DATA MINING

Stratégie & techniques

@Mohamed ZWAWA

SUP DE VINCI

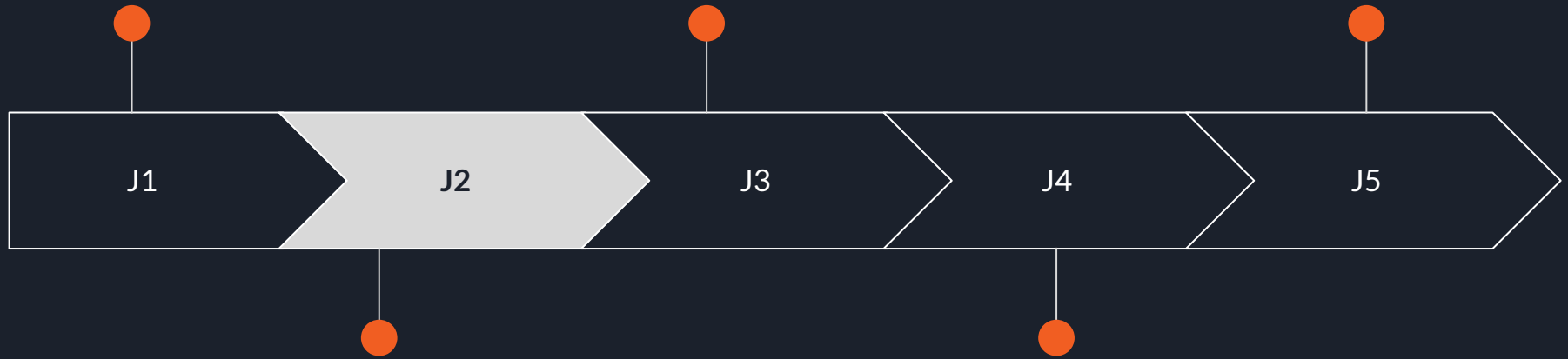
Jan 2023

Exposés

Organisation du cours
Introduction générale
TP (Prise main des outils de travail)
Résumé

La modélisation décisionnelle
Les méthodes factorielles
TP (ACP & AFD)
Résumé

Révision générale
Evaluation
Bonus



La stratégie du data mining
Les techniques du data mining
La réduction des variables
TP (Techniques de transformation & Feature selection)
Résumé

La régression linéaire
Autres modèles: RL, KNN, SVM
TP (Mini projet/Exposé)
Résumé

Contenu

- La stratégie du data mining
- Les techniques du data mining
- Travaux pratiques
  - TP4: Les techniques de transformation
  - TP5: La sélection des variables
- Résumé

La stratégie du data mining

- Le data mining se base sur l'exploration des données puis l'apprentissage.
- Exploration des données
- Apprentissage à partir des données explorées afin de construire un modèle prédictif qui répond à un besoin métier donné. (Apprentissage supervisé)

- Les étapes de l'exercice du data mining:
  - Extraction des données (⚠️ Data understanding)
  - Exploration des données: Typologie des données, recherche de corrélation, détection des valeurs aberrantes, manquantes, erronées (⚠️ Business understanding)
  - Partitionnement des données (Entrainement, Test)
  - Atelier de création d'un estimateur (tester plusieurs modèles + optimisation)
  - Comparaison des modèles et choix du meilleur

Les techniques du data mining

- Classification: analyse qui permet de classer les données dans des classes bien définies.

- Régression: analyse qui vise à affecter une valeur à une variable spécifique.

- Clustering: analyse qui permet d'identifier des similarités et des différences entre les données.

- Outlier: analyse des valeurs aberrantes (détection de fraude)

et pas que…

Problématique de réduction des variables

Objectifs de la réduction du nombre des variables d'entrée:

- Réduire le volume de données à traiter tout en conservant le maximum d'information utile
- Supprimer les variables non pertinentes (bruit)
- Réduire la complexité (coût) d'un modèle décisionnel
- Améliorer la lisibilité des données (mieux comprendre comment les décisions sont prises par le modèle)
- Diminuer le problème de malédiction de la dimension.

Tout dépend du besoin métier:

Objectifs différents + critères différents ⇒ Méthodes différentes

La réduction des variables → Deux grandes familles

La sélection des variables

- Sélection d'un sous ensemble de k variables parmi les n variables d'entrée.
- Avantages:
  - Les k variables sélectionnées gardent leurs significations.
- Inconvénients:
  - Complexité de calcul
  - Sous optimale

La réduction des dimensions

- Réduction de n dim à k dim
- k dim construites à partir des variables initiales via une combinaison (linéaire ou non linéaire).
- Exemples: ACP ou AFD
- Avantages:
  - Flexibilité
- Inconvénients:
  - Les nouvelles variables ne sont pas interprétables

Quizz

https://forms.gle/Awq47EHkrPycwBYi6

TP n°4
Les techniques de transformation

TP n°5
Feature selection

Résumé

Qu’est-ce qu’on a appris aujourd’hui ?

https://forms.gle/TDUTg2KaycNkzFVJ6

Références

- Data mining I Exploration Statistique
- Data Mining et Statistique
- 5 outils data mining pour mieux analyser vos données - Codeur Blog
- Data Mining définition : Qu'est-ce que l'exploration des données ?
- Difference in Data Mining Vs Machine Learning Vs Artificial Intelligence

Merci 🙂

formateur_mohamed.zwawa@supdevinci-edu.fr