APPRENTISSAGE AUTOMATIQUE

sept 2021

Apprentissage automatique

- ~~Approches symboliques~~
- Approches statistiques
  - Représentation numérique des données par un ensemble de caractéristiques
  - Utilisation d'algorithmes d'apprentissage statistique

  → Acquérir une connaissance sur les données : combinaison entre algorithme et données d'apprentissage
- Apprentissages
  - Supervisé
  - Non-supervisé
  - ~~Par renforcement~~

Types d'apprentissage

- **Apprentissage supervisé**
  - À partir de données annotées (par des humains)
  - Apprendre pour annoter de nouvelles données
  - Problèmes de classification, régression ou de segmentation/étiquetage
- Apprentissage non supervisé
  - À partir de données non annotées
  - Apprendre pour créer des annotations sur ces données
  - Problèmes de clustering

Types d'apprentissage

- Apprentissage semi-supervisé
  - À partir de données annotées et non-annotées
  - Apprendre pour annoter de nouvelles données
  - Problème de classification et de clustering
- Apprentissage par renforcement
  - À partir d'une situation donnée, d'un ensemble d'expérience et d'un ensemble d'actions possibles
  - Évaluer la meilleure décision à prendre (récompense)

Faiblement supervisé …

Apprentissage supervisé

- Deux types de techniques:
  - Inductives
    - Apprentissage : Construction d'un modèle
  - Transductives
    - Sans apprentissage : Classement effectué en fonction des données déjà classées
    - Manipulation de toutes les données lors d'un nouveau classement

Apprentissage automatique

- Trois grands temps
  - À partir d'un ensemble de **données**
    - Description *pertinente* des données
  - Mise en œuvre *efficace* d'un **algorithme**
    - Fonction du type de problème à résoudre
  - **Evaluation** et/ou analyse des résultats obtenus
    - Fonction de l'application visée

Les Données

Qu'est-ce qu'une donnée?

- *Instance* de la *population* caractérisée par un ensemble de *descripteurs*
- Représentation plus formelle
  - *x* une donnée de l'ensemble des données *X*
  - chaque donnée *x* est définie par *p* descripteurs
  - chaque descripteur *d* prend sa valeur dans *Vd*
  - Toute donnée appartient alors à un espace euclidien à p dimensions

Types de descripteurs (1/2)

- Descripteurs qualitatifs
  - Variable discrète
    - Ensemble de valeurs prédéfinies
  - Pas d'application d'opérations arithmétiques habituels
  - Exemples :
    - une couleur, une marque, une ville, …
  - Nature de la valeur :
    - nominale
      - Ensemble de valeurs arbitraires, incomparables *a priori*
      - Ex couleur : rose et orange

Types de descripteurs (2/2)

- Descripteurs quantitatifs
  - Type : entier, réel, date
    - ≠ numérique et réciproquement
  - Possibilité d'appliquer des opérateurs arithmétiques habituels
  - Nature de la valeur :
    - Ordinale
      - Ensemble de valeurs arbitraires mais comparables SELON une unité de mesure
    - Absolue
      - Ensemble de valeurs non arbitraires

Sélection des descripteurs

- Pertinence
  - Importance de la sélection des descripteurs en fonction de **l'application visée**

    → Définir le problème et les objectifs
- Préparation des données
  - Inventaire, collecte et intégration
  - Sélection
    - Suppression d'individus
    - Suppressions de descripteurs
    - Création de descripteurs

Sélection des descripteurs

- Fiabilité
  - Représentation complète
    - Tous les descripteurs pour toutes les données ?
  - Validité des valeurs des descripteurs
    - Une donnée peut être ? bruitée ?
- Quantité
  - Peu : apprentissage simplifié… performance?
  - Beaucoup : apprentissage complexe… performance?

Organisation des données

- CORPUS : Ensemble des données disponibles
- Corpus d'apprentissage (APP)
  - Entraînement du modèle
- Corpus de développement (DEV) (facultatif)
  - Optimisation des paramètres d'ajustement du modèle (si nécessaire)
- Corpus de test (TEST)
  - Évaluation des performances du modèle en généralisation

!! La taille est critique …

L'algorithme – La construction du modèle

Apprentissage supervisé

- Création automatique d'un *modèle* à partir d'un corpus de données d'apprentissage *annotées*
  - Prédire une classe par donnée « connue »
  - Généraliser : Prédiction sur une donnée non connue

→ Modèle permettant d'associer à toute donnée correctement décrite une valeur définie

Classification supervisée

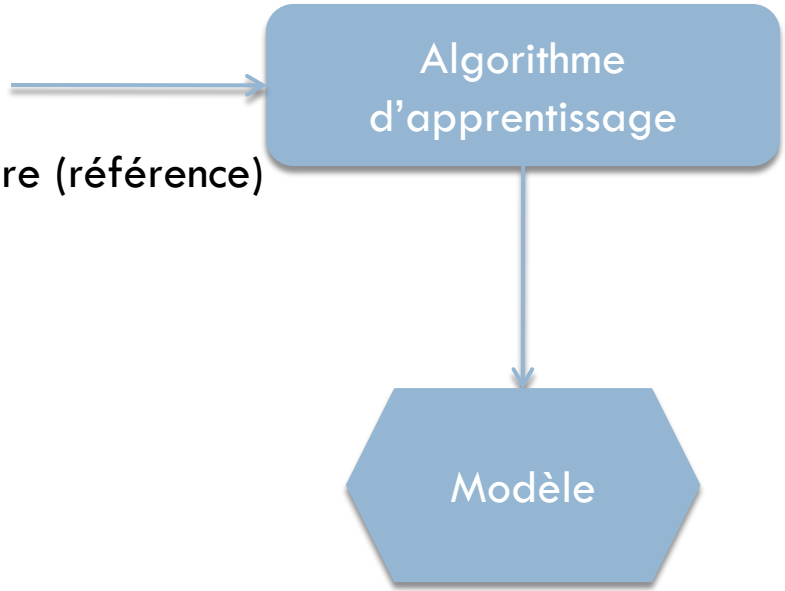
Plus formellement :

- Ensemble de couples donnée/classe : (x_i, y_i)
  - Avec $x_i \in X$, l'ensemble des données d'apprentissage
  - Avec $y_i \in Y$, l'ensemble des classes à prédire
  - Tel que : $y_i = f(x_i) + w_i$ (w_i bruit de mesure)
- Construction d'un modèle
  - Déterminer la représentation compacte de f par g appelée fonction de prédiction.
  - Tel que : $y_i = g(x_i) + \varepsilon_i$, ε_i erreur de prédiction

Classification supervisée

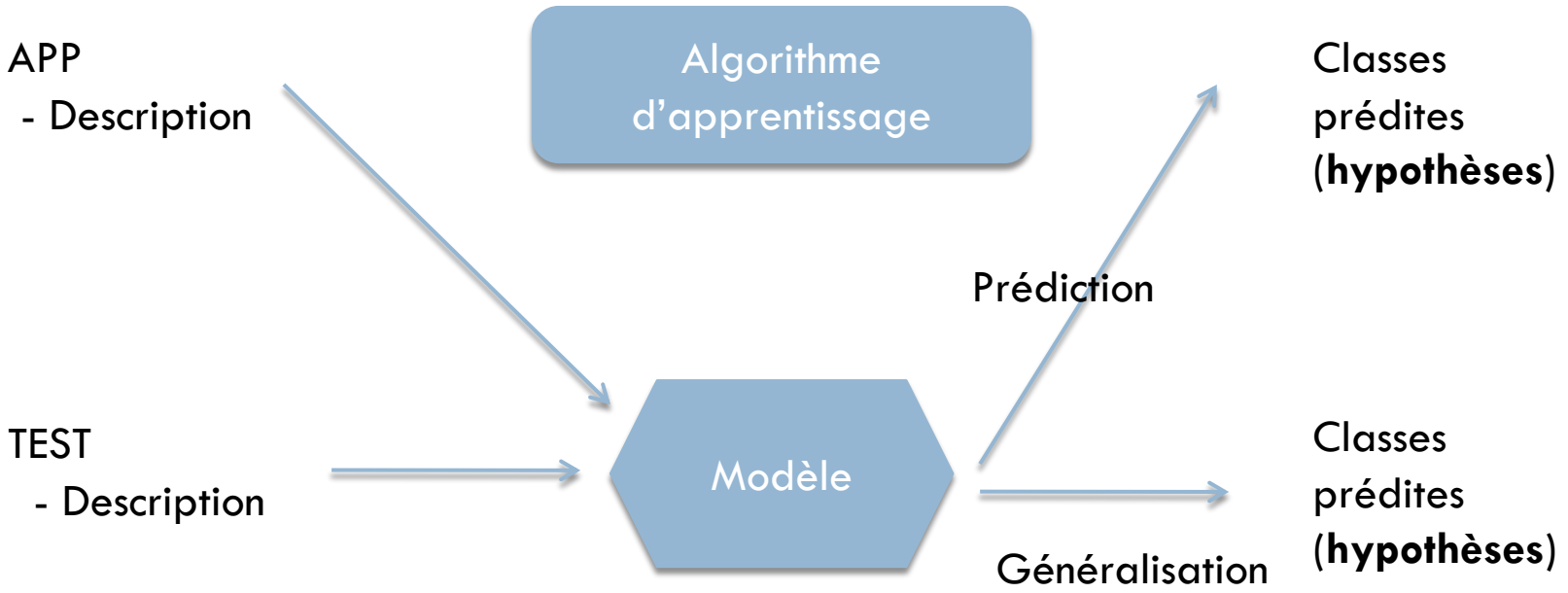
- Apprentissage

APP
- Description
- classe à prédire (référence)



Classification supervisée

- Classement (ou test)



Domaine de définition des classes

- Y est un ensemble fini : Problème de *Classification*
  - Associer une donnée à une valeur discrète parmi plusieurs classes prédéfinies
  - **Classification binaire** : ¥ $\{0,1\}$
  - **Classification multi-classes** : ¥ $\{0,1,...,I\}$

- Y est un ensemble infini : Problème de *régression*
  - Associer une donnée à une valeur continue
  - **Régression** : $Y \subset \Re$

Classification multi-classes : Cas particulier

- Possibilité d'associer plusieurs classes à une seule donnée
  - *Ensemble de classes discrètes non exclusives*

$$Y = \{a, b, c, d, ...\}$$

  - *Si une donnée n'est associée qu'à une seule classe*
    - Classification ***uni-label***
  - *Si une donnée peut être associée à plusieurs classes*
    - Classification ***multi-labels***
- Cas proposé par peu d'algorithmes
- Correspond souvent à plusieurs classifications binaires.

À propos du modèle

- Peut-être considéré comme une boite noire
  - Simple utilisateur… mais
- Selon l'algorithme choisi :
  - Différentes représentations possibles
  - Différents paramètres à ajuster

→Meilleur choix et optimisation de l'apprentissage si on connaît l'algorithme

L’évaluation

Validation classique des résultats

- Cas classique : Assez de données annotées
  - Ex : 1 APP (70%) et 1 TEST (30%)
  - Estimation de l'erreur de prédiction
    - Évaluation du modèle sur l'APP
    - Taux de mauvaise classification sur l'APP
  - Estimation de l'erreur de généralisation
    - Evaluation du modèle sur le TEST
    - Taux de mauvaise classification sur le TEST
- Remarque : mise en production
  - Ré-apprentissage du modèle sur TOUT le corpus annoté

Mesures de performance du modèle

- Évaluation de l'erreur
  - Soit le couple (x_i, y_i), y_i classe de **référence**
  - Soit le modèle g
  - Soit **l'hypothèse** $y'_i = g(x_i)$

  Est-ce que $y'_i = y_i$?
- Comment évaluer l'erreur?
  - Dépend de l'objectif de l'application visée

Matrice de confusion

- Aussi appelée Tableau de contingence
- Représentation des données en fonction de leur association à la classe… mais laquelle?
- Alignement des données REF et HYP
  - Hypothèse : HYP
  - Référence : REF
- Remplir par comptage un tableau
  - Chaque donnée doit appartenir à l'effectif d'une case

Mesures classiques globales

- Taux de bonne classification

$$\underset{\text{(Accuracy)}}{\text{Acc}} = \frac{\#\text{ instances bien classées}}{\#\text{ instances classées}}$$

- Taux d'erreur (mauvaise classification)

$$\underset{\text{(Classication Error Rate)}}{\text{CER}} = \frac{\#\text{ instances mal classées}}{\#\text{ instances classées}}$$

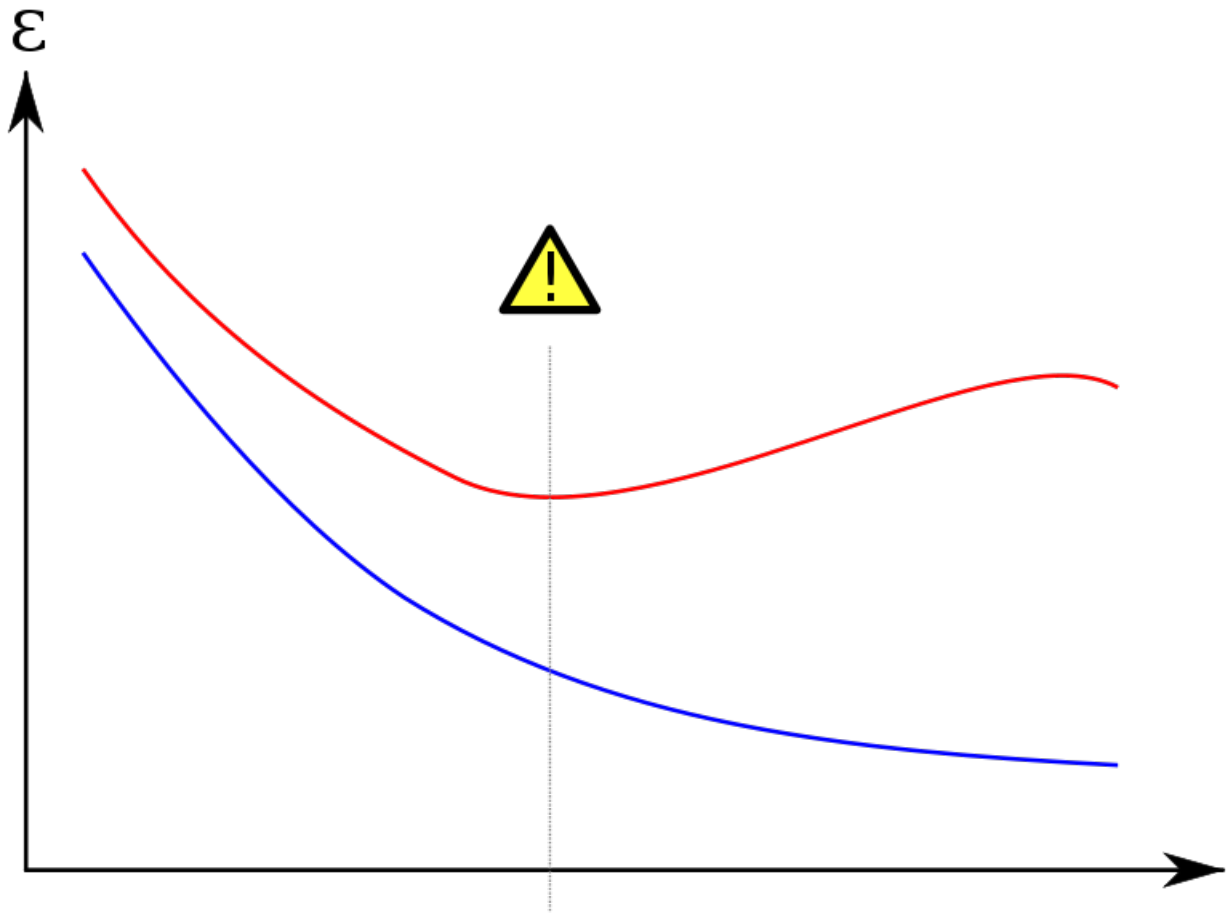
Problème de sur-apprentissage

- Que signifie MEILLEUR choix des paramètres?
- Quand « arrêter » d'apprendre?
- Comment choisir la représentation ?

Problème de sur-apprentissage

- Deux critères à considérer
  - Erreur de prédiction
  - Erreur de généralisation
- Quand « arrêter » d'apprendre?
  - Erreur de prédiction diminue ET l'erreur de généralisation augmente
- Comment faire?
  - Taille du corpus d'apprentissage
  - Paramètres d'ajustement du modèle

Problème de sur-apprentissage



Confiance dans l'estimation de l'erreur?

- Erreur = variable aléatoire
  - Après classification, 2 valeurs possibles pour la donnée : bien ou mal classé
    - Erreur = probabilité de l'événement « mal classé »
  - En déterminer la moyenne? Un intervalle?
- CER : Calcul de l'Erreur sur *1* corpus de test
  - Sur 100 exemples de test, 15 sont faux
    - Le taux d'erreur du système est de 15%?

Intervalle de confiance

- Estimation du taux d'erreur *réel E* du système à partir du taux d'erreur observé *CER* sur un ensemble de test T
  - Approximation de la loi binomiale par la loi normale
    Intervalle de confiance à 95%
  - On estime l'erreur par l'intervalle de confiance :

$$CER \pm 1.96 \sqrt{\frac{CER.(1-CER)}{\#instances\ classées}}$$

  - !! Nombre d'exemples du jeu de test suffisant

Mesures en Recherche d'Information

- Précision : pourcentage de documents pertinents

$$\text{précision}_i = \frac{\#\text{ instances correctement classées i}}{\#\text{ instances classées i}}$$

$$\text{précision} = \frac{\sum_i \text{précision}_i}{\text{nombre de classes}}$$

- Précision élevée, moins de bruit

Mesures en Recherche d'Information

- Rappel : pourcentage de documents pertinents retrouvés

$$\text{rappel}_i = \frac{\#\text{ instances correctement classées i}}{\#\text{ instances réellement i}}$$

$$\text{rappel} = \frac{\sum_i \text{rappel}_i}{\text{nombre de classes}}$$

- Rappel élevé, moins de silence

Mesures en Recherche d'Information

- F-mesure : combinaison de la précision et du rappel

$$fmesure = \frac{(1+\beta^2)rappel * précision}{\beta^2(rappel + précision)}$$

généralement $\beta=1$

Données d'apprentissage

- Le point sensible de l'apprentissage automatique
  - Nécessité *suffisamment* de données annotées
  - *Suffisamment?* Dépend de :
    - La difficulté de la tâche
    - La complexité de représentation des données
- Problème:
  - L'annotation du corpus d'apprentissage/test est humaine

  → Coût très élevé
- Mais les méthodes ont fait leurs preuves!

Validation croisée

- Problème : manque de données annotées
- Approche par « leave one out »
  - Soit un ensemble de *P* données annotées
  - Construction de *P* modèles différents sur (APP-1donnée)
  - Test de chacun des modèles sur la donnée mise de côté
- Généralisation au « N-fold »
  - Découpage de l'APP en N sous-ensembles distincts
  - Apprentissage sur N-1 fold et test sur le fold restant
- Erreur : moyenne des erreurs de chaque fold

Difficultés inhérentes à l'apprentissage

- Données en entrées …
  - Nombre d'exemples trop faible p/r nombre de descripteurs
  - Ensemble des descripteurs incomplet pour caractériser les concepts
  - Données « bruitées » : fausses ou mal étiquetées
- L'algorithme d'apprentissage fonctionne mal …
  - Mauvais paramétrage du système
  - Impossibilité d'apprendre les concepts
- L'évaluation n'est pas satisfaisante …

Conclusion

- Beaucoup de choix …
  - Choix de la représentation des données
  - Choix de l'algorithme utilisé
  - Choix de la répartition des données, de la méthode d'évaluation
- **Tout ceci dépend de l'application visée**
  - **Bien définir le problème, les objectifs, les ressources**
- ça marche souvent bien … encore faut-il avoir suffisamment de données d'apprentissage de qualité suffisante

Quelques sources

Livres :

- « Apprentissage artificiel, concepts et algorithmes », A.Cornéjuols et L.Miclet

Cours sur le web :

- http://www.grappa.univ-lille3.fr/~ppreux/fouille/
- http://www.dsi.unive.it/~marek/files/06%20-%20datamining.pdf
- http://www.public.asu.edu/~jye02/
- http://freedownloadb.com/ppt/data-mining-data-warehousing-lecture-notes