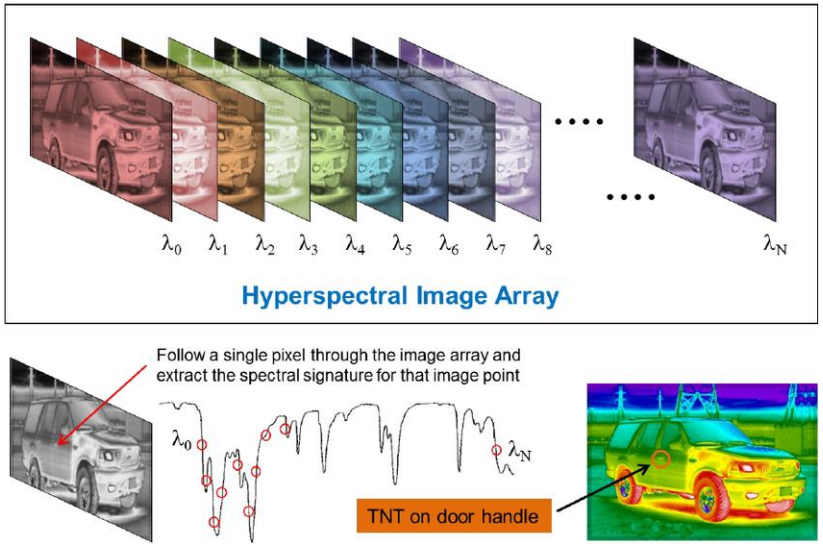
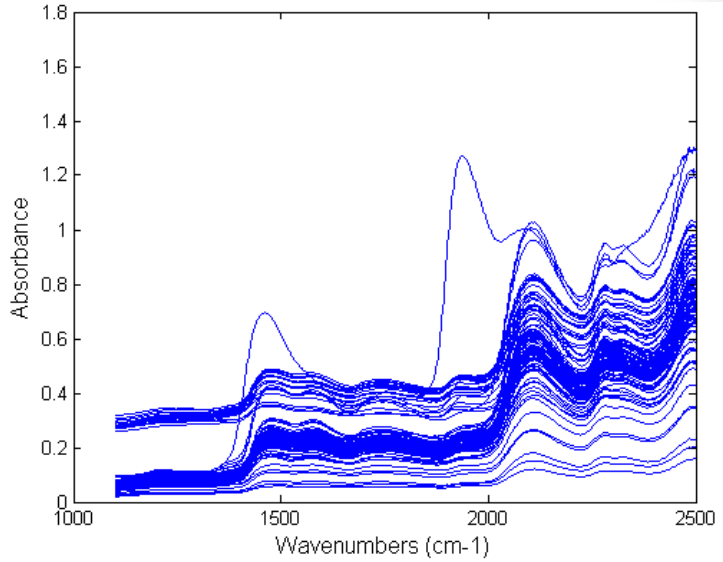


Image Hyperspectrale



Spectroscopie

features

observations

	Gene feature 1	Gene feature 2	Gene feature 3	Gene feature 4	...	...
Individual 1						
Individual 2						
Individual 3						

Génomique

Dimensionnalité

- Difficulté à analyser les données de grande dimension
  - Outils d'analyse conçus pour les données de faible dimension
- Le fléau de la dimensionnalité
  - Si 10 échantillons sont raisonnables à apprendre un modèle de 1-dim; 100 pour 2-dim. et 1000 pour 3-dim…..!!!???
- En régression linéaire, si le nombre de paramètres est plus grand que les coordonnées de l'espace => problème indéfini.

Fléau de dimensionnalité

- Le **fléau de la dimension** ou **malédiction de la dimension** (*curse of dimensionality*) est un terme inventé par Richard Bellman en 1961 pour désigner divers phénomènes qui ont lieu lorsque l'on cherche à analyser ou organiser des données dans des espaces de grande dimension, alors qu'ils n'ont pas lieu dans des espaces de dimension moindre..

- Plusieurs domaines sont concernés et notamment l'apprentissage automatique, la fouille de données, les bases de données, l'analyse numérique ou encore l'échantillonnage. L'idée générale est que lorsque le nombre de dimensions augmente, le volume de l'espace croît rapidement si bien que les données se retrouvent « isolées » et deviennent éparses. Cela est problématique pour les méthodes nécessitant un nombre significatif de données pour être valides, les rendant alors peu efficaces.

Solution ⇒ Réduction de la dimensionnalité

Pourquoi le Réduction de la dimensionnalité

- Certaines caractéristiques peuvent ne pas être pertinentes
- Nous voulons visualiser des données de grande dimension
- La dimensionnalité "intrinsèque" peut être plus petite que le nombre de caractéristiques

Evaluation des caractéristiques

- Information mutuelle entre l'attribut et la sortie
- indépendance entre caractéristiques et classe (sortie désirée)
- Performance de classification / régression

Sélection non-supervisée des Caractéristiques

Diffère de la sélection des caractéristiques de deux manières:

Au lieu de choisir un sous-ensemble de caractéristiques.

- Créer de nouvelles caractéristiques définies en fonction de toutes les caractéristiques originales
- Ne tient pas compte des étiquettes de classe, mais seulement les caractéristiques de données

Idée:

- Compte tenu des points de données dans l'espace *d*-dimensionnel.
- ***Projeter*** dans l'espace de dimension inférieure tout en préservant autant d'informations que possible

- ***Exemple:*** trouver la meilleure approximation planaire aux données 3D
  - **Solution:** choisissez une projection qui minimise l'erreur quadratique dans l'espace original reconstruit des données

L' Analyse en Composantes Principales (ACP)

But: on cherche à définir *k* nouvelles variables combinaisons linéaires des *d* variables initiales qui feront perdre le moins d'information possible

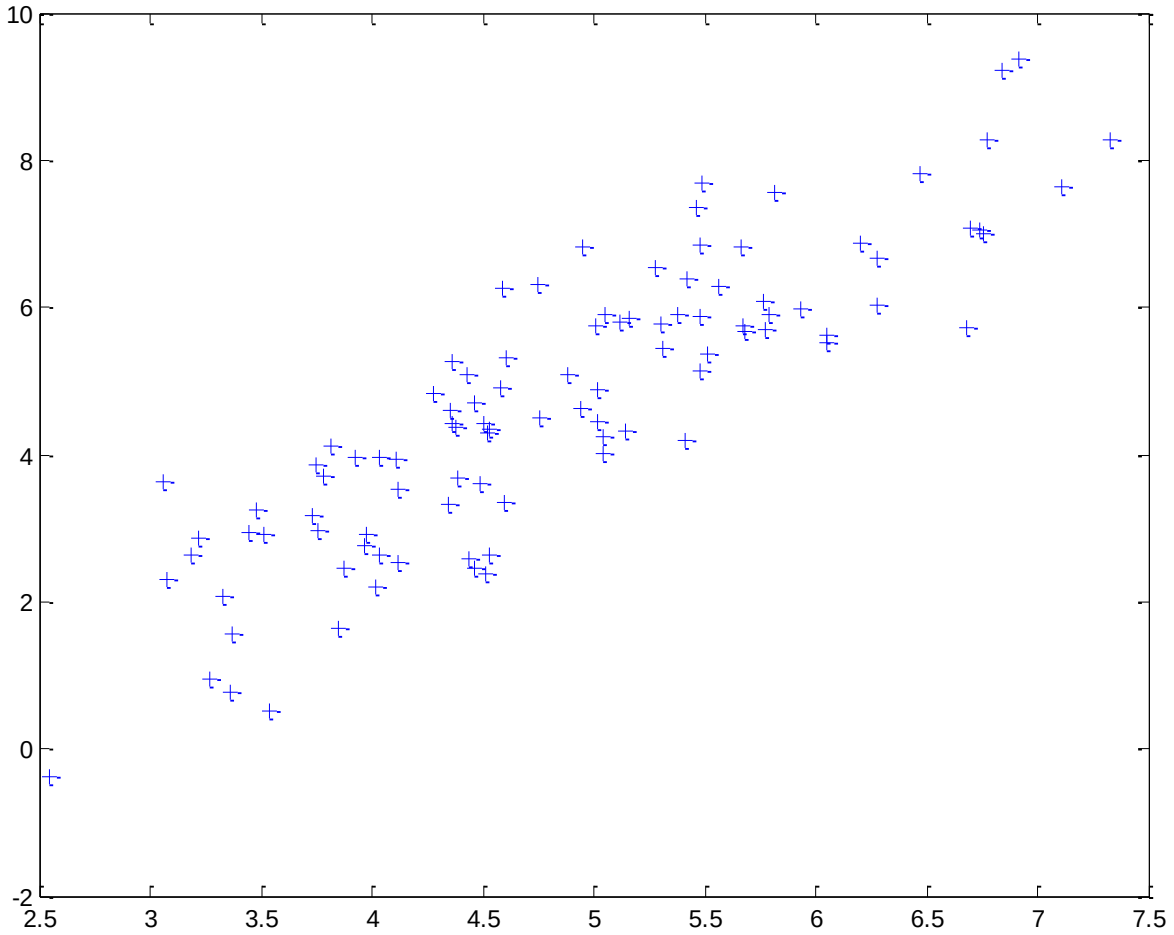
Algorithme:

1. Créer une matrice de données X (N x d), avec un vecteur de ligne x_n par point de données
2. Soustraire la moyenne de X de chaque vecteur de ligne x_n dans X
3. Trouver la matrice de covariance C de X
4. Trouvez les vecteurs propres et les valeurs propres de C
5. Déterminer les CP: les *k* vecteurs propres avec les plus grandes valeurs propres (**seuil?**)

Code Matlab du PCA

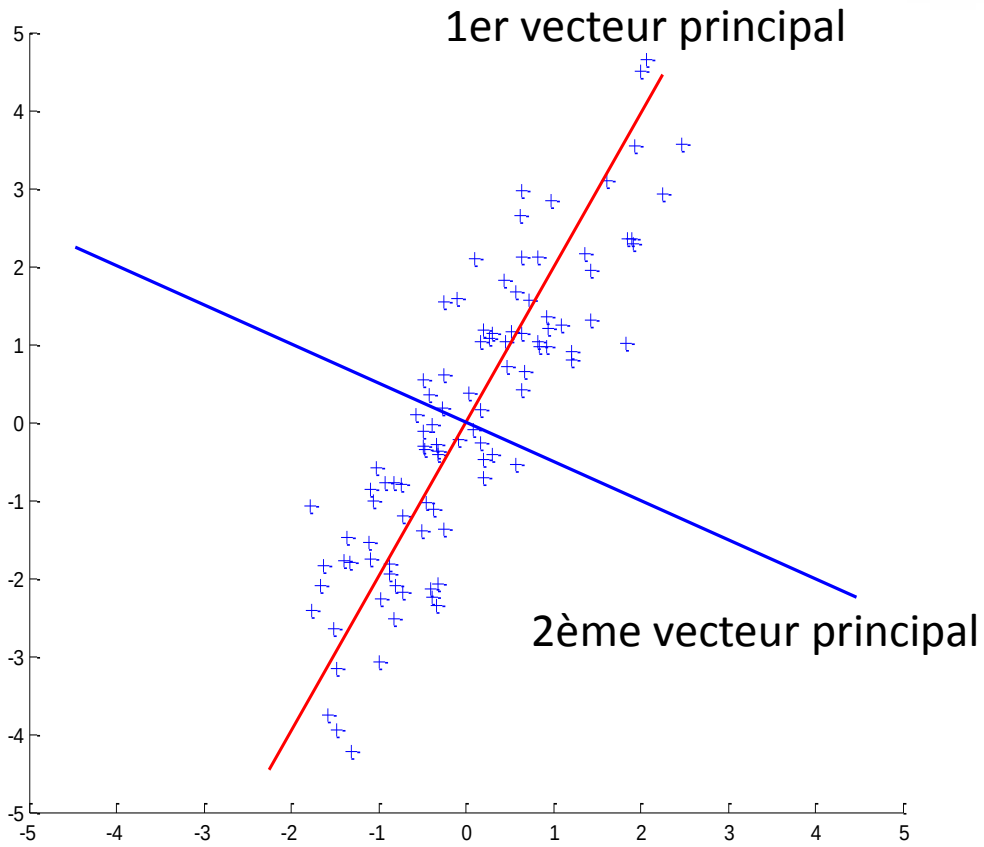
```
% generer les données
Data = mvnrnd([5, 5],[1 1.5; 1.5 3], 100);
figure(1); plot(Data(:,1), Data(:,2), '+');
% centrer les données
for i = 1:size(Data,1)
  Data(i, :) = Data(i, :) - mean(Data);
end
% matrice de covariance
DataCov = cov(Data);

%Valeurs Prpores
[PC, variances, explained] = pcacov(DataCov);

% Tracer les composantes principales
figure(2); clf; hold on;
plot(Data(:,1), Data(:,2), '+b');
plot(PC(1,1)*[-5 5], PC(2,1)*[-5 5], '-r')
plot(PC(1,2)*[-5 5], PC(2,2)*[-5 5], '-b'); hold off

% Projeter à une dimension
PcaPos = Data * PC(:, 1);
```

Données 2D



Composantes Principales

- Donne le meilleur axe de projection
- Erreur RMS minimale
- Les vecteurs principaux sont orthogonaux



Combien de Composantes Principales?

- Vérifier la distribution des valeurs propres
- Prenez suffisamment de vecteurs propres pour couvrir 80-90% de la variance