Trois domaines de l'analyse de données

I : Classification (Approche supervisée)

Il s'agit de la découverte de règles permettant de ranger des objets ou individus (c'est-à-dire des données représentées par des vecteurs multi-dimensionnels) dans des classes prédéfinies.

Exemples :

- Assurance ou Médecine : analyse de risques. Deux classes pourraient être Y1 = "Individu à risque" et Y2 = "Individu sans risque" que l'on cherche à prédire à partir de données médicales ou sociologiques (pour un individu i ces données sont formalisées par un vecteur multidimensionnel Xi).
- Marketing : segmentation de clients pour le marketing direct. Deux classes pourraient être Y1="Client à fort potentiel" et Y2="Client à faible potentiel" que l'on cherche à prédire à partir de données personnelles (pour un client i ces données sont formalisées par un vecteur multidimensionnel Xi).

Principe de la classification :

Il faut au départ disposer d'un jeu de données dont la classe de chaque donnée est connue (ensemble de valeurs connues appelé vérité terrain ou Ground Truth).
On fonctionne alors en trois étapes :

- construction d'un modèle de prédiction basé sur une certaine proportion (par exemple les trois quarts) du jeu de données disponibles : c'est ce qu'on appelle la phase d'entraînement ou d'apprentissage.
- validation du modèle obtenu en le testant sur la proportion restante (par exemple le quart restant) du jeu de données disponible.
- prédiction de la classe – inconnue – de nouvelles données.

Efficacité d'une méthode de classification :

Au final, l'efficacité d'une méthode tient essentiellement à deux choses :

- la vitesse d'apprentissage c'est-à-dire la complexité des algorithmes utilisés.
- la qualité de l'apprentissage c'est-à-dire la proportion de prédictions correctes.

Algorithmes usuels :

Il existe de nombreux algorithmes permettant d'obtenir des modèles de prédiction. Plusieurs nous ont été présentés : l'algorithme des k-plus proches voisins (ou k-NN), la construction d'arbres de décision, les réseaux bayésiens basés sur des probabilités conditionnelles, les réseaux de neurones …

L'algorithme des k-plus proches voisins a en particulier retenu toute notre attention puisqu'il figure au programme de la spécialité NSI en première. Il en a été de même pour les explications apportées sur le fonctionnement des réseaux de neurones qui sont régulièrement cités par les médias lorsqu'il s'agit d'évoquer les recherches en intelligence artificielle (notons d'ailleurs que les deux termes ne sont pas synonymes quand bien même il est tentant de faire le raccourci entre *"intelligence"* et *"neurones"*).

II : Clustering (Approche non supervisée)

Cette fois-ci il s'agit de faire de la classification d'objets (c'est-à-dire de données formalisées par des vecteurs multidimensionnels) sans connaître les classes a priori. Il faut donc disposer de méthodes algorithmiques permettant de créer des classes. Le principe général est le suivant :

Dans un premier temps on définit une mesure de similarité entre objets.
Dans un second temps il faut chercher à créer des classes en :

- maximisant la similarité *intra*-classes
- minimisant la similarité *inter*-classes

Les méthodes pour créer des classes satisfaisantes sont nombreuses. Deux d'entre elles nous ont été présentées plus en détail :

- **Par partitionnement (exemple des K-means):**
  Pour N objets, il s'agit de créer k partitions que l'on raffine itérativement jusqu'à obtenir une similarité satisfaisante. L'algorithme le plus utilisé est celui des K-moyennes :
    (1) : choisir aléatoirement k objets initiaux appelés graines
    (2) : assigner chacun des N objets à la graine la plus proche : cela donne k classes
    (3) : remplacer les k graines par les k "centres" des k classes
    (4) : réassigner chacun des N objets à la graine la plus proche : cela donne k classes
    (5) : Si des objets ont changé de classe à l'étape (4), retourner à l'étape (3)

- **Par densité (exemple de DBSCAN):**
  La densité se formalise grâce à la notion de voisinage (deux objets sont voisins s'ils sont à une distance inférieure à une valeur fixée). À partir de là, un objet est considéré comme dense si le nombre de ses voisins dépasse un certain seuil.

  L'algorithme pour trouver une nouvelle classe est alors le suivant :
    (1) : Choisir aléatoirement un objet dense
    (2) : Créer une nouvelle classe composée de cet objet et de tous ses voisins
    (3) : Rajouter dans la classe tous les voisins des objets denses qui viennent d'y être rajoutés
    (4) : Si des objets ont été rajoutés à l'étape (3), recommencer l'étape (3)

  Cette fois-ci une classe est donc obtenue par des élargissements successifs à partir de ses points denses.

On notera pour finir que les classes obtenues par ces méthodes sont parfois peu intelligibles au sens où elles ne correspondent pas forcément à des critères unidimensionnels.

III : Règles d'association

On ne cherche plus à déterminer des classes d'appartenance d'objets mais à déterminer des règles liant des données avec des probabilités "élevées". L'exemple qui nous a été présenté était celui du panier de la ménagère.

Il s'agit, à partir des compositions de tous les paniers qui sont déjà passés en caisse, de déterminer des implications telles que "Si un panier contient le produit X alors le panier contient le produit Y". Plus précisément il s'agit de trouver de telles implications ayant un "bon niveau de probabilité".