Supervisée

On considère une population comportant q groupes. On observe p variables $X_1, \ldots, X_p$ pour chaque individu/objet de la population. On cherche à obtenir un modèle/algortihme pour classer de nouveaux individus/objets dans les bons groupes, c'est-à-dire de prédire Y à partir de $X_1, \ldots, X_p$.

Example

- Revenue Québce désire identifier les déclarations fiscales méritant d'être examinées de façon plus approfondies (détection de fraude).
- Reconnaissance automatique des chiffres et des lettres des codes postaux écrits à la main.
- Identification de nouveaux clients potentiels.
- Filtrage de courriels indésirables.
- Reconnaissance d'images.

Approche générale

1. Sélectionner un certain nombre d'individus dont on connaît le groupe d'appartenance.
2. Mesurer p caractéristiques $X_1, \ldots, X_p$ sur ces individus.
3. Diviser ce jeu de données en deux :

- Un jeu de données pour la modélisation (entrainement, "train")
- Un jeu de données pour la vérification (valisation, "test")

4. Développer un modèle/algorithme pour classer le mieux possible les individus du jeu de données d'entrainement.
5. Évaluer notre modèle/algorithme sur le jeu de données de valisation.
6. (Répéter étapes 3-4-5 avec d'autres modèles/algorithmes et choisir le meilleur).

Quelques méthodes :

- Analyse discriminante
- Arbre de classification

— Régression
— Classificateur naïf de Bayes
— Méthode des k plus proches voisins
— Support vector machine
— Réseaux de neurones.

Il n'y a aucun algorithme qui garantit le meilleur classificateur pour toute situation donnée. Chaque problème est nouveau et on doit tenter de trouver la meilleure façon de procéder par essai et erreur. Ceci étant dit, certains principes s'appliquent plus généralement * Commencer par une exploration des données (p.ex. statistiques descriptives sur toutes les variables prises individuellement, ACP, ACB/ACM, classification non-supervisée) * Tirer avantage de la connaissance du sujet des experts qui nous entourent * Voir si certaines méthodes n'ont pas déjà eu du succès dans des analyses similaires

La principale difficulté vient habituellement de la dimension du problème : le nombre de modèles/méthodes possibles pour un problème donné est énorme et croît rapidement avec le nombre de variables disponibles. Parfois, réduire la dimension du problème (ACP, ACB/ACM, classification non-supervisée) peut aider : on applique ces techniques à un sous-ensemble des variables, et on utilise ensuite les scores produits comme prédicteurs dans les algos de classifications Il n'y a pas de recette générale pour savoir quel sous-ensemble choisir … Allez voir sur des sites de concours d'analyse de données (p.ex. Kaggle, KD Nuggets, etc.) et regardez les approches utilisées par les gagnants des concours pour lesquels le problème à résoudre s'apparente un peu au vôtre.