Fiche de Révision : Réduction de Dimension et Visualisation

Introduction

- **Apprentissage supervisé** : Création de modèles prédictifs à partir de données étiquetées pour une bonne généralisation.
- **Apprentissage non supervisé** : Exploration et analyse des données pour extraire motifs et structures.
- **Objectifs de la réduction de dimension** :
  - **Visualisation** : Identifier des anomalies ou regroupements ($q = 2$ ou $q = 3$).
  - **Représentation** : Réduction de bruit, pré-traitement et détection de structures cachées ($q < d$).

Méthodes de Réduction de Dimension

Principe général

Projeter des points $x_i \in \mathbb{R}^d$ sur $z_i \in \mathbb{R}^q$ ($q < d$) tout en préservant la topologie des données (distances ou voisinages).

PCA (Analyse en Composantes Principales)

- **Objectif** : Maximiser la variance des projections et minimiser l'erreur de reconstruction.
- **Modèle** : $X = ZP^\top + B$, où P est une matrice de projection orthogonale.
- **Formule** : Trouver $P \in \mathbb{R}^{d \times q}$ qui minimise :

$$J(P) = \sum_{i=1}^{N} \|x_i - PP^\top x_i\|^2$$

- **Étapes** :
  1. Normaliser les données : $x_{ij} = (x_{ij} - \bar{x}_j)/\sigma_j$.
  2. Calculer la matrice de corrélation : $C = \frac{1}{N} X^\top X$.
  3. Trouver les valeurs propres λ_j et vecteurs propres p_j de C.
  4. Sélectionner les q plus grandes valeurs propres.

Méthodes non linéaires

- **SNE (Stochastic Neighbor Embedding)** : Minimise la divergence de Kullback-Leibler entre les distributions des espaces d'origine et projeté.
- **t-SNE** : Variante utilisant une fonction de probabilité alternative pour mieux gérer les longues distances.
- **Autres méthodes** : UMAP, auto-encodeurs, embeddings personnalisés (word2vec, etc.).

Applications

- **Visualisation** : Identifier anomalies et regroupements, exemple avec le dataset MNIST ($d = 784$, $q = 2$).
- **Pré-traitement** : Réduction de bruit pour simplifier les calculs des algorithmes d'apprentissage.
- **Représentation** : Détection de structures cachées dans les données.

Hugo Tondenier

Évaluation et Critères

- **Choix de q** :
  - Analyse du pourcentage de variance expliqué (graphique de l'"elbow").
  - Fixer un seuil de variance récupérée (par exemple 95 %).
- **Limites de la PCA** :
  - Projection uniquement linéaire.
  - Basée uniquement sur des statistiques d'ordre 2 (moyenne et variance).