MST IASD-S1 2023-2024 : Apprentissage automatique

Lecture 1: Qu'est-ce que l'apprentissage automatique?

Département Génie Informatique, FST de Tanger

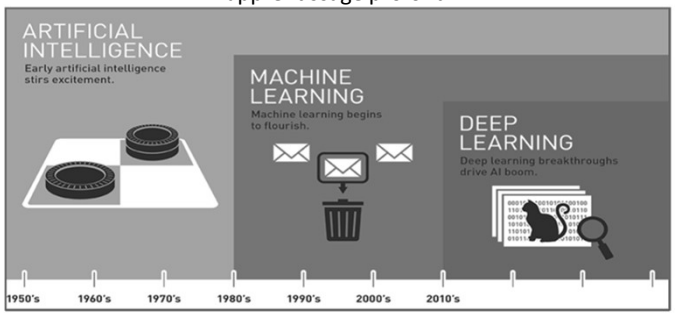
M. AIT KBIR

# Plan

- Introduction à la science des données
- Notions d'apprentissage automatique
- Bayes Naïf
- Arbres de décision
- Regression logistique
- Kmeans
- …
- Labs + 2 Devoirs (.ipynb)
- 1 CC



Intelligence Artificielle/Apprentissage automatique/ apprentissage profond



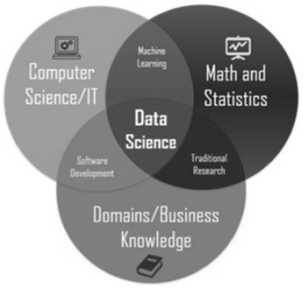

L'avantage de l'apprentissage profond, par rapport à l'apprentissage automatique, est que le programme est capable d'élaborer le jeu de caractéristiques, d'une façon non supervisé, à partir des données brutes.



# Science des données

La science des données permet l'extraction des connaissances d'un ensemble des données, ou encore de construire l'analyse au-dessus des données, on peut y trouver:

- La modélisation des données
- L'apprentissage automatique
- L'élaboration des algorithmes
- Les tableaux de bord



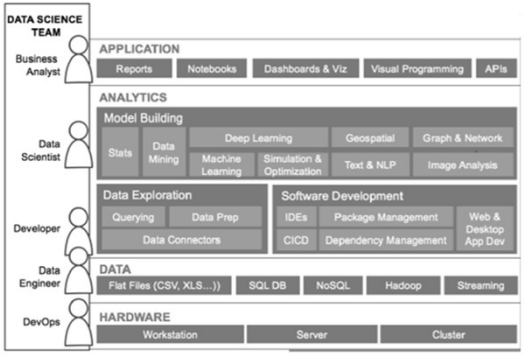

Drew Conway's Venn diagram

# Science des données : étapes clés

1. Acquisition: Acquérir des données auprès de diverses sources…
2. Exploration et compréhension: Comprendre les données à utiliser et le mode de collecte …
3. Transformation et manipulation: une des étapes les plus longues et les plus importantes …
4. Analyse et modélisation: explorer la relation entre les variables contenues dans les données et utilise une panoplie d'astuces **d'apprentissage automatique** pour regrouper, classer …
5. Communiquer et présenter: restituez les données sous une forme et une structure convaincantes …



# Science des données : équipe






# Science des données : Data Scientist

Les compétences d'un Data Scientist se situent à l'intersection de celles des mathématiciens, des informaticiens et des statisticiens. Il utilise souvent ses compétences pour construire des modèles et des algorithmes complexes pour prédire les résultats ou découvrir les motifs (Patterns) sous-jacents aux données. Ceci, dans le but d'acquérir des connaissances, qui se traduit par la capacité d'analyse des différents scénarios et la simplicité de calcul des décisions optimales.



# Apprentissage automatique

- Comment écrire des programmes qui résolvent des problèmes comme la reconnaissance des caractères manuscrits ?.

# Apprentissage automatique

- Réponse:
  - Construire des programmes qui reproduisent les mécanismes réalisés par le cerveau humain.
- Problème:
  - Peut d'informations sur comment notre cerveau réalise cette tâche.
  - Même en présence d'informations, le programme pourrait être très compliqué.
- Une autre solution consiste à recueillir un ensemble d'exemples qui spécifient la sortie correcte pour une entrée donnée.
  - Un algorithme d'apprentissage automatique exploite donc ces exemples et produit un programme qui réalise la tâche demandée.
  - Si la phase d'apprentissage se passe bien, le programme fonctionne pour les nouveaux exemples aussi bien que pour les exemples d'apprentissage.



# Apprentissage automatique: exemples d'application

Reconnaissance des formes:
- Expressions faciales
- Signature
- Images médicales

Reconnaissance des défauts ou des fraudes:
- Séquences inhabituelles de transactions par carte de crédit
- Pièces défectueuses dans une chaine de fabrication

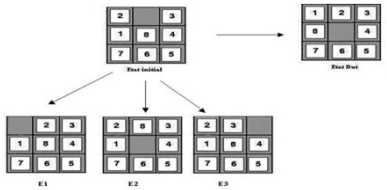
Prédiction:
- Futures cours des actions ou taux de change
- Le temps qu'il va faire dans deux prochains jours

Exemples liés au Web:
- Filtrage de spam, détection de fraude
- Systèmes de recommandation



# Apprentissage automatique vs IA symbolique

- La représentation des connaissances qui travaille avec un espace d'états et un système de production avec développement des règles d'inférence logique, l'apprentissage et l'incertitude sont généralement ignorés.
- **Exemple jeu de taquin** (arbre à explorer à l'aide d'une stratégie de recherche)**:**





# Apprentissage automatique vs IA symbolique

- Systèmes experts : à partir d'entretiens avec des experts, on extrait une base des connaissances spécifique à un domaine. Le système d'inférences logiques, logique des prédicats, établit une conclusion grâce à l'information fournie.
  - DENDRAL (en chimie) : utilisation de connaissances expertes pour déduire la structure de molécules acycliques
  - MYCIN (en médecine): système expert en infections bactérienne.
  - HERSAY II (en compréhension de la parole): Travaux sur la reconnaissance et la compréhension de la parole .
  - PROSPECTOR (en géologie) géologie, aide le géologue à évaluer l'intérêt d'un site en vue d'une prospection minière. (1600 règles).

## Statistiques vs Apprentissage automatique

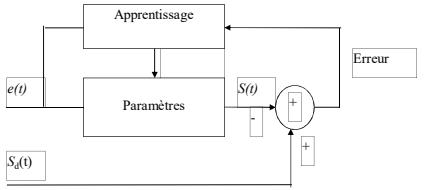
- Données de faible dimension ( moins de 100 dimensions)
- Données trop bruités
- La structure des données non complexes (représentable par des modèles simples)
- Le problème consiste en la distinction de la vraie structure du bruit.

- Grande dimensionnalité
- Le bruit n'est pas suffisent pour cacher les structures dans les données, si on les traite correctement.
- Les données comprennent beaucoup de structures, trop compliquées pour être représentées par un simple modèle.
- Mettre au point une manière pour représenter la structure complexe des données pour faciliter l'apprentissage.



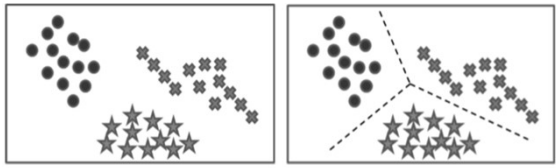
## Types d'apprentissage automatique

- Apprentissage supervisé
  - Apprendre à prédire/produire une sortie lorsqu'un vecteur entrée se présente (Les exemples d'apprentissage sont étiquetés).






## Types d'apprentissage automatique

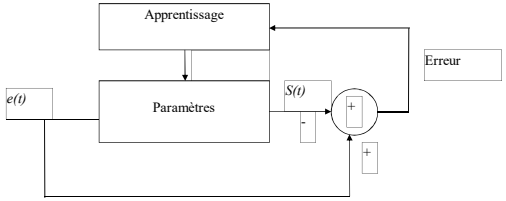


Apprentissage supervisé: trois classes d'exemples, chacun possède une classe d'appartenance (étiquette). Après un apprentissage réussi, les frontières entre les classes sont trouvées (traits en pointillés).

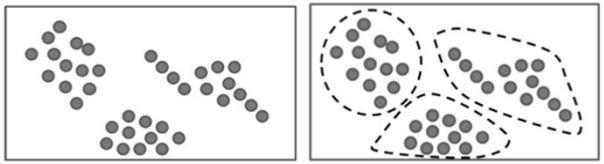


## Types d'apprentissage automatique

- Apprentissage non supervisé:
  - Créer une représentation internes des entrées par création de groupement des données (Clusters) (Les exemples d'apprentissage ne sont pas étiquetés).



## Types d'apprentissage automatique



Apprentissage non supervisé: Aucune information à priori sur les exemples. Après un apprentissage réussi, des groupements sont trouvés ( délimités par les traits en pointillés).



## Types d'apprentissage automatique

- Apprentissage par renforcement

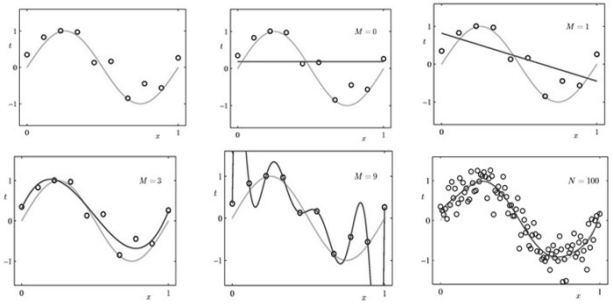
  Apprendre des actions pour maximiser les récompenses: L'apprentissage par renforcement consiste à considérer un agent autonome, plongé au sein d'un environnement, et qui doit prendre des décisions en fonction de son état courant. En retour, l'environnement procure à l'agent une récompense (un ensemble de valeurs scalaires).



## Apprentissage automatique: Généralisation

Exemple : approximation d'une fonction par un polynôme de degré M, à partir des données d'apprentissage issues de cette dernière.

$f(x)=sin(2\pi x)$






## Apprentissage automatique: Généralisation

Courbe continue montre la fonction utilisée pour générer les données ( $t=f(x)=sin(2\pi x)$+faible bruit ). Le but est de calculer $t$ pour une nouvelle valeur $x$ *par l'approximation de la fonction par un polynôme de degré* $M$:

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Ceci peut être réalisé par le calcul de du vecteur $w=(w_0, w_1, \ldots, w_M)$ qui minimise:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$$

Cette fonction mesure la différence entre $y(x, w)$ et la réponse des $N$ exemples d'apprentissage $x_n$.

M=0 et M=1 donnent une mauvaise approximation (sous-apprentissage). On remarque un bon résultat pour M=3. Pour un degré élevé, exemple M=9, l'approximation est bonne pour les données d'apprentissage, mais donne une mauvaise représentation de la fonction ailleurs, ce qui connu sous le terme sur-apprentissage. En effet, on cherche à modéliser les régularités présentes dans les données et non pas à les mémoriser.

La dernière figure correspond à N=100 et M=9, on observe une bonne approximation.

## Apprentissage automatique: Généralisation

La grande complexité du modèle par rapport aux données d'apprentissage est à l'origine du sur-apprentissage. Parmi les solutions pour réduire ce phénomène on peut prendre un nombre d'exemples élevé. Une autre solution consiste à chercher des nouveaux attributs ou à les créer à partir de ceux disponibles. Par exemple, pour augmenter la dimension d'un vecteur ($x_1$, $x_2$, $x_3$), on peut ajouter des nouveaux attributs de type polynomial($x_1$, $x_2$, $x_3$, $x_2x_1^2$, $x_1x_2x_3$, $x_3^3$).

Dans le cas du sous-apprentissage, c'est la faible complexité du modèle par rapport aux données d'apprentissage qui est la cause du problème. Parmi les solution on peut chercher à réduire la dimension des données en éliminant les attributs non influents.



## Apprentissage automatique : Généralisation

- Le but de l'apprentissage supervisé est de bien classer les exemples test non vus lors de la phase d'apprentissage.
- On cherche par l'apprentissage automatique à modéliser les vraies régularités dans les données et ignorer le bruit.
  - Comment donc être sûr que l'apprentissage automatique généralise correctement ?
- Quelques suppositions qui permettent que les régularités présentes dans la base des exemples d'apprentissages soient conservées.
  - Les exemples d'apprentissage sont extraits indépendamment de l'ensemble des exemples possibles.
  - Chaque fois qu'on extrait un exemple d'apprentissage, il vient de la même densité de probabilité.
  - Les exemples tests sont extrait de la même façon que les exemples d'apprentissages.



## Données de validation (validation croisée)

La base des exemples est devisée à trois ensembles:

- Exemples d'apprentissage pour la mise au point des paramètres du modèle.
- Les exemples de validation pour décider quel est le type de modèle et/ou le degré de régularisation qui marche mieux.
- Les exemples de test pour estimer les performances du modèle. Une estimation sans biais est obtenue par l'utilisation de cet échantillon.



## Validation croisée:


La technique dite "k-fold cross-validation", permet de diviser la base des exemples d'apprentissage en k échantillons ( ici k est égal à 4). Dans le cas simple les échantillons de même taille. k-1 groupements sont utilisés et le dernier groupe pour l'évaluation. Cette procédure est répétée pour tous les autres groupes, la performance est la moyenne des k scores.

# Mesure de performances

La matrice de confusion donne à quel point le classifieur peut identifier les exemples des différentes classes.

| | |
|---|---|
| **TP** | **FP** |
| **FN** | **TN** |

- Vrais positifs (True Positives) (TP) : exemples positifs correctement classés
- Vrais négatifs (True Negatives) (TN) : exemples négatifs correctement classés
- Faux positifs (False Positives)(FP) : exemples négatifs mal classés
- Faux négatifs (False Negatives)(FN) : exemples positifs mal classés

Les bonnes performances d'un classifieur se traduisent par une matrice de confusion où plus exemples sont présents sur la diagonale.



# Évaluation de l'apprentissage

```
Accuracy = (Tp+Tn)/(Tp+Tn+Fp+Fn)
```

Calcule le nombre de résultats corrects que la solution a réussi à identifier parmi le nombre total des cas examinés. C'est est une bonne mesure quand la base des exemples est symétrique, le nombre de cas positifs et négatifs est presque le même.

```
Precision = Tp/(Tp+Fp)
```

Une mesure qui représente la fraction des résultats positifs corrects par rapport aux résultats positifs. Mesure la capacité du système à écarter les solutions non-pertinentes.

```
Recall = Tp/(Tp+Fn)
```

Une mesure qui représente la fraction des résultats positifs corrects par rapport aux résultats corrects. Mesure la capacité du système à retourner toutes les solutions pertinentes.

```
F = 2*Precision*Recall/(Precision+Recall)
```

F-score (ou F-measure) est une mesure qui permet de faire un compromis entre la précision et le rappel en faisant approximativement la moyenne des deux valeurs.



# Apprentissage automatique: démarche

a. Extraction des données
b. Exploration des données, pour la détection de valeurs aberrantes, atypiques ou incohérentes.
c. Partition aléatoire de l'échantillon (apprentissage, validation, test)
d. Apprentissage
e. Éventuellement procéder par partitionnement aléatoire de l'échantillon et reprendre depuis l'étape c (nouvelle itération), pour améliorer l'erreur de prédiction et s'assurer de la robustesse du modèle.
f. Généralisation sur des données nouvelles.



# Apprentissage automatique

Documentation:

Livre "Machine learning in action", Peter Harrington. Voir le chapitre 4 "Classifying with probability theory: naïve Bayes" qui concerne le prochain cours.

Voir aussi:
https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html
https://numpy.org/doc/stable/user/basics.creation.html