

Machine Learning Avancée

MLA ITI 5

Présentation générale

1 septembre 2025

Stéphane Canu

Objectifs du cours Moodle

moodle.insa-rouen.fr/

Ouvrir l'index du cours **eLearning Avancée (ITI5)**

Cours Paramètres Participants Notes Rapports Plus ▾



Généralités

Tout replier



Forum des nouvelles

Le cours aborde les méthodes avancées d'apprentissage statistique et les problèmes d'optimisation qui se posent dans le domaine de l'apprentissage statistique, ainsi que des algorithmes efficaces pour les résoudre. Plus exactement, nous considérerons le processus d'entraînement automatique comme un problème d'optimisation. En particulier nous aborderons les questions liées à l'apprentissage d'optimisation convexes, différentiables, non convexes et non différentiables.

Nous aborderons notamment :

- les questions récentes liées aux progrès récents de l'apprentissage statistique
- les moindres carrés et les pénalités usuelles associées (LASSO, MCP...)
- les méthodes de sélection de variables (Lasso, best subset, PCR et PLS...)
- les problèmes de réglage d'hyper paramètres
- AutoML
- l'Apprentissage de représentation
- les méthodes de clustering
- les méthodes de Factorisation

Scikit learn sera utilisé pour la mise en œuvre de ces méthodes.

Les TD auront lieu sous python à l'aide du notebook jupyter

Vous pourrez aussi utiliser google colab

<https://colab.research.google.com>

Evaluation

1. Pour tous : un examen écrit de 3h

2. Pour les ITI, il faut aussi effectuer une contribution à un logiciel open source (typiquement scikit learn) :

- <https://scikit-learn.org/stable/developers/contributing.html>
- <https://github.com/scikit-learn/scikit-learn/issues>

2.1 Définissez le point sur lequel vous voulez travailler (pour le 2 novembre)

2.2 Rendez votre travail (avant le dimanche 11 janvier)

Contribution

scikit-learn / scikit-learn

Public

Sponsor

Notifications

Fork 26.2k

☆

<> Code 1.6k Issues Pull requests 590 Discussions Actions Projects 15 Wiki Security Insights

Cython and memviews creation
#17299 · NicolasHug opened on May 21, 2020

Local testing of global_random_seed is not enough
#28959 · jeremiedbb opened on May 6, 2024

is:issue state:open

Open 1580 Closed 10087

Author Labels Projects Milestones Assignees Types

The dcg_score and ndcg_score documentation are hard to understand Documentation Needs Triage
#32049 · BrianMacCarvill opened 2 days ago

Leiden Clustering Needs Triage New Feature
#32048 · martinalex000 opened 2 days ago

rendering of 'routing' note in the documentation Documentation Needs Triage
#32046 · Remi-Gau opened 2 days ago

make sphinx directive about version more sklearn specific Documentation Needs Triage
#32045 · Remi-Gau opened 2 days ago

Classification metrics don't seem to support sparse? Bug Documentation
#32036 · lucyleeow opened 3 days ago

Setting weights on items when passing list of dicts to RandomizedSearchCV Needs Decision - Include Feature New Feature
#32032 · jeromedockes opened 4 days ago

Implementing Divisive Analysis Hard module:cluster Needs Decision - Include Feature New Feature
#31989 · G4ll4rd0 opened last week

Different Results on ARM and x86 when using RFECV(RandomForestClassifier()) Bug Needs Investigation
#31988 · Oddant1 opened last week

ValueError in PLSRegression.fit() with zero-variance predictor Bug
#31971 · SauersML opened 2 weeks ago

a11y - scikit-learn docs accessibility audit and remediation Documentation
#31965 · trallard opened 2 weeks ago

UserWarning: X has feature names. but PowerTransformer was fitted without feature names Bug

Rapport de contribution sklearn: Issue #24862: Make automatic validation for all scikit-learn public functions

Exemple de rapport

Abstract

Ma contribution à la librairie open-source Scikit-learn traite de la validation automatique des fonctions publiques de Scikit-learn.

Pour cela j'aurai contribué à la validation de deux fonctions à savoir celles de [cluster.cluster_optics_xi](#) et de [covariance.graphical_lasso](#).

Ces tâches de validation correspondent à l'issue suivante : "[Make automatic validation for all scikit-learn public functions](#)". Au travers de ce rapport je décrirais les tâches que j'ai effectué et la problématique initiale.

1. Introduction

1.1. Problématique

Le but de la suivante "[Make automatic validation for all scikit-learn public functions](#)" est de faire une validation automatique de l'ensemble des fonctions publiques de scikit-learn en rajoutant avant chaque fonction publique le décorateur `sklearn.utils._param_validation.validate_params`. Ce décorateur prend en entrée un dictionnaire Python où chaque clé correspond au nom des différents paramètres de la fonction à valider et les valeurs correspondantes aux contraintes associées.

Voici un exemple pour `kmeans_plusplus` ci-dessous :

Le but donc de ce décorateur est de valider les paramètres de la fonction `shrunk_covariance` sans devoir le faire avec des conditions IF. Ceci permet d'utiliser correctement les fonctions et de les simplifier au maximum pour optimiser les calculs.

1.2. Fonctions choisies

La première fonction choisie est la fonction de `sklearn.covariance.graphical_lasso` elle permet d'estimer la covariance inverse éparsée avec un estimateur l1-pénalisé. Cette fonction contient 11 paramètres qu'il a fallu les rajouter de manière individuelle en respectant bien leurs types de données ainsi que les valeurs que prenaient

```
@validate_params(
    {
        "emp_cov": ["array-like"],
        "shrinkage": [Interval(Real, 0, 1, closed="both")],
    }
)
def shrunk_covariance(emp_cov, shrinkage=0.1):
    """Calculate a covariance matrix shrunk on the diagonal.

    Read more in the :ref:`User Guide <shrunk_covariance>`.

    Parameters
    -----
    emp_cov : array-like of shape (n_features, n_features)
        Covariance matrix to be shrunk.

    shrinkage : float, default=0.1
        Coefficient in the convex combination used for the computation
        of the shrunk estimate. Range is [0, 1].
```

Figure 1: Décorateur pour la fonction `shrunk_covariance`

La deuxième fonction choisie est la fonction `cluster.cluster_optics_xi`, cette fonction sert à effectuer un clustering en utilisant Optics.

2. Contribution

Le travail que j'ai effectué était donc de rajouter pour la fonction 23 paramètres au sein du décorateur

Exemple de rapport

1. Intitulé du problème	3
2. Description du problème	3
3. Correction proposée par l'auteur	3
4. Etude du problème	3
5. Etude du fonctionnement de sklearn	5
6. Premier pull request : ajout dans la documentation	6
7. Deuxième pull request : ajout d'un exemple dans la galerie	7
Références	10

Les questions récentes liées aux progrès récents de l'apprentissage statistique

- Actualité de l'IA
 - GPT 5/ Grok 4 -> leaderboard
 - Hallucination -> grok
 - Meta's super intelligence lab
 - Open gpt-oss-120b et gpt-oss-20b (20 Go) / grok 2.5
 - local
 - Spécialisé : Introducing OwkinZero, a biological reasoning model designed to accelerate drug discovery.
 - la polémique Julia
 - Worldwide Kids Vibe Coding Hackathon (<https://kidsaicoding.com/kids-hackathon-worldwide>)

L'entraînement des LLM

