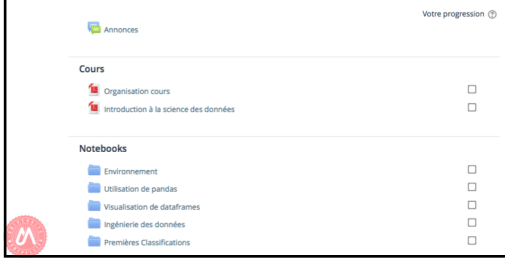
Méthodes pour la science des données

HMIN232M

Pascal Poncelet
LIRMM
Pascal.Poncelet@lirmm.fr
http://www.lirmm.fr/~poncelet

Les informations sont disponibles sous Moodle

- Nom : Méthodes pour la science des données
- Code : HMIN232M



2

Les informations sont disponibles sous Moodle

- Contenu des cours, des TP
- Devoir à rendre via Moodle
- Informations diverses via Moodle

- Des ressources disponibles : les notebooks

3

Les intervenants

- Dino Ienco, CR Irstea
- Christophe Menichetti, Ingénieur IBM
- Pascal Poncelet, Prof UM
- Konstantin Todorov, MCF UM

4

Le planning

- Les cours ont lieu le jeudi
- Attention ADE n'est pas forcément à jour
- EDT : Voir Moodle
- En cas de modification, nous vous prévenons par un message via Moodle

5

Le programme

- Le processus d'extraction de connaissances
- Comprendre les données
  - *ingénierie des données, statistiques descriptives, visualisation, choix des dimensions importantes, normalisation*
- Trouver les meilleurs modèles et mesures
  - *classification supervisée, classification non supervisée, pattern mining*
- Evaluation des modèles
  - *precision, rappel, F-measure, AUC,*
- Mise en place d'un pipeline complet pour utiliser les modèles appris sur de nouvelles données

6

Un projet

- Thématique : en cours de définition - données textuelles
- L'année dernière : détection d'opinions
- Les données sont fournies pour faciliter les traitements et se focaliser sur la partie fouille
- 2 fichiers Excel : les données et la classe des données
  - I saw this movie at a drive-in in 1959. Until "Howard the Duck" I considered this the worst movie I had ever seen. This movie tried to combine all the genera in one.
  - Negatif
- Phrase positive ou négative ?

7

Un projet

- Travail en groupe
- Trouver la meilleure classification
- Un challenge
- Evaluation :
  - Qualité du rapport,
  - Les prétraitements, choix des représentations, choix des descripteurs, choix des classifieurs, …
  - Explication des résultats du challenge

8

Evaluation

- 2 notes :
  - 1 note individuelle : article à lire et réponse à des questions (Examen final : 30%)
  - 1 note commune (Projet avec soutenance : 70%)
    - Attention en fonction de la présentation, les notes peuvent être différentes
- Pour le projet :
  - Par groupe à saisir sur Moodle
  - Remise d'un notebook, des données informations supplémentaires, etc.

9

Contacter les enseignants

- Les adresses mails :
  - Dino.Ienco@irstea.fr
  - Christophe.Menichetti@fr.ibm.com
  - Pascal.Poncelet@lirmm.fr
  - Konstantin.Todorov@lirmm.fr

- S'il vous plait mettre dans le sujet :
- [HLIN232M] ….

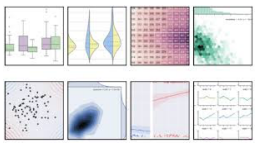
10

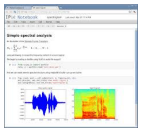
Environnement de travail

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

- Les outils
  - Sickit learn : bibliothèque libre Python destinée à l'apprentissage automatique
  - Pandas, SciPy : pour manipuler les données
  - Seaborn pour visualiser les statistiques sur les données
  - Jupyter Notebook : pour faire de la science des données





11

Une démo de notebook

12

- A faire rapidement :
  - S'inscrire sur Moodle
  - Installer jupyter (cf fichier environnement.pdf et environnement.ipynb)
- Des questions ?

13