Apprentissage statistique

Master parcours SSD - UE Apprentissage Statistique I

Pierre Mahé - bioMérieux & Université de Grenoble-Alpes

Apprentissage statistique ?

Apprentissage statistique = apprentissage automatique

- (statistical learning, machine learning)

Wikipedia : Machine learning is the subfield of computer science that gives computers the ability to learn without being explicitly programmed. [...] Machine learning explores the study and construction of algorithms that can learn from and make predictions on data – such algorithms overcome following strictly static program instructions by making data driven predictions or decisions, through building a model from sample inputs.

⇒ Apprentissage à partir d'exemples

- par opposition aux **systèmes experts**.

Motivation

Il est très dur de définir ce qui "fait" un 2 :

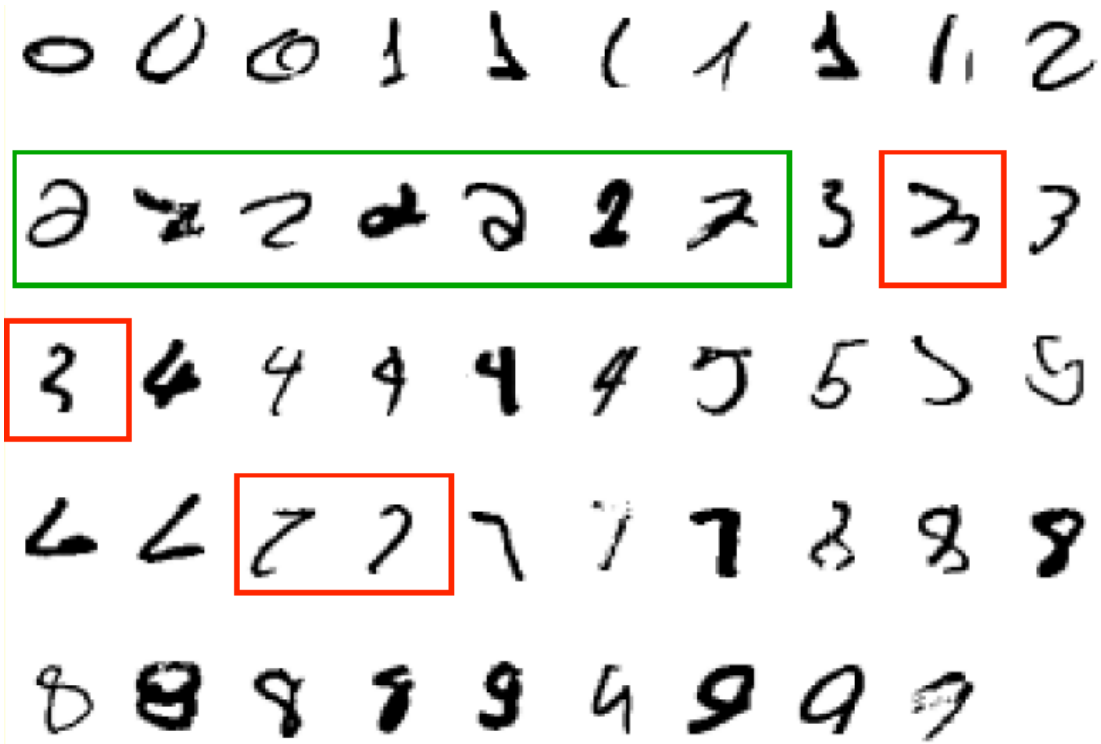
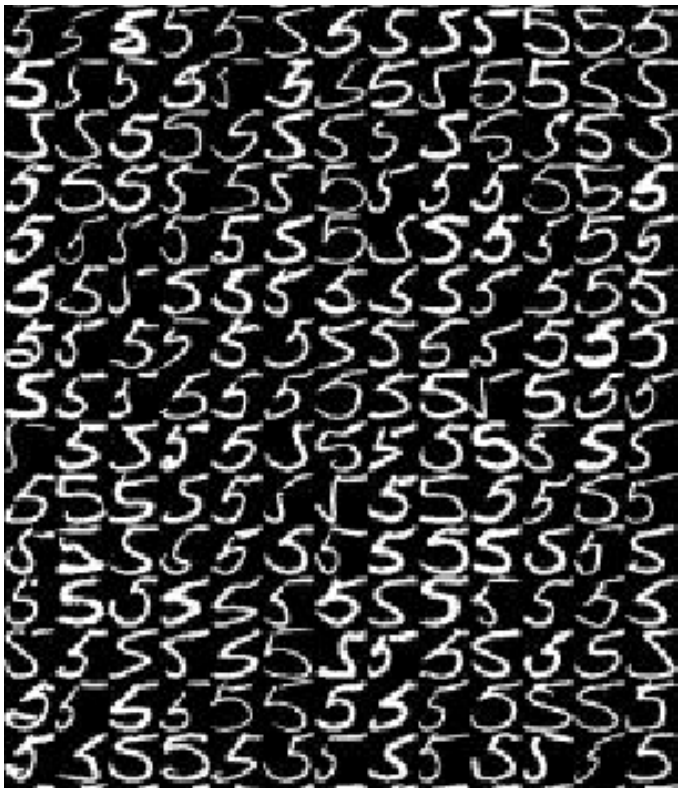


Figure: Exemple tiré d'un cours de G. Hinton

...mais on sait très bien apprendre à les reconnaître.

L'approche Machine Learning [1]



1. USPS dataset : `http://www.cs.nyu.edu/~roweis/data.html`

Apprentissage statistique

Principe = trouver des régularités dans les données

Pourquoi faire ?

- découvrir des structures "cachées" dans les observations
  - apprentissage **non-supervisé**
- prédire une propriété pour de nouvelles observations
  - apprentissage **supervisé**

A l'interface de nombreux domaines :

- statistiques
- informatique ("computer science")
- intelligence artificielle
- mathématiques (e.g., optimisation numérique)
- théorie de la décision
- ...

Intérêt

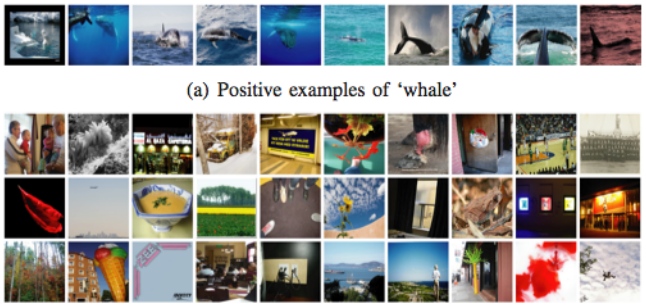
L'apprentissage statistique prend tout son sens quand :

- l'expertise humaine est absente
  - e.g., analyse de l'ADN
- il est très difficile de l'expliciter
  - e.g., reconnaissance de caractères
- les quantités de données à traiter sont trop importantes
  - e.g., applications web / réseaux sociaiux
- les données évoluent dynamiquement
  - e.g., prédire le cours d'actions financières
- ...

Enormément d'applications dans de nombreux domaines !

Analyse d'images

- Catégorisation



(a) Positive examples of 'whale'

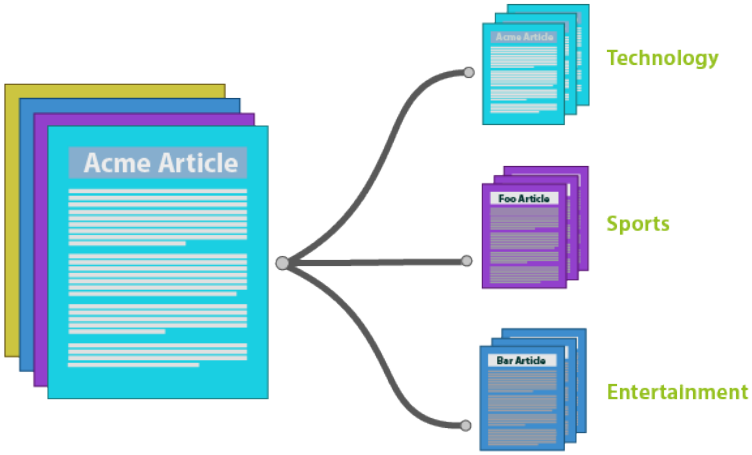
(b) Negative examples of 'whale' obtained by random sampling

(c) Relevant negatives of 'whale' (this paper)

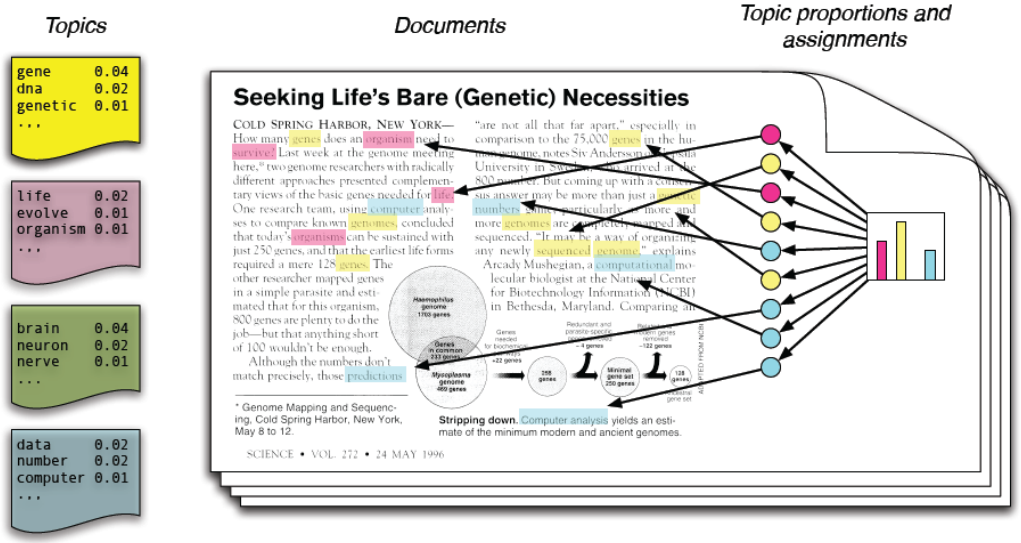
- Interprétation de scènes

Analyse de texte

- Catégorisation automatique



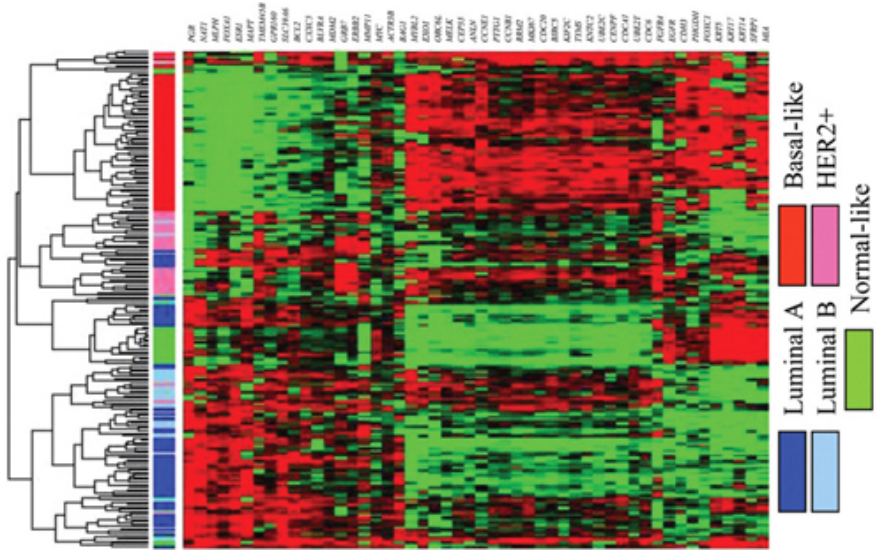
- Détection de thèmes (topics) "cachés"



Biologie

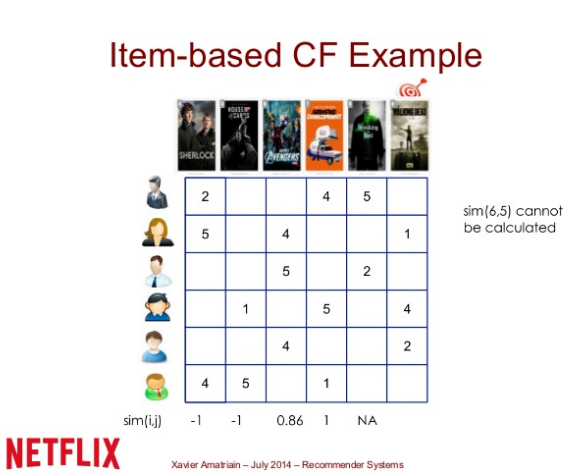
- Prédire la fonction d'une protéine à partir de sa structure

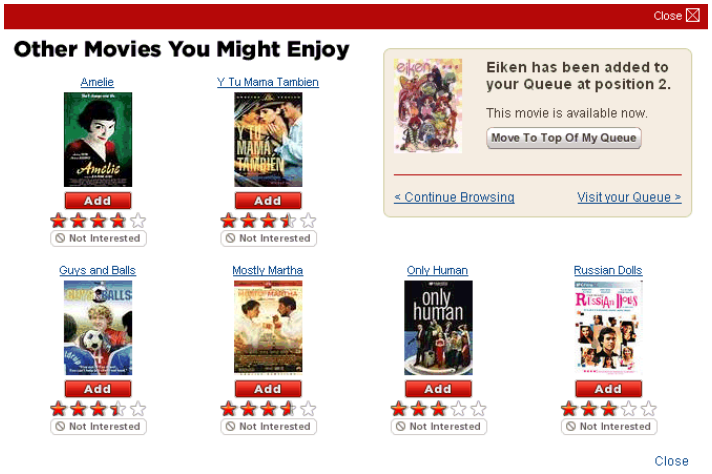
- Diagnostic/prognostic à partir de puces à ADN



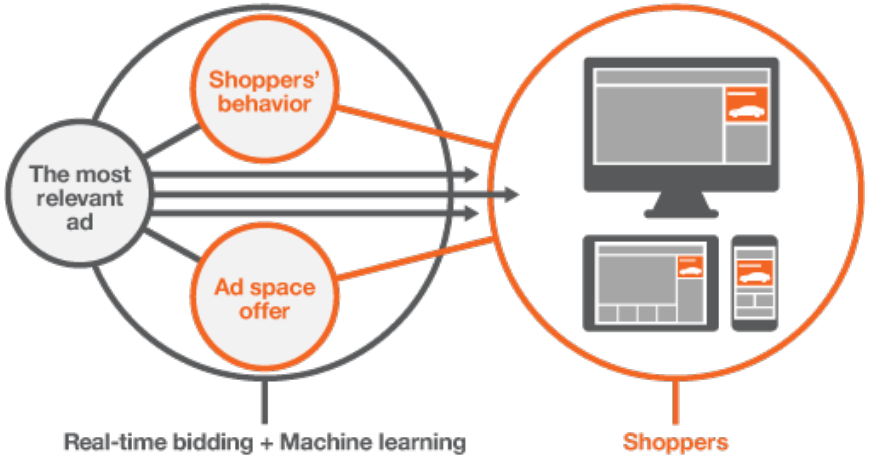
Web / Internet

- Recommandation



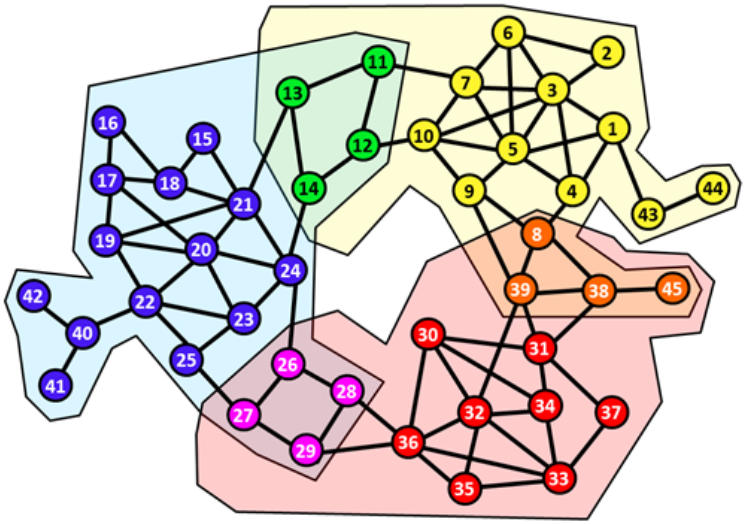


- Publicité ciblée



Réseaux sociaux

- Détection de communautés



Et plein d'autres....

- Audio : reconnaissance de la parole, séparation de sources
- Vidéo : suivi d'objets, surveillance
- Finance, économie
- Sciences de la Vie : climatologie, planétologie
- Génétique
- ...

Défis

- Apprendre des comportements qui soient valables pour d'autres observations
  - notion de **généralisation**
- Faire face à des types de données variés
  - vecteurs, matrices, courbes, séquences, arbres, graphes
- Faire face à des volumes de données conséquents
  - apprentissage : large-scale learning & haute dimension
  - prédiction : problématiques temps-réel

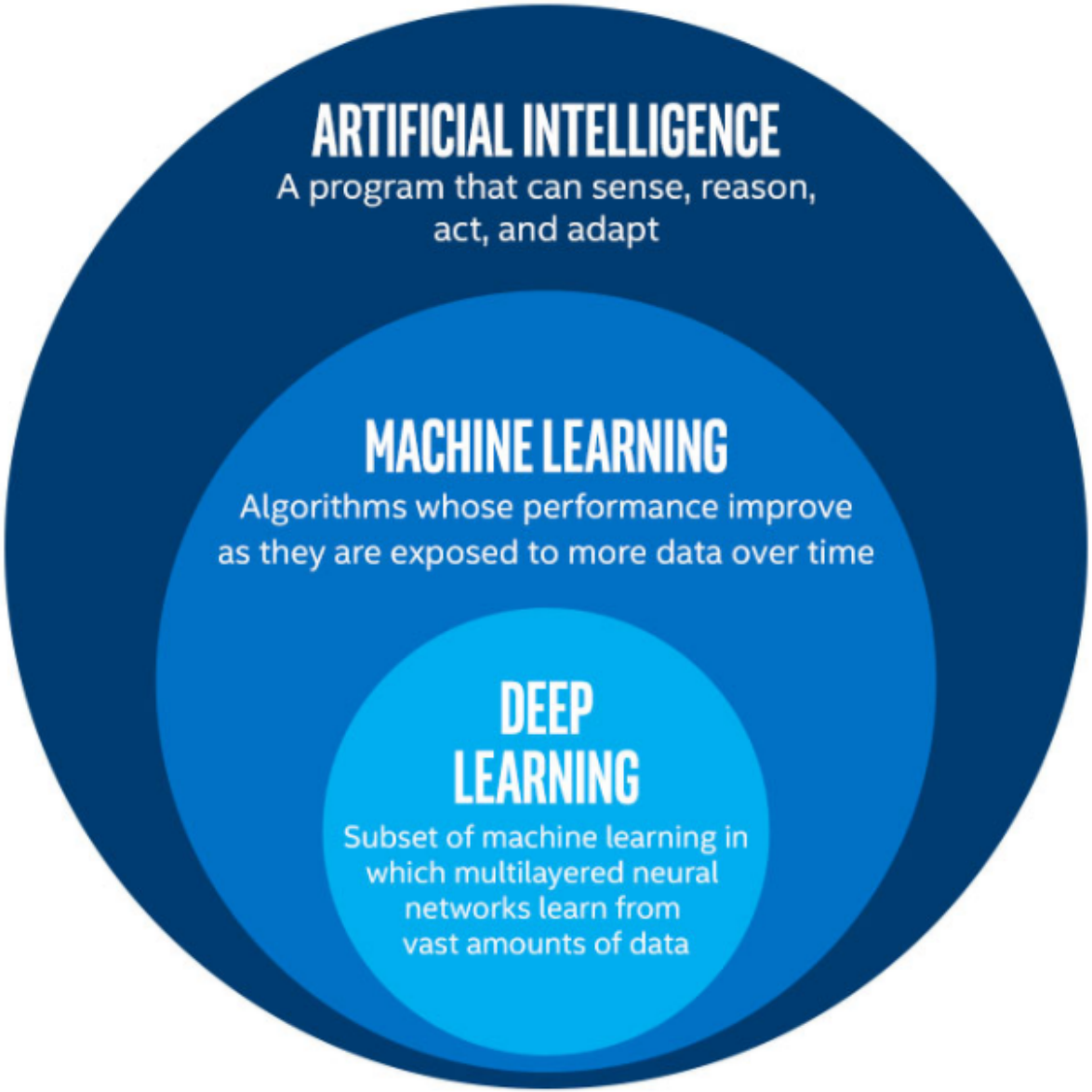
Machine learning & Big Data ?

Wikipedia : Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy. The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.

Les 5 "V" du big-data :

- Volume, Variety, Velocity, Variability, Veracity

Machine learning & Intelligence Artificielle ?



Les deux cadres d'apprentissage principaux

Apprentissage supervisé :

- données : observations $\{(x_i, y_i)\}$, $i = 1, ..., n$.
  - descripteurs / variables explicatives + variable d'intérêt
- objectif(s) : prédiction
  - (+ compréhension du lien entre X et Y)

Les deux cadres d'apprentissage principaux

Apprentissage supervisé :

- données : observations $\{(x_i, y_i)\}$, $i = 1, ..., n$.
  - descripteurs / variables explicatives + variable d'intérêt
- objectif(s) : prédiction
  - (+ compréhension du lien entre X et Y)

Apprentissage non-supervisé :

- données : observations $\{x_i\}$, $i = 1, ..., n$.
  - pas de variable à expliquer
- objectif : identifier des "structures" dans les données
  - moins clairement formalisé que le cadre supervisé

Les deux cadres d'apprentissage principaux

Apprentissage supervisé :

- données : observations $\{(x_i, y_i)\}$, $i = 1, ..., n$.
  - descripteurs / variables explicatives + variable d'intérêt
- objectif(s) : prédiction
  - (+ compréhension du lien entre X et Y)

Apprentissage non-supervisé :

- données : observations $\{x_i\}$, $i = 1, ..., n$.
  - pas de variable à expliquer
- objectif : identifier des "structures" dans les données
  - moins clairement formalisé que le cadre supervisé

Mais aussi :

- apprentissage par renforcement, semi-supervisé, transductif, actif, online, ...

Apprentissage supervisé

Apprentissage supervisé - principe

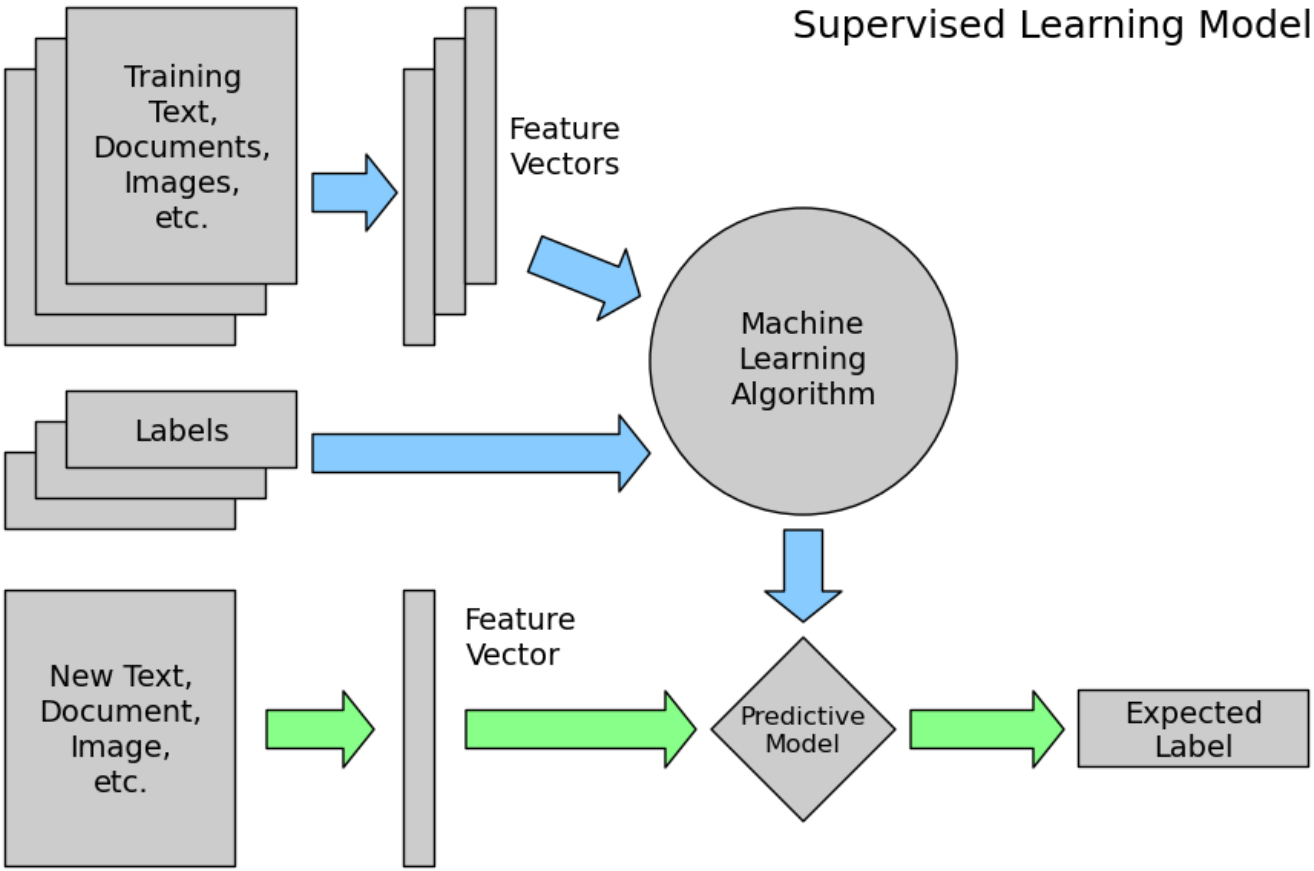


Figure: Image tirée de `http://www.astroml.org/sklearn_tutorial/general_concepts.html`

Apprentissage supervisé - formalisation

On dispose d'un échantillon $\{(x_i, y_i)\}$, $i = 1, ..., n$, :

- des observations $x_i \in \mathcal{X}$,
- des réponses associées $y_i \in \mathcal{Y}$.

$\Rightarrow$ ce sont les données (ou le jeu) d'apprentissage.

Apprentissage supervisé - formalisation

On dispose d'un échantillon $\{(x_i, y_i)\}$, $i = 1, ..., n$, :

- des observations $x_i \in \mathcal{X}$,
- des réponses associées $y_i \in \mathcal{Y}$.

$\Rightarrow$ ce sont les données (ou le jeu) d'apprentissage.

Typiquement :

- $\mathcal{X} = \mathbb{R}^p$: on parle de vecteurs de descripteurs.
  - features, attributes, input variables
- Si $\mathcal{Y} = \mathbb{R}$, on parle de régression.
- Si $\mathcal{Y} = \{1, ..., K\}$, on parle de classification.
- Si $\mathcal{Y} = \{-1, +1\}$, on parle de classification binaire.
  - on note parfois également $\mathcal{Y} = \{0, 1\}$

Apprentissage supervisé - illustration

Classification :

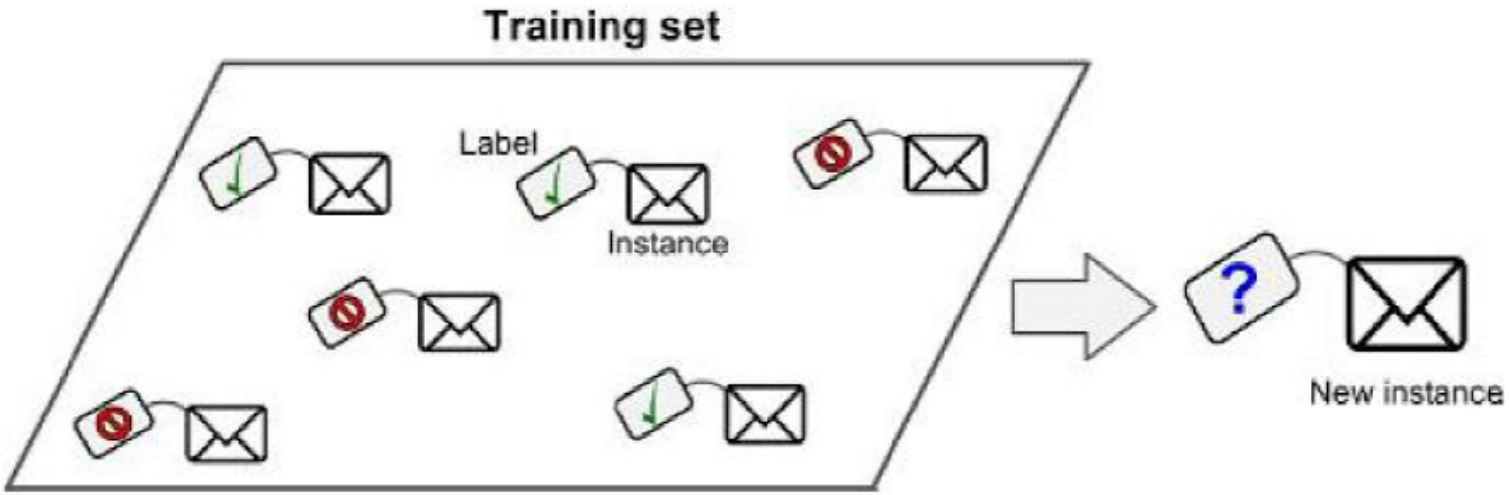


Figure: Image tirée de Géron (2017)

- $\mathcal{X} = \{\text{e-mails}\}$
- $\mathcal{Y} = \{1, ..., K\}$ (ici : spams / non-spams)

Apprentissage supervisé - illustration

Régression :

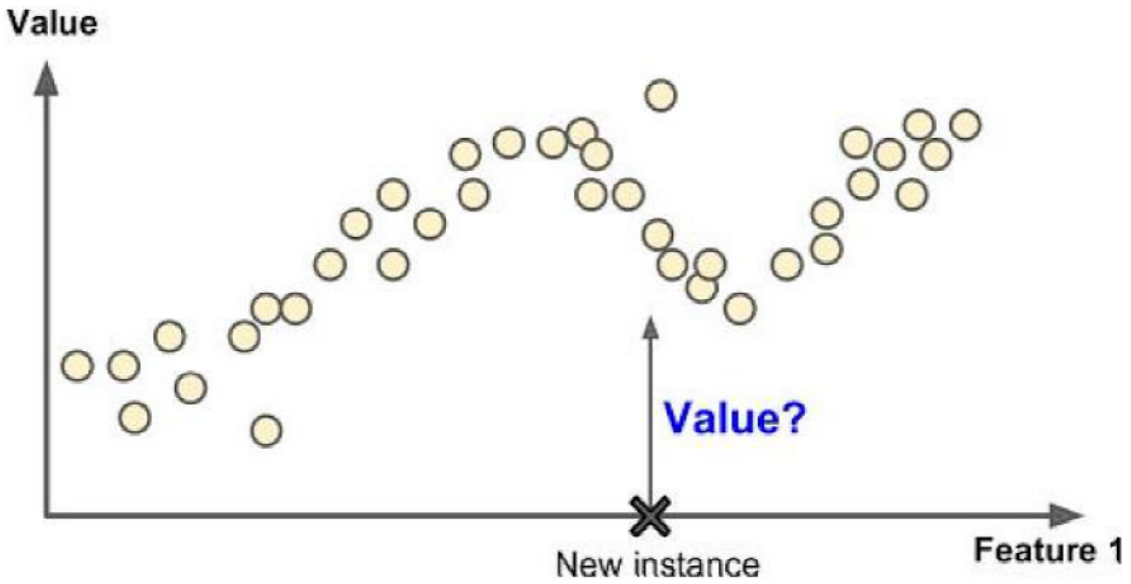


Figure: Image tirée de Géron (2017)

- $\mathcal{X} = \mathbb{R}$ (Feature 1)
- $\mathcal{Y} = \mathbb{R}$ (Value)

Apprentissage supervisé - formalisation

Données d'entrée : échantillon $\left\{(x_i, y_i)\right\}_{i=1,\dots,n} \in \mathcal{X} \times \mathcal{Y}$.

Objectif : apprendre une fonction $f : \mathcal{X} \to \mathcal{Y}$ permettant de prédire la réponse associée à une nouvelle observation.

Apprentissage supervisé - formalisation

Données d'entrée : échantillon $\{(x_i, y_i)\}_{i=1,\dots,n} \in \mathcal{X} \times \mathcal{Y}$.

Objectif : apprendre une fonction $f : \mathcal{X} \to \mathcal{Y}$ permettant de prédire la réponse associée à une nouvelle observation.

Critère : une fonction de perte L (pour "loss") mesurant l'erreur entre y et $f(x)$.

Typiquement :

- l'erreur quadratique pour la régression :

$$L(y, f(x)) = (y - f(x))^2$$

- le coût 0/1 pour la classification :

$$L(y, f(x)) = \mathbf{1}(y \neq f(x))$$

Apprentissage supervisé - formalisation

Cadre probabiliste : on considère que nos observations (x_i, y_i) sont des variables aléatoires régies par une loi jointe $P(X, Y)$.

Apprentissage supervisé - formalisation

Cadre probabiliste : on considère que nos observations (x_i, y_i) sont des variables aléatoires régies par une loi jointe $P(X, Y)$.

$\Rightarrow$ L'objectif de l'apprentissage supervisé est donc de trouver la fonction f minimisant l'espérance de la fonction de perte :

$$R(f) = E_{X,Y}\big[L(Y, f(X)\big],$$

à partir d'un échantillon $\big\{(x_i, y_i)\big\}$, $i = 1, ..., n$.

Apprentissage supervisé - formalisation

Cadre probabiliste : on considère que nos observations (x_i, y_i) sont des variables aléatoires régies par une loi jointe $P(X, Y)$.

$\Rightarrow$ L'objectif de l'apprentissage supervisé est donc de trouver la fonction f minimisant l'espérance de la fonction de perte :

$$R(f) = E_{X,Y}\big[L(Y, f(X)\big],$$

à partir d'un échantillon $\big\{(x_i, y_i)\big\}$, $i = 1, ..., n$.

$R(f)$ est appelée le risque (ou la perte) de la fonction f.

Apprentissage supervisé - risque empirique

A minima, un "bon" prédicteur devrait bien se comporter sur les données d'apprentissage...

Apprentissage supervisé - risque empirique

A minima, un "bon" prédicteur devrait bien se comporter sur les données d'apprentissage...

On s'intéresse donc en premier lieu au risque empirique :

$$R_{emp}(f) = \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big).$$

Apprentissage supervisé - risque empirique

A minima, un "bon" prédicteur devrait bien se comporter sur les données d'apprentissage...

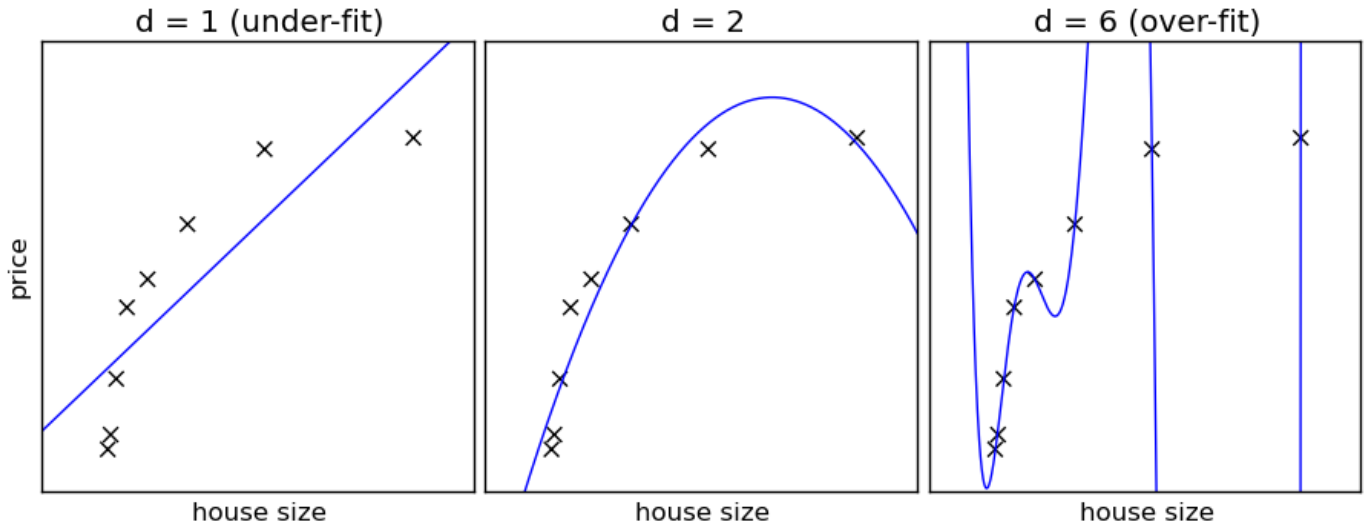
On s'intéresse donc en premier lieu au risque empirique :

$$R_{emp}(f) = \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, f(x_i)\big).$$

Mais minimiser le risque empirique n'est pas suffisant, il faut également contrôler la complexité du modèle pour éviter le sur-apprentissage.

Risque empirique et sur-apprentissage

Illustration de sous-apprentissage et sur-apprentissage sur un problème (jouet) de régression (polynomiale) :

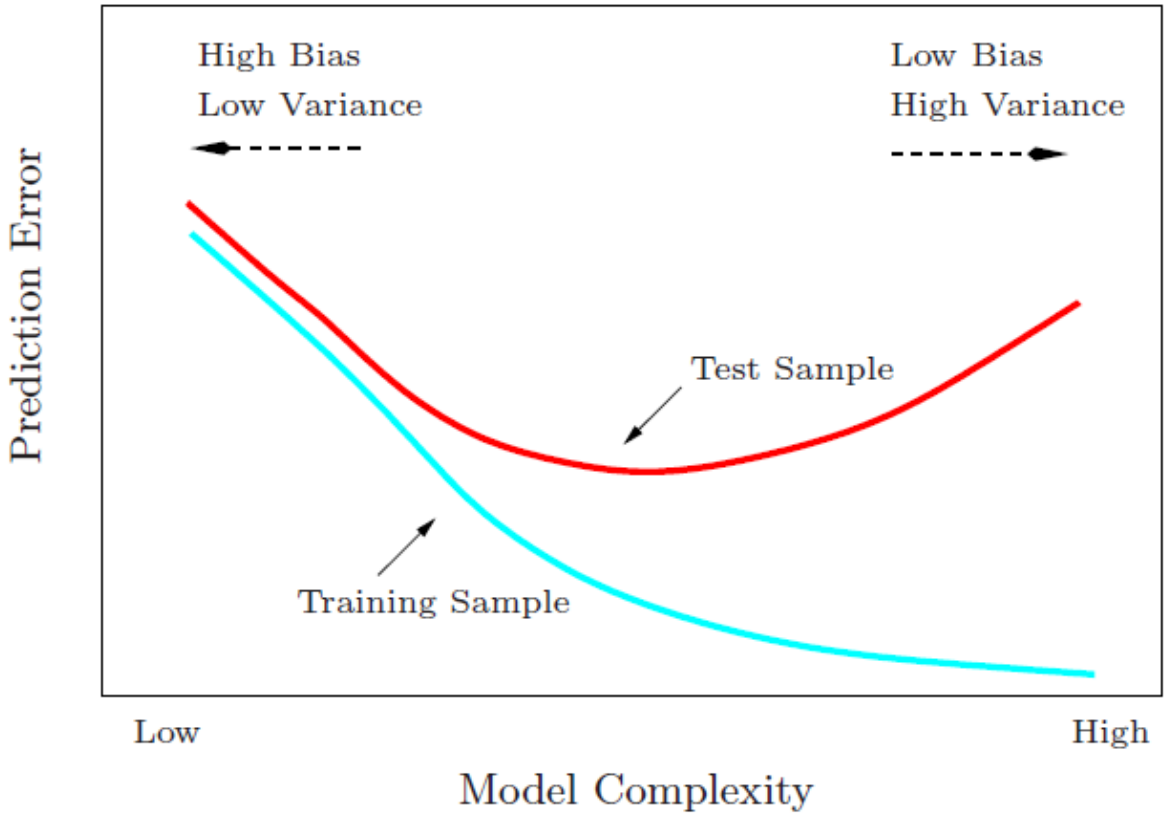


$\Rightarrow$ le risque empirique décroît de gauche à droite

- i.e., quand on augmente le degré du polynome

http://www.astroml.org/sklearn_tutorial/practical.html

Compromis biais/variance [2]



⇒ question clé de l'apprentissage supervisé

2. Image tirée de Hastie et al. (2001) (Fig.2.11)

Exemple de la régression linéaire

Méthode de base d'apprentissage supervisé :

- données d'entrée $= (x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, pour $i = 1, ..., n$.
- modèle $f_\theta(x) = \alpha + \beta x$, pour $\theta = (\alpha, \beta) \in \mathbb{R}^2$.

Exemple de la régression linéaire

Méthode de base d'apprentissage supervisé :

- données d'entrée $= (x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, pour $i = 1, ..., n$.
- modèle $f_\theta(x) = \alpha + \beta x$, pour $\theta = (\alpha, \beta) \in \mathbb{R}^2$.

$\Rightarrow$ estimation de θ par moindres carrés = minimisation du risque empirique (défini par la perte quadratique).

Exemple de la régression linéaire

Méthode de base d'apprentissage supervisé :

- données d'entrée $= (x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, pour $i = 1, ..., n$.
- modèle $f_\theta(x) = \alpha + \beta x$, pour $\theta = (\alpha, \beta) \in \mathbb{R}^2$.

$\Rightarrow$ estimation de θ par moindres carrés = minimisation du risque empirique (défini par la perte quadratique).

Inférence vs Prédiction :

- Inférence : s'intéresse (surtout) à $\hat{\theta}$
  - facteurs influents, significativité, ampleur d'effet, ...
- Prédiction : s'intéresse (surtout) à $f_{\hat{\theta}}$
  - qualité de la prédiction, généralisation

$\Rightarrow$ compromis complexité ("boîte noire") / interprétabilité

Apprentissage non-supervisé

Apprentissage non-supervisé

Données d'entrée : échantillon $\left\{x_i\right\}_{i=1,\ldots,n}$.

$\Rightarrow$ pas de variable de sortie !

Apprentissage non-supervisé

Données d'entrée : échantillon $\{x_i\}_{i=1,\dots,n}$.
$\Rightarrow$ pas de variable de sortie !

Objectif : identifier des "structures" dans les données
$\Rightarrow$ "comprendre" les données.

Apprentissage non-supervisé

Données d'entrée : échantillon $\{x_i\}_{i=1,\dots,n}$.
$\Rightarrow$ pas de variable de sortie !

Objectif : identifier des "structures" dans les données
$\Rightarrow$ "comprendre" les données.

Par exemple :

- sous-groupes dans les observations
- relations entre les variables (corrélation, redondance)
- données aberrantes
- représentations et visualisations informatives

$\Rightarrow$ souvent mené dans un cadre exploratoire
$\Rightarrow$ pas de critère objectif : fortement empirique

Apprentissage non-supervisé - illustration

Clustering :

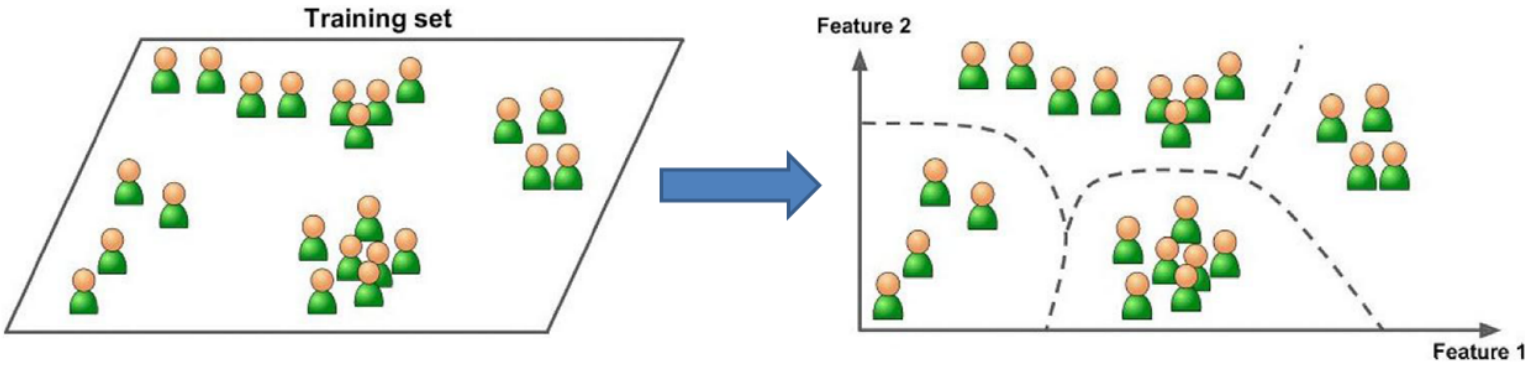


Figure: Images tirées de Géron (2017)

- catégoriser des instances en groupes homogènes
- groupes = "clusters"

Apprentissage non-supervisé - illustration

Détection de données aberrantes :

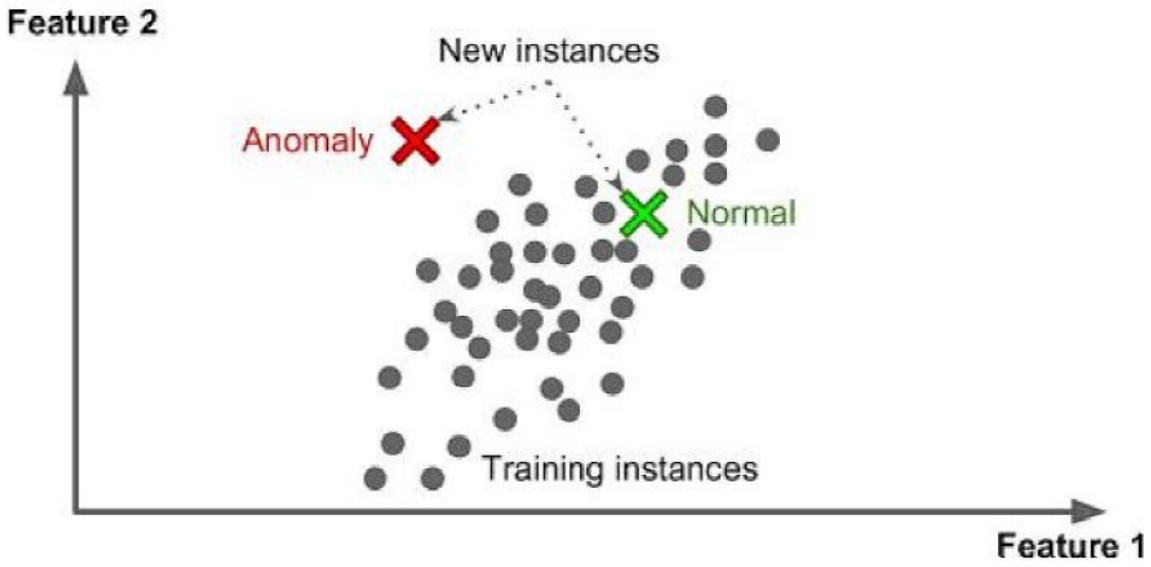


Figure: Image tirée de Géron (2017)

- données aberrante = "outlier"
- outlier / novelty / anomaly detection

Exemple de l'Analyse en Composantes Principales

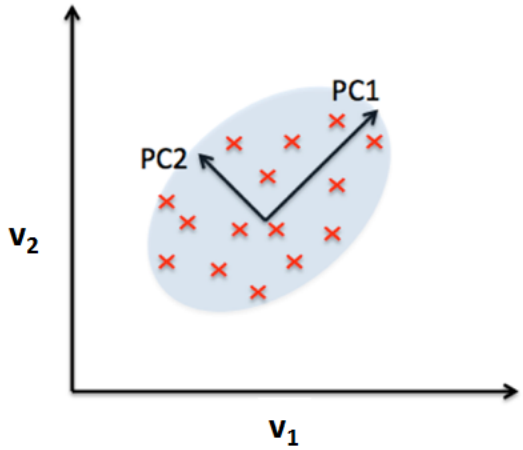
Méthode de base d'apprentissage non-supervisé :

- données d'entrée $= \{x_i\} \in \mathbb{R}^p$, pour $i = 1, ..., n$
- transforme les p variables en $\min(n, p)$ composantes principales orthogonales par combinaisons linéaires

Exemple de l'Analyse en Composantes Principales

Méthode de base d'apprentissage non-supervisé :

- données d'entrée $= \{x_i\} \in \mathbb{R}^p$, pour $i = 1, ..., n$
- transforme les p variables en $\min(n, p)$ composantes principales orthogonales par combinaisons linéaires

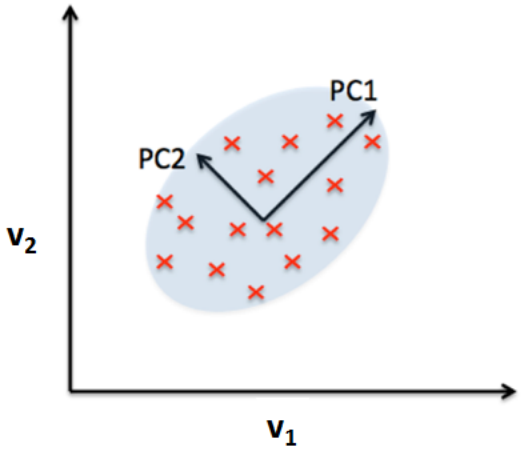


- $PC_i = a_{i1} v_1 + a_{i2} v_2$
- PC_1 = la plus forte variance
- PC_2 = la plus forte variance résiduelle

Exemple de l'Analyse en Composantes Principales

Méthode de base d'apprentissage non-supervisé :

- données d'entrée $= \{x_i\} \in \mathbb{R}^p$, pour $i = 1, ..., n$
- transforme les p variables en $\min(n, p)$ composantes principales orthogonales par combinaisons linéaires



- $PC_i = a_{i1} v_1 + a_{i2} v_2$
- PC_1 = la plus forte variance
- PC_2 = la plus forte variance résiduelle

$\Rightarrow$ quand p est grand : souvent la première étape de l'analyse

- visualisation des données, identification de structures, de sous-groupes, de données aberrantes

Conclusion

Ce cours...

Plan du cours :

1. Introduction
2. Apprentissage non-supervisé
3. Clustering hiérarchique et critères de similarité
4. k-means et modèles de mélanges
5. Apprentissage supervisé et k plus proches voisins
6. Modèles probabilistes de classification
7. Lasso et modèles pénalisés

Objectifs principaux :

- Panorama des concepts et méthodes clés du domaine
- Appréhender les bases théoriques sous-jacentes
- Savoir mettre en oeuvre ces méthodes en R

Rappels R

Opérations de base en R

- Installer des packages : `install.packages`
- Lire / écrire des tableaux de données
  - `read.table, read.csv2, read.delim`
  - `write.table, write.csv2`
- Types de base : `factor`, `data.frame`, `list`
  - facteur `f` : `levels(f), nlevels(f)`
  - liste `l` : `lapply(l, ...), sapply(l, ...)`
    - `l.dims = lapply(l, dim)` $\rightarrow$ une liste
    - `l.dims = sapply(l, dim)` $\rightarrow$ une matrice
    - `l.dims = sapply(l, function(x){dim(x)})`
  - matrice `M` : `apply(M, 1/2, ...)`
    - `row.means = apply(M, 1, mean)`
    - `col.means = apply(M, 2, mean)`
    - `col.means = apply(M, 2, function(x){mean(x)})`

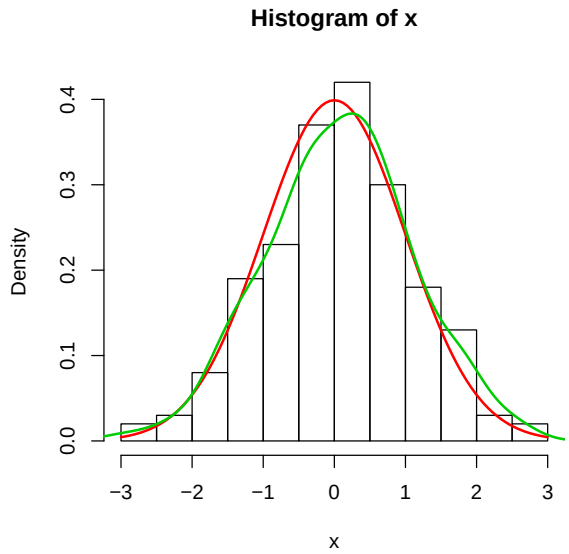
Opérations de base en R - loi normale

Distribution normale :

- tirage : `rnorm`
- densité : `dnorm` (`dnorm(0)` $= 1/\sqrt{2\pi}$)
- fonction de répartition : `pnorm` (`pnorm(0) = 0.5`)
- quantile : `qnorm` (`qnorm(0.5) = 0`)

Exemple :

```
> n = 1000
> x = rnorm(n)
> hist(x, prob=TRUE)
> curve(dnorm, add=TRUE, col=2)
> lines(density(x), col=3)
```



Opérations de base en R - régression linéaire

Régression linéaire : `lm.fit = lm(y ~ x)`

- `x` : variable(s) d'entrée
- `y` : variable de sortie / réponse

Attributs importants de l'objet `lm.fit` :

- coefficients : `lm.fit$coefficients`
- valeurs estimées : `lm.fit$fitted.values`
- résidus : `lm.fit$residuals`

Fonctions associées importantes :

- visualisation : `plot(lm.fit)`
- prédiction : `predict(lm.fit, newdata = ...)`
- résumé : `summary(lm.fit)`

Opérations de base en R - régression linéaire

Régression linéaire : `lm.fit = lm(y ~ x)`

```
> summary(fit.lm)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-30.9726  -9.5429   0.3133   9.2285  29.2748

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.6672     3.2665  -2.653   0.0103 *
x             9.8235     0.6087  16.140   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.36 on 58 degrees of freedom
Multiple R-squared:  0.8179,    Adjusted R-squared:  0.8147
F-statistic: 260.5 on 1 and 58 DF,  p-value: < 2.2e-16
```

Opérations de base en R - ACP

Analyse en Composantes Principales : `pca = prcomp(X)`

- `X` : matrice de taille n (observations) $\times$ p (variables)
- #PC $= \min(n, p)$

Attributs de l'objet `pca` :

- projections : `pca$x`
  - matrice de taille $n \times$ #PC
- axes de projection : `pca$rotation`
  - matrice de taille $p \times$ #PC
- (racine carrée de la) variance expliquée : `pca$sdev`
  - vecteur de longueur #PC

Opérations de base en R - visualisations

Visualisations :

- scatterplot : `plot(x,y)`
  - densité locale : `plot(x, y, col = densCols(x,y))`
- boîte à moustache : `boxplot(x)`, `boxplot(x ~ y)`
- Diagramme en bâtons : `barplot`
- histogramme et densité : `hist`, `plot.density`

Customisations :

- marges : `par(mar = c(bottom,left,up,right))`
- plusieurs figures : `par(mfrow = c(nrow,ncol))`
- options graphiques : `pch`, `lty`, `lwd`

Références

A. Géron. *Hands-On Machine Learning with Scikit-Learn & TensorFlow*. O'Reilly, 2017.

T. Hastie, R. Tibshirani, and J.. Friedman. *The Elements of Statistical Learning*. Springer, 2001.