REsearch and methodology in Data Science

Cours 3 – Sélection de variables

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

Master DAC
Sorbonne Université

2023-2024

Séléction de caractéristique

Sélection de caractéristiques sélectionner un sous-ensemble des caractéristiques existantes:

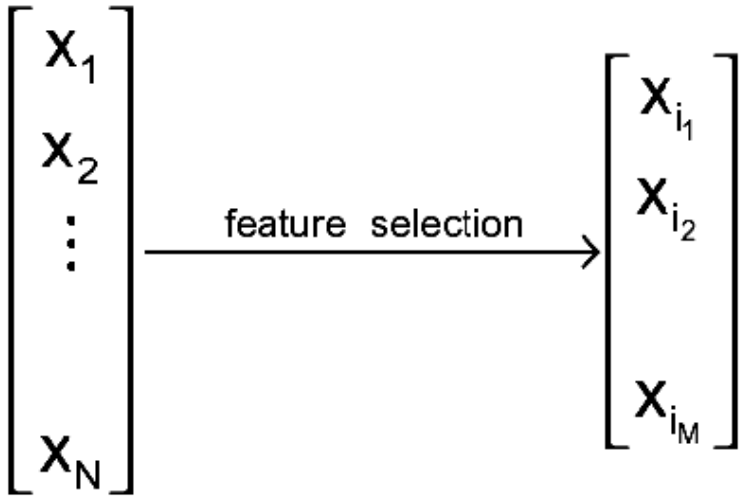
- Approches de type **Filtering**
- Approches de type **Wrappers**

Extraction de caractéristiques combiner des caractéristiques existantes pour obtenir un (petit nombre) de caractéristiques pertinentes:

- Approches de type **PCA**
- Approches de type **Auto-Encodage**
- Approches de type **Representation Learning (Deep Learning)**

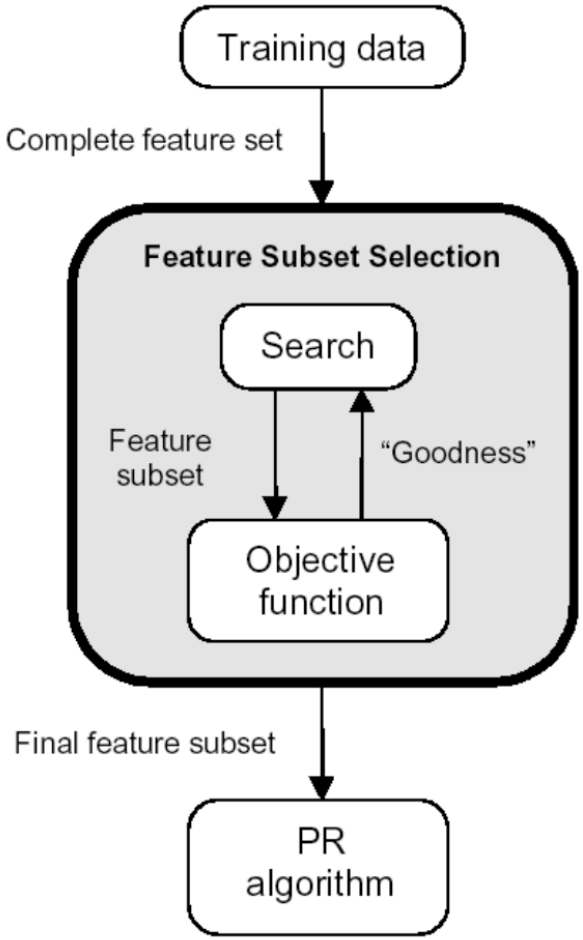
Séléction de caratéristiques

- Soit un ensemble d'entrée $\mathcal{X} = \mathbb{R}^n$ tel que $x = (x_1, x_2, ..., x_n)$
- On cherche à trouver un sous-ensemble de dimensions caractérisé par un ensemble $\mathcal{I}$ d'index dans $[1; n]$
- Etant donné $\mathcal{I} = (i_1, ...i_M)$, le nouvel espace d'entrée sera caractérisé par $x = (x_{i_1}, x_{i_2}, ..., x_{i_M})$



Sélection de caratéristiques

- Très grand espace de recherche
- Besoin de méthodes approchées



Deux approches

Méthodes de filtrage: sélection *a priori*

- Estimation du pouvoir prédictif de chaque caractéristique
- Étude mono-dimensionnelle de chaque caractéristique
- Sélection de celles avec le pouvoir prédictif le plus élevé

Méthodes de wrappers: sélection *a posteriori*

- Choix basé sur la qualité du modèle obtenu

Corrélation

Mesure de l'intensité de la liaison entre deux variables

Corrélation linéaire Soit la variable X_i (caractéristique) et la variable Y (étiquette):

$$Corr(X_i, Y) = \frac{Cov(X_i, Y)}{\sqrt{Var(X_i)Var(Y)}}$$

- $Cov(X_i, Y) = E[X_i Y] - E[X_i]E[Y] = E\left[(X_i - E[X_i])(Y - E[Y])\right]$
- $Cov(X_i, Y) = 0$ ssi X_i et Y sont indépendantes

Corrélation empirique

Comme d'habitude lois inconnues pour X_i et Y

Estimateur

$$R(i) = \frac{\sum_{k=1}^{N} (x_i^k - \bar{x}_i)(y^k - \bar{y})}{\sqrt{\sum_{k=1}^{N} (x_i^k - \bar{x}_i)^2 \sum_{k=1}^{N} (y^k - \bar{y})^2}}$$

- Dépendance linéaire
- Versions non-linéaires
- **Corrélation n'est pas causalité**

Méthodes de filtrage

- Tri des variable par ordre de pertinence
- Conservation des caractéristiques les plus pertinentes

Avantage

- Chaque caractéristique est analysée **indépendamment des autres**.
- Rapide

Limite

- Chaque caractéristique est analysée **indépendamment des autres**.
- Une variable pourrait être utile en combinaison avec une autre

Méthodes de wrappers

- Choisir un sous-ensemble de caractéristiques
- Entraîner un modèle et l'évaluer
- Choisir le sous-ensemble qui donne les meilleurs performances

Coûteux:

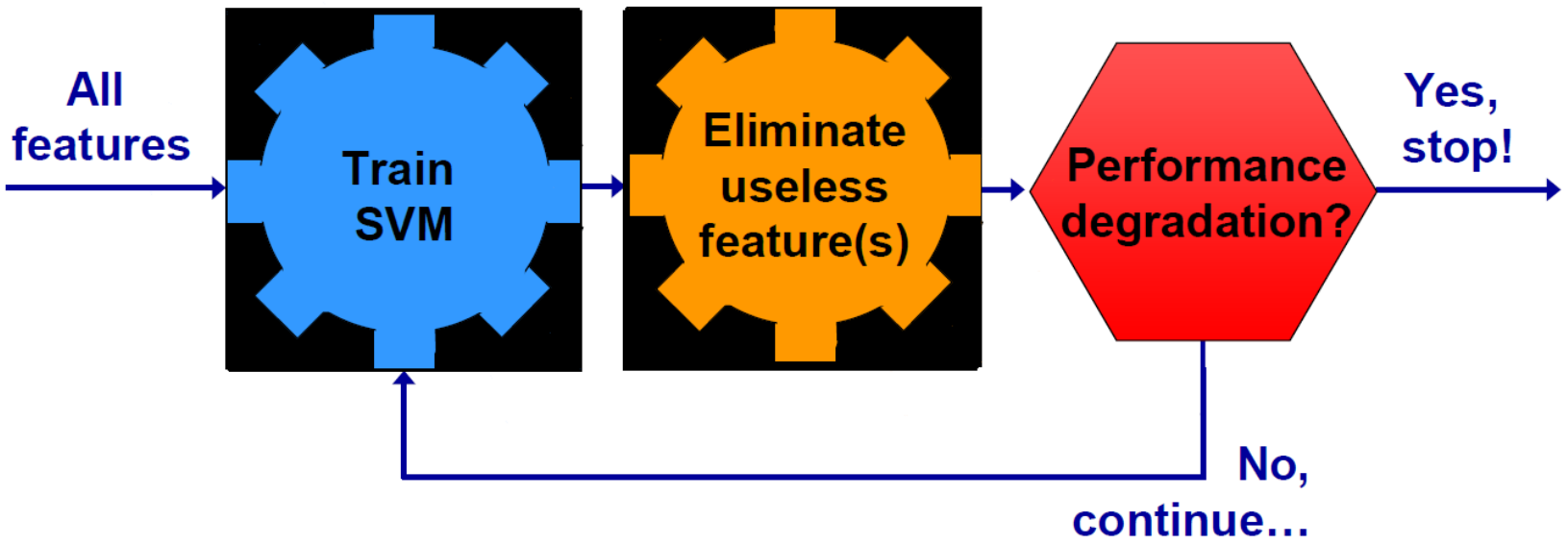
- Nombre exponentiel de sous-ensembles
- Entraînement des modèles

Recherche gloutonne: ajout graduel de caractéristiques basé sur un score à chaque pas de l'algorithme

Attention: le score doit refléter la performance du système (en généralisation)

Méthodes embarquées

De moins en moins de caractéristiques



Recursive Feature Elimination (RFE) SVM. *Guyon-Weston, 2000. US patent 7,117,188*

Conclusion

Grandes lignes d'un projet ML

- Définir la tâche et y réfléchir
- Analyser les données (statistiques descriptives, feature engineering)
- Établir un protocole expérimental
- Choisir un modèle et évaluer ses performances
- Présenter le travail