

LU3I026 - Science des données

Représentations des données

Olivier Schwander <olivier.schwander@sorbonne-universite.fr>

Sorbonne Université

2024-2025

Modèles non-linéaires

Typiquement

- ▶ Passage linéaire/affine (ajout de la colonne de 1)
- ▶ Polynômes
- ▶ Gaussiennes

Réseaux de neurones (deep learning)

- ▶ Très grosses fonctions non-linéaires

Modèles polynomiaux

[Au tableau]

Astuce du noyau: kernelisation

[Au tableau]

Feature engineering

Travail sur les variables

- ▶ Moins de variables
- ▶ Meilleures variables
- ▶ Pour faire plus facilement la prédiction

Intérêts

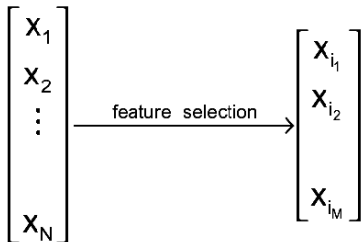
- ▶ Contourner les limites des modèles
- ▶ Guider l'apprentissage en cas de manque de données
- ▶ Limiter la dimension du problème

Importance des pré-traitements

[Au tableau]

Sélection de caractéristiques

Sélectionner un sous-ensemble des caractéristiques existantes



Comment faire ?

Exhaustivement

- ▶ Tous les sous-ensembles possibles ?
- ▶ Trop coûteux: exponentiel

Besoin de méthodes approchées

- ▶ Filtrage: a priori
- ▶ Wrapper: a posteriori

Méthodes de filtrage

Étapes

- ▶ Analyser chaque variable individuellement
- ▶ Estimer son pouvoir prédictif
- ▶ Trier les variables et ne garder que les meilleures

Exemple

- ▶ Corrélation entre chaque variable et la sortie à prédire

Limites

- ▶ Chaque variable est analysée toute seule
- ▶ Qualité de l'estimation du pouvoir prédictif ?
- ▶ Pas directement lié à la tâche

Méthodes de wrapper

Principe

- ▶ Choisir un sous-ensemble
- ▶ Évaluer les performances de ce sous-ensemble
- ▶ Ajouter ou supprimer des variables
- ▶ Recommencer

Limites

- ▶ Stratégies d'ajout ou de suppression
- ▶ Coûteux

Extraction de caractéristiques

Construire de nouvelles variables

- Complexe

Analyse en composantes principales: motivation

ACP ou PCA pour Principal Components Analysis

[Au tableau]

Formulation

Minimiser l'erreur de reconstruction

$$\min_P \sum_i \|x_i - P(x_i)\|^2$$

Projection P

- ▶ Espace de départ \mathbb{R}^d
- ▶ Espace d'arrivée: sous-espace vectoriel E_p de rang p
- ▶ avec $p < d$
- ▶ P projection orthogonale des points sur E_p

Centrer et réduire

Centrer: indispensable

[Au tableau]

Réduire: facultatif

- ▶ En fonction des cas
- ▶ En particulier: échelles très différentes ou unités différentes

Changement de base

Projection linéaire sur un vecteur

Pour un vecteur unitaire m (de norme 1)

$$\hat{x}_i = (m_i \cdot x_i)m$$

Sur un sous-espace

Pour une base orthogonale m_1, \dots, m_p tq $m_i \cdot m_j = 0$ si $i \neq j$

$$\hat{x}_i = \sum_{j=1}^p (m_j \cdot x_i)m_j$$

Sous forme matricielle

Avec $M = (m_1, \dots, m_p)$

$$\hat{x}_i = MM^t x_i$$

Représentation dans le sous-espace

Dans la base M

$$\tilde{x}_i = M^t x_i$$

Algorithme

Avec la covariance empirique

- ▶ Estimer $\Sigma = \frac{1}{n}X^tX$
- ▶ Décomposer en valeur propres $\Sigma = Q\Lambda Q^t$ (Λ valeurs propres et Q vecteurs propres)
- ▶ Trier les valeurs propres $\lambda_1, \dots, \lambda_n$ et les vecteurs propres associés
- ▶ Prendre les p premières colonnes de Q : $\tilde{Q} = (Q_1, \dots, Q_p)$ (avec Q_m une colonne de Q)
- ▶ \tilde{Q} est la projection qu'on cherche

$$\tilde{x}_i = \tilde{Q}^t x_i$$

En pratique

Décomposition en valeurs singulière

Singular Value Decomposition ou SVD

- ▶ $X = USV^t$
- ▶ avec $U \in \mathbb{R}^{n \times n}$
- ▶ et $V \in \mathbb{R}^{d \times d}$
- ▶ et $S \in \mathbb{R}^{n \times d}$ diagonale (valeurs singulières)

Pour l'ACP

- ▶ On veut éviter de calculer X^tX
- ▶ SVD sur la matrice des points
- ▶ $X^tX = V(S^tS)V^t$

Ou sinon

- ▶ Descente de gradient

Variance expliquée

[Au tableau]

Conclusion

Analyse des données indispensable

- ▶ Manuellement
- ▶ Avec un expert
- ▶ Automatiquement

Dualité

- ▶ Modèle plus souple
- ▶ Données transformées

Analyse en composante principale

- ▶ Outil basique mais classique
- ▶ Plein de variantes
- ▶ Utilisable aussi pour la visualisation