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.