
Examen Apprentissage automatique - FISA

— 1h30
— Tous documents autorisés
— Documentations en ligne : `https://numpy.org` et `https://scikit-learn.org`
— Vos réponses et analyses doivent être rédigées et éventuellement accompagnées de figures dans un fichier à part.
— Le rendu se fait sur Ametice : un zip avec le code et le pdf.

Les données de travail ont été récoltées dans une université de renommée internationale pour étudier les habitudes de travail des étudiants :

— Temps de révision hebdomadaire en heure
— Nombre d'absences non justifiées pendant le semestre
— Ratio temps de lecture / temps de sommeil quotidien
— Vie avec un chat ou non
— Note finale obtenue à l'examen
— Examen validé ou non

Les données sont disponibles dans le fichier .csv fourni. Pour les charger en Python, vous pourrez utiliser le code suivant :

```
import numpy as np
data = np.loadtxt("data.csv")
X = data[:,:4]
Ynote = data[:,4]
Yval = data[:,5]
```

Partie 1 (Régression)

L'objectif de cet exercice est d'analyser la dépendance entre les variables X et Ynote avec l'algorithme de régression `Ridge`.

Questions

1. Ecrivez la fonction de coût qui est optimisée sur l'ensemble d'apprentissage.
2. Quel est le rôle de l'hyperparamètre `alpha` (notation d'après la doc sklearn) ?
3. Écrivez le code pour :
   — séparer les données en des données d'apprentissage (70%) et de test (30%),
   — normaliser les variables explicatives (utilisez la classe `MinMaxScaler`)

— sélectionner l'hyperparamètre `alpha` sur l'échantillon d'apprentissage par validation croisée (**ne pas** utiliser la classe `RidgeCV`),
— afficher une estimation des performances empirique et en généralisation de la régression Ridge sur ce jeu de données,

4. Commentez les résultats obtenus. Quelles sont les variables les plus prédictives de la note finale ?
5. Y a t-il un intérêt à utiliser la version régularisée de la régression linéaire ?

Partie 2 (Classification)

On s'intéresse maintenant à une tâche de classification binaire pour prédire la variable Yvalid $\{-1; +1\}$ à partir des données de X.

Justifiez vos réponses et écrivez le code nécessaire (avec ou sans sklearn).

Questions

1. Ce problème est-il linéairement séparable ?
2. Quelle est la performance de l'algorithme des K-plus-proches-voisins sur ce problème de classification ?