ML 2 : Classification & Clustering

KIT APPRENANT

le campus numérique in the ALPS

Ce projet est cofinancé par l'Union européenne
L'EUROPE S'ENGAGE en région Auvergne-Rhône-Alpes avec le FSE
La Région Auvergne-Rhône-Alpes

OBJECTIFS

Objectifs pédagogiques

A l'issue de ce module, vous serez capable de :

- Entrainer et évaluer un modèle de classification
- Utiliser des méthodes d'ensemble
- Mettre en évidence les phénomènes de sur/sous apprentissage
- Entrainer et évaluer un modèle de clustering

Compétences développées :

Vous apprendrez à choisir les bonnes métriques pour un problème de classification, entrainer des modèles de classification dont des modèles de bagging et de boosting et réaliser des prédictions, ainsi qu'à évaluer l'importance des features utilisées. Et enfin à réaliser un clustering et évaluer sa qualité.

Démarche pédagogique (projet, ressources ...)

- Durée du projet : 3 jours
- Travail en autonomie, mais échangez autant que possible entre vous !
- Produire vos propres scripts et mémos individuels pour terminer le projet

Compétences

Itération 1

- Entrainer un modèle de classification et faire des prédictions
- Choisir les bonnes métriques d'évaluation pour un problème de classification

Itération 2

- Entrainer un modèle de classification en utilisant les techniques de bagging et de boosting
- Trier les paramètres d'un problème par ordre d'importance

Itération 3

- Évaluer les performances d'un modèle de clustering

MODALITÉS

Durée

4 jours soit 28 heures au total.
Lancement le 7/11/23 et clotûre le 10/11/23.

Formateur(s)

Théo Trouillon, Cyril François

ITÉRATION 1

Classification

Modalités

- Travail individuel en autonomie
- 1/1.5 jours en présentiel

Livrables

- ❏ Répondre aux questions du fichier mémo
- ❏ Le notebook rempli, permettant d'évaluer les performances d'un classifieur par k plus proches voisins

Objectifs

- Se familiariser avec la bibliothèque scikit-learn
- Savoir entraîner un modèle de classification et faire des prédictions
- Connaître les différentes métriques d'évaluation pour les problèmes de classification
- Mettre en place une procédure de sélection de modèle par grid-search et cross-validation

Compétences

- Entrainer un modèle de classification et faire des prédictions
- Choisir les bonnes métriques d'évaluation pour un problème de classification

Ressources

- https://scikit-learn.org/stable/tutorial/basic/tutorial.html
- https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html
- https://scikit-learn.org/stable/modules/cross_validation.html
- *"Hands on machine learning ..."*, chapitres 2 et 3

- *"Introduction to statistical learning"*, chapitre

ITÉRATION 2

Ensemble methods

Modalités

- Travail individuel en autonomie
- 1/1.5 jours en présentiel

Livrables

- ❏ Visualisation du classement des paramètres sous forme d'histogramme
- ❏ Utilisation des méthodes d'ensemble
  - ❏ Notebook complété.
  - ❏ Mémo/Schéma sur les méthodes d'ensemble comprenant:
    - ❏ Schéma de fonctionnement des méthodes de bagging et de boosting.
    - ❏ Avantage/Inconvénients de chacunes des méthodes.

Objectifs

- Entraîner un modèle de classification en utilisant les techniques de bagging et de boosting.
- Trier les paramètres d'un problèmes par ordre d'importance.
- Évaluer les performances d'un modèle de classification.

Compétences

- Entrainer un modèle de classification en utilisant les techniques de bagging et de boosting
- Trier les paramètres d'un problème par ordre d'importance

Ressources

- https://scikit-learn.org/stable/modules/ensemble.html
- https://martin-thoma.com/ensembles/
- https://medium.com/@rrfd/boosting-bagging-and-stacking-ensemble-methods-with-sklearn-and-mlens-a455c0c982de
- https://xgboost.readthedocs.io/en/latest/index.html
- https://www.lpsm.paris/pageperso/has/source/Hand-on-ML.pdf *(chapitre 7)*

ITÉRATION 3

Clustering

Modalités

- Travail individuel en autonomie
- 1/1.5 jours en présentiel

Livrables

❏ Analyses de silhouette commentée de résultats de clusterings qui doit comprendre:
  ❏ Une silhouette avec des clusters non uniformes.
  ❏ Une silhouette avec des clusters uniformes.

Objectifs

- Utiliser des méthodes de partitionnement
- Trouver le nombre de cluster optimal
- Créer des partitions à partir d'un jeu de données en utilisant des méthodes mise à disposition dans scikit-learn
- Visualiser les partitions créées par un algorithme de partitionnement

Compétences

- Évaluer les performances d'un modèle de clustering

Ressources

- https://scikit-learn.org/stable/modules/clustering.html#clustering
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- https://scikit-image.org/