IA et science des données

Cours 10 – mardi 5 avril 2022

Christophe Marsala
Vincent Guigue

Sorbonne Université

LU3IN026 - 2021-2022

Programme du jour

Le projet

Apprentissage non-supervisé

Plan du cours

Le projet

Apprentissage non-supervisé

Exemples d'études (non exhaustif...)

- Choisir une classe à prédire
  - est-ce qu'une application est populaire ?
  - prédire le prix
  - ...
- Trouver des clusters
  - qu'est-ce qui rend populaire une application ?
  - caractéristiques des applis beaucoup téléchargées
  - ...
- Algorithmes de base
  - supervisé : kppv, perceptron (Rosenblatt, biais), arbres de décision (numériques
    - données catégorielles et numériques
    - classe binaire et multi-classes
  - non-supervisé : K-moyennes (notebook séance 10)

Évaluation d'un classifieur : bonnes pratiques

- Train vs Test
- Validation croisée

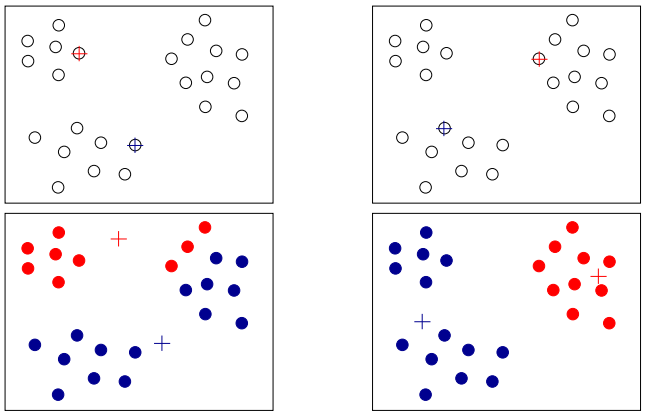
Plan du cours

Le projet

Apprentissage non-supervisé
- évaluation du résultat
- en pratique
- conclusion

Clusters différents au final

- Le choix initial des centres est important ! (ici avec $K = 2$)



Algorithme K moyennes (2)

- Prérequis
  - X : un ensemble de données (base d'apprentissage)
  - un entier naturel $K > 0$ (le nombre de clusters à trouver)
  - une mesure de distance d entre deux exemples x et y : $d(x, y)$
- Algorithme :
  1. choisir aléatoirement K exemples dans X comme premiers centres de clusters c_1, c_2,..., c_K
     - chaque centre c_k définit un cluster C_k
  2. affecter chaque x de X au cluster dont il est le plus proche
     - calculer $d(x, c_1)$,..., $d(x, c_K)$
     - affecter x au cluster C_k pour lequel $d(x, c_k)$ est la plus petite
  3. mettre à jour les centres des clusters
     - c_k est la moyenne des descriptions du cluster C_k
  4. retourner à l'étape 2 jusqu'à ce que l'inertie globale ne change plus beaucoup
- Résultat
  - un ensemble de clusters C_1,..., C_K

Évaluation du résultat d'un clustering

- Évaluer la partition obtenue : mesurer sa qualité
  - différentes approches
  - utilisation des caractéristiques des clusters
- Compacité d'un cluster
  - évaluer combien les exemples sont proches les uns des autres
  - compacité intra-cluster
- Séparabilité des clusters
  - évalue combien les clusters sont éloignés les uns des autres
  - distance inter-clusters
- Mesure globale : index d'une partition
  - index de Dunn
  - index de Xie-Beni
  - ...

Les K-moyennes en pratique

- Algorithme très simple à mettre en œuvre
  - algo ancien : James McQueen 1967
  - mais encore très utilisé !
  - nombreuses variantes...
- Quelques problèmes
  - quelle valeur pour K ?
  - convergence : la stabilisation peut être très longue à venir
  - sensible au choix initial des centres
  - quelle mesure de distance ?

K-moyenne : ce que l'on a vu

- Objectif : trouver une partition en K groupes (ou clusters)
  - bonne partition : minimise l'inertie globale intra-cluster
  - mesure de l'inertie d'un cluster : densité autour de son centre
- Algorithme : itérations successives jusqu'à convergence
  - affectation des exemples aux clusters
  - mise à jour des centres
  - arrêt si convergence ou itérations max