IA et science des données

Cours 8 – mardi 22 mars
Ensembles. Apprentissage non-supervisé

Christophe Marsala
Vincent Guigue

Sorbonne Université

LU3IN026 - 2021-2022

Plan du cours

L'approche BAGGING

- Bootstrap AGGregatING
- Construire un ensemble de classifieurs de même type
- Agréger leurs résultats lors d'une classification
- $\Longrightarrow$ approche très efficace !
  - la variance globale est plus faible que la variance de chaque classifieur
- Si les classifieurs sont des arbres de décision : forêt

L'approche BAGGING : apprentissage et classification

Apprentissage :

- Soit $\mathbf{X}$ une base d'apprentissage avec n exemples
- Soit B le nombre de classifieurs souhaités et $m < n$ le nombre d'exemples à choisir
  1. Extraire B sous-bases de $\mathbf{X}$: $\mathbf{X}_1, \ldots, \mathbf{X}_B$
     - sélection aléatoire de m exemples de $\mathbf{X}$
     - avec ou sans remise
  2. Construire un classifieur f_k pour chaque sous-base $\mathbf{X}_k$
- Au final : on obtient un ensemble de B classifieurs $f_1, \ldots, f_B$

Classification :

- Soit un ensemble de B classifieurs $f_1, \ldots, f_B$
- Soit un exemple $\mathbf{x}$ à classer
  1. calculer $f_k(\mathbf{x})$ pour chaque classifieur $k = 1, \ldots, B$
  2. classe finale prédite de $\mathbf{x}$: classe majoritaire parmi les $f_k(\mathbf{x})$

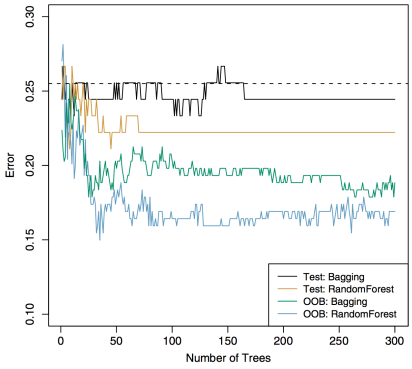
Les forêts aléatoires (random forest)

- Idée : plus les arbres sont diversifiés, meilleur sera le score global
- Augmenter la diversité : choisir aléatoirement les variables à utiliser !
- Bagging modifié : random forest
- Soit $\mathbf{X}$ une base d'apprentissage avec n exemples
- Soit B le nombre de classifieurs souhaités, $m < n$ le nombre d'exemples à choisir et $p \leq d$ variables de description à choisir
  1. Extraire B sous-bases de $\mathbf{X}$: $\mathbf{X}_1, \ldots, \mathbf{X}_B$
     - sélection aléatoire de m exemples de $\mathbf{X}$
     - sélection aléatoire de p variables de descriptions
  2. Construire un classifieur f_k pour chaque sous-base $\mathbf{X}_k$
- Remarque : B, m et p sont des hyper-paramètres de l'algorithme

Evaluation d'ensembles

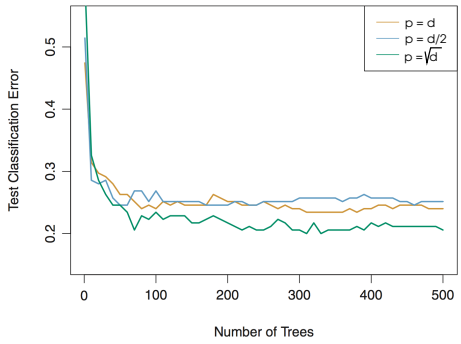
- Pour évaluer un ensemble construit par Bagging / random forest
- Validation croisée
  - très coûteuse pour évaluer un ensemble
  - il faut construire B classifieurs à chaque fois !
- Evaluation Out Of the Bag (OOB)
  - adaptée aux ensembles et suffisante pour les évaluer
  - évaluer f_k sur les exemples de $\mathbf{X}$ non sélectionnés pour le construire
  - chaque $\mathbf{x}$ est évalué par les f_k pour lesquels il n'a pas été utilisé en apprentissage
  - compter le nombre de fois où il est bien classé sur le nombre de fois où il est classé

Performances bagging vs random forest



(source : "An introduction to statistical learning", Gareth et al.)

Performances random forest : choix du nombre d'attributs



(source : "An introduction to statistical learning", Gareth et al.)

Ensembles de classifieurs

- Approches de construction d'ensembles
  - bagging : bootstrap aggregating
  - random forests
- Evaluation :
  - validation croisée
  - approche Out Of Bag

Plan du cours

Rappels : notations (1)

- Ensemble de n exemples (ou cas, ou individus) : $\mathbf{x}_1, \ldots, \mathbf{x}_n$
  - chaque individu $\mathbf{x}_i$ est décrit par d variables.
    $x_{i,j}$ (ou x_{ij}) est la valeur de la variable j pour l'exemple $\mathbf{x}_i$
- Base d'apprentissage
  - ensemble d'exemples $\mathbf{X} \in \mathbb{R}^{n \times d}$

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,d} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,d} \end{pmatrix}$$

  - apprentissage supervisé : chaque $\mathbf{x}_i$ est associé à un label y_i
    - ensemble de labels associés à $\mathbf{X}$

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

    - classification binaire : $y_i \in \{-1, +1\}$

Rappels : notations (2)

- Pour un seul exemple : $\mathbf{x} = (x_1, x_2, \ldots, x_d)$
- Terminologie : un label y_i = une classe
- Classifieur f : $f(\mathbf{x})$ est la classe donnée par f à l'exemple $\mathbf{x}$
  - cas binaire :
    - $f : \mathbb{R}^d \longrightarrow \{-1, +1\}$
      $\mathbf{x} \longmapsto f(\mathbf{x})$
    - ou aussi : $f : \mathbb{R}^d \longrightarrow \{l_1, l_2\}$ avec l_1 et l_2 deux labels donnés
  - cas multiclasses :
    - $f : \mathbb{R}^d \longrightarrow \{l_1, l_2, \ldots, l_k\}$

Exemple : classifier des emails

- Objectif : construire un modèle pour détecter un spam
  - à partir d'un ensemble d'emails déjà reçus
- Description pour chaque email : valeurs d'attributs
  - statistiques sur son contenu
  - pourcentage de mots qui sont référencés comme caractéristiques de spam (par exemple : *money*,...), utilisation d'un dictionnaire
  - nombre moyen de lettres consécutives en majuscules
  - syllabes présentes, ...
- Base d'apprentissage : constituée d'emails déjà reçus
  - 4601 emails répartis en 2 catégories : 1813 spams, et 2788 non spams
  - Par exemple : spam (1) / non spam (-1)
    - 0,0.64,0.64,0.32,0,0,0,0,0,0,0.64,$\cdots$,0.778,0,3.756,61,278 $\longrightarrow$ 1
    - 0.49,0,0.49,0.49,0,0,0,0,0,0,0.99,$\cdots$,0.091,0,1.214,5,51 $\longrightarrow$ -1

Apprentissage artificiel

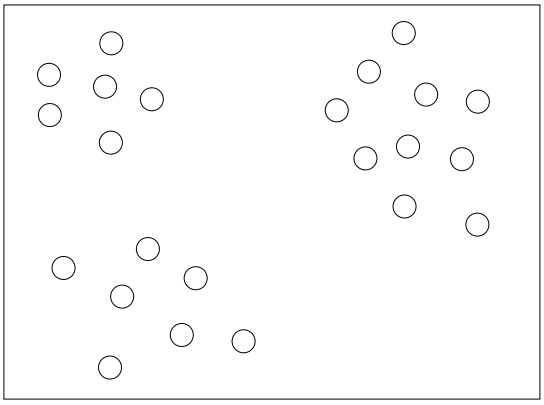
- Ensemble de données décrites par des attributs
  - la description de la donnée est fournie
  - les attributs fournissent des valeurs connues pour la description, ils sont mesurables ou observables
- Éventuellement : existence de classes
  - une description peut être associée à une classe
  - la classe est une catégorie, une variable particulière, souvent non mesurable ou non observable directement
- Classification des descriptions, deux situations possibles :
  - soit une classe est connue pour certaines données
    - apprentissage supervisé
  - soit il n'y a pas de classe connue mais on souhaite former des groupes de données qui se ressemblent
    - apprentissage non supervisé

Exemple d'apprentissage non supervisé : groupes d'emails

- Objectif : grouper des emails selon leur ressemblance
- Description pour chaque email : statistiques sur son contenu
  - pourcentage de mots pertinents (selon le domaine)
  - nombre moyen de lettres consécutives en majuscules
  - ...
- Aucune classe fournie
  - des classes seront mises en évidence à l'issue de l'apprentissage
- Par exemple :
  - email 1 : 0,0.64,0.64,0,0.32,0,0,0,0,0,0,0.64,0,0,0,0.32,0,1.29,1.93,0,0.96,0,0,0,0,0,0,0,0,0,0,0,0,0, 0,0, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0.778,0,0,3.756,61,278
  - email 2 : 0.49,0,0.49,0,0.49,0,0,0,0,0,0,0.99,0,0,0,0,0,0,0.99,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0, 0.49,0,0,0,0,0,0,0,0.99,0.49,0,0,0,0,0,0.091,0,0,1.214,5,51

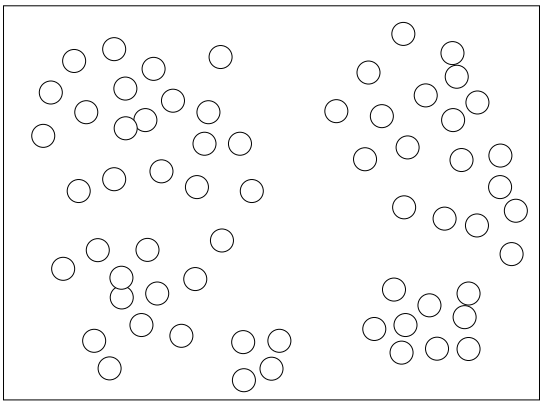
Un petit exemple

- Un ensemble de données quelconque : combien de groupes ?



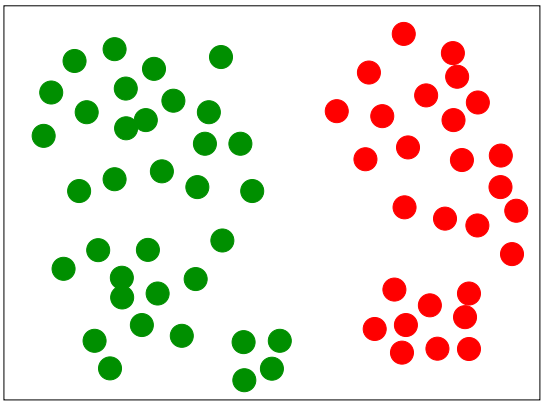
Un autre petit exemple

- D'autres données quelconques : combien de groupes ?



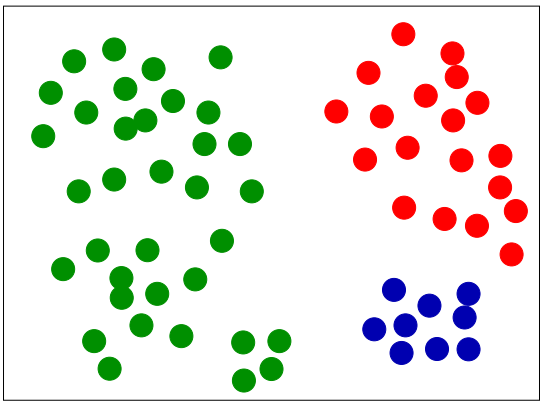
Un autre petit exemple

- D'autres données quelconques : 2 groupes ?



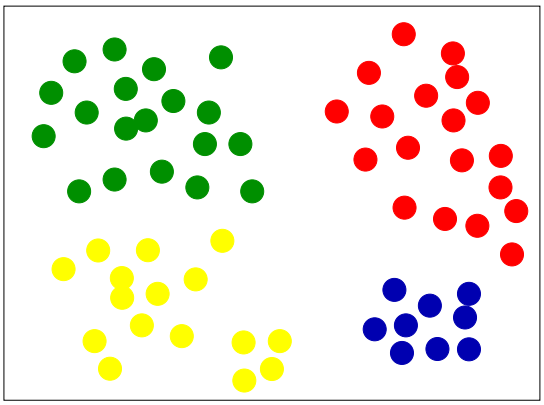
Un autre petit exemple

- D'autres données quelconques : 3 groupes ?



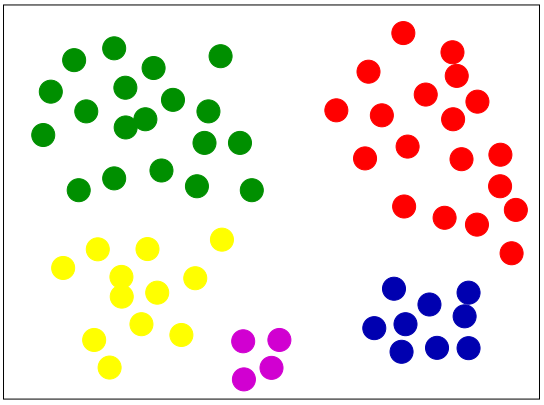
Un autre petit exemple

- D'autres données quelconques : 4 groupes ?



Un autre petit exemple

- D'autres données quelconques : 5 groupes ?



L'apprentissage non supervisé

1. Phase d'apprentissage
   - construction d'un modèle
   - utilisation d'un ensemble de données d'apprentissage
     - créer des groupes homogènes selon les descriptions
     - mettre en évidence des ressemblances
   - évaluation du modèle obtenu
     - groupes homogènes ?
2. Phase de test
   - validation du modèle
   - utilisation sur un ensemble de données de référence
     - vérifier que les groupes restent homogènes
3. Phase d'utilisation
   - mise en œuvre du modèle
   - utilisation sur des données quelconques

La classification en apprentissage non supervisé

- Classification : trouver des classes de descriptions
- Un ensemble de données sans classe connue
  - on recherche à faire des regroupements de descriptions similaires
  - on souhaite mettre en évidence des classes, des catégories
- But : former des groupes de données qui se ressemblent
  - clustering : faire des groupes parmi les données
  - cluster : ensemble de données regroupées ensemble
- Exemple :
  - le clustering hiérarchique
  - l'algorithme des K-moyennes

Le clustering hiérarchique

- But : obtenir des groupes d'exemples
- Idée : grouper petit à petit les exemples qui se ressemblent
- Question : Comment mesurer la ressemblance entre 2 exemples ?
- On possède un espace de représentation des exemples
  - calculer des distances entre les exemples
  - deux exemples se ressemblent d'autant plus qu'ils sont proches
- Mesurer une distance : fonction $d : \mathbf{X}^p \times \mathbf{X}^p \rightarrow \mathbb{R}^+$
  - séparation : $\forall x, y \in \mathbf{X}^d$, $d(x, y) = 0$ ssi $x = y$
  - symétrie : $\forall x, y \in \mathbf{X}^d$, $d(x, y) = d(y, x)$
  - inégalité triangulaire : $\forall x, y, z \in \mathbf{X}^d$, $d(x, y) + d(y, z) \geq d(x, z)$
- À connaître : distance ultramétrique
  - on remplace l'inégalité triangulaire par $\forall x, y, z \in \mathbf{X}^d$, $\max(d(x, y), d(y, z)) \geq d(x, z)$

Mesurer la distance entre 2 clusters

- Utiliser une distance entre 2 exemples : $d(x_1, x_2)$
  - Euclidienne, Manhattan, Minkowski, "infinie",...
  - étape de normalisation nécessaire
- Distances entre 2 clusters : $dist(A, B)$
  - A) complete linkage
  - B) average linkage
  - C) simple linkage
  - D) centroid linkage
- Centre de gravité (centroid) d'un cluster