Data Challenge IASO Dauphine

Encadrants : Tony Bonnaire, Kimia Nadjahi

Conçu avec : Nicolas Schreuder, Alexandre Allauzen

En collaboration avec l'équipe de ChallengeData de l'ENS

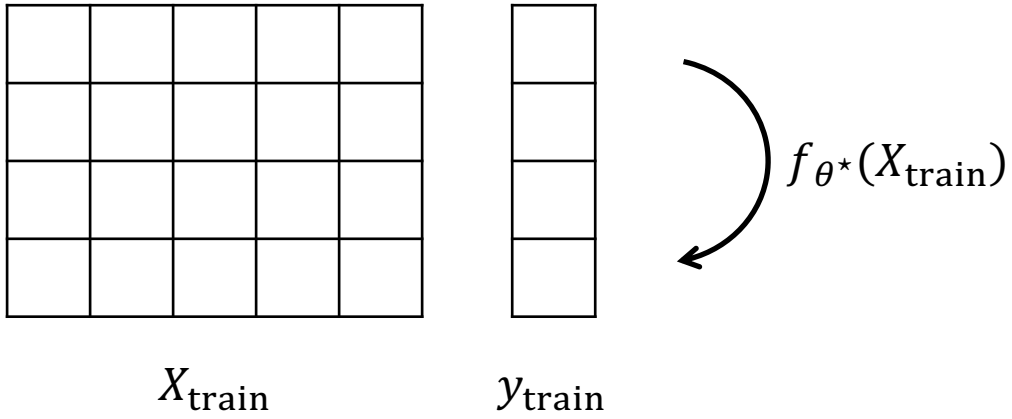
Paris Santé Campus, 3-7 juin 2024

Généré par Microsoft AI avec le prompt « A data challenge of students from economics, finance and artificial intelligence »

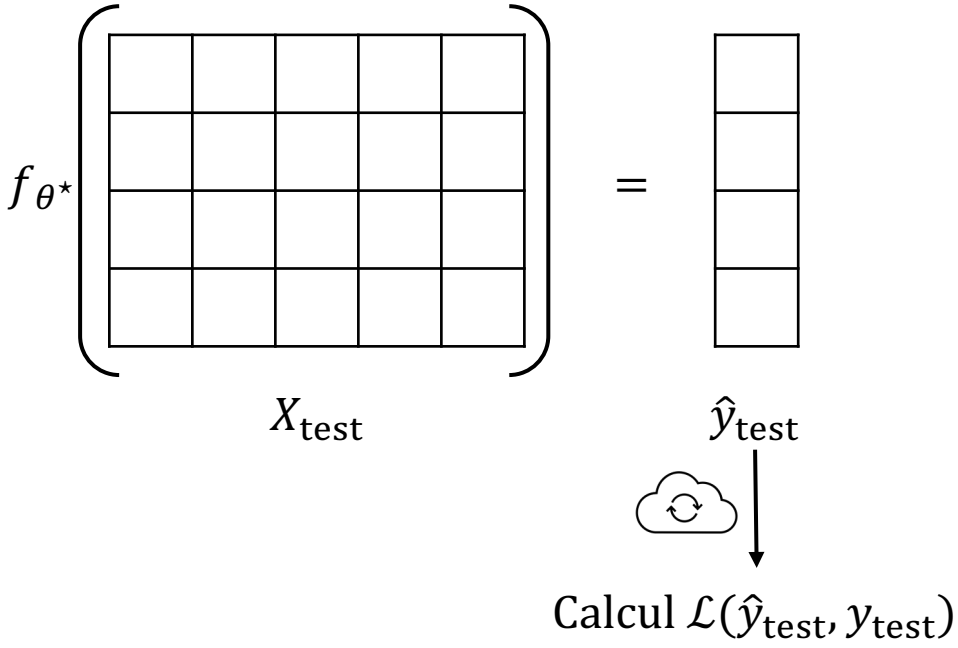
Qu'est-ce qu'un data challenge ?

Répondre à une problématique industrielle ou scientifique **à partir de données** en battant l'**algorithme de benchmark** selon une **métrique** choisie par l'organisateur $\mathcal{L}$

1 **Ensemble d'entrainement** pour créer et évaluer des modèles f_θ



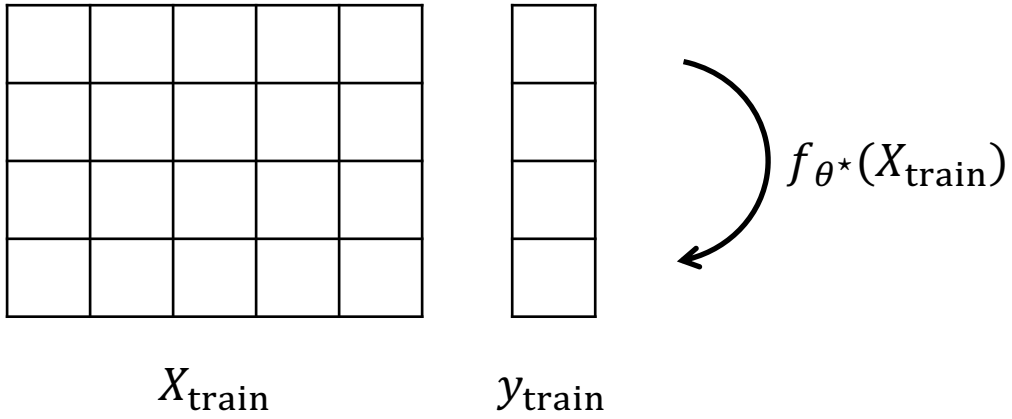
2 **Ensemble de test** (labels inconnus) pour comparer les participants



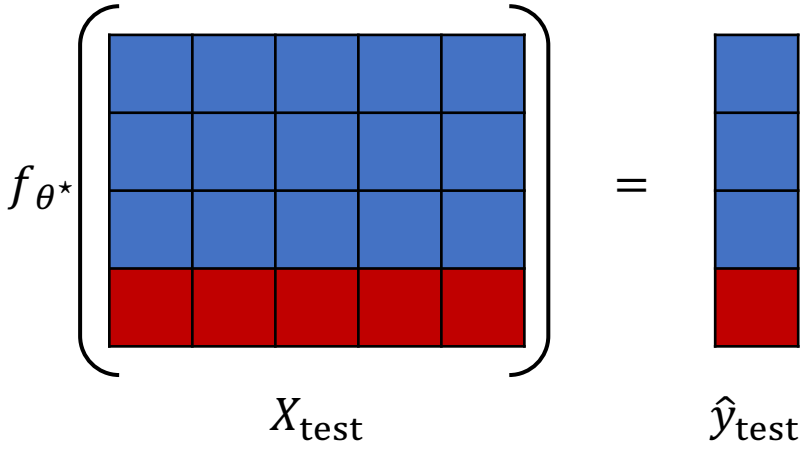
Qu'est-ce qu'un data challenge ?

Répondre à une problématique industrielle ou scientifique **à partir de données** en battant l'**algorithme de benchmark** selon une **métrique** choisie par l'organisateur $\mathcal{L}$

1 **Ensemble d'entrainement** pour créer et évaluer des modèles f_θ



2 **Ensemble de test** (labels inconnus) pour comparer les participants



Séparation de l'ensemble de test en un ensemble public et un ensemble privé (sur lequel vous aurez les résultats une fois par jour) pour **éviter le sur-apprentissage**

Prédiction du prix de l'immobilier

https://challengedata.ens.fr/challenges/68

INSTITUT Louis Bachelier

Prédire le prix d'un bien immobilier à partir de ses caractéristiques visibles (surface, position, exposition, nombre de chambres, de salles de bain, etc.)

50,000 offres immobilières (40,000 pour l'entrainement et 10,000 pour le test)

Données tabulaires qui incluent 27 variables réparties dans trois principales catégories :

- Description du bien : taille, type, nombre de chambres, etc.,
- Localisation : latitude/longitude (bruitées), ville, code postale, performance énergétique, etc,

Données d'images (optionnel !) : entre 1 et 6 images pour chaque bien

Métrique Mean Absolute Percentage Error (MAPE) $\mathcal{L}(\boldsymbol{y}) = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$

Benchmark Modèle de régression XGBoost sur les caractéristiques tabulaires et sur un embedding simple des images ($\mathcal{L}_{\text{test}} = 36.78$)

Qui sont les traders haute-fréquence

https://challengedata.ens.fr/challenges/50

AUTORITÉ DES MARCHÉS FINANCIERS AMF

Classer les participants aux transactions de marchés financiers et identifier les traders haute-fréquence (HFT), qui utilisent des algorithmes et méthodes automatiques

Comportement de 80 traders sur plusieurs jours (40 en entrainement, 40 en test)

Données tabulaires qui incluent 35 variables décrivant la statistique des temps de transactions entre deux événements sur plusieurs actions du marché (moyenne, min, max, médiane, quartiles, etc.)

3 classes possibles : HFT, non-HFT et MIX

Métrique F1-score micro-moyenné

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Benchmark Forêt aléatoire pour prédire un score HFT. Si plus de 85% des événements sont classés HFT alors le trader est HFT. Si c'est 50% alors il est MIX. Autrement, il est non-HFT.

($\mathcal{L}_{\text{test}} = 0.9048$)

Objectif et étapes à suivre

Proposer et présenter vos solutions de ML à l'un des problèmes précédents

Ne vous lancez pas dans un algorithme compliqué dès le départ : analysez les données, construisez des modèles simples et comprenez pourquoi ils fonctionnent/ne fonctionnent pas, puis améliorez-les !

ETAPES

1. **Créer un compte** sur le site https://challengedata.ens.fr
2. **S'inscrire au cours** « Data Challenge IASO Dauphine – Juin 2024 » via ce lien
3. Constituer des **groupes de 3** pour travailler
4. Lire la page du challenge et télécharger les données
5. Chercher des solutions !

A PROPOS DE L'EXAMEN

- Dates et heures du challenge : 3/4/5 juin, salle réservée de 9h/17h
- **Date de l'examen : 7 juin de 9h30 à 12h30**
- Présentation orale de 15 minutes de la/les solution(s) retenue(s) + 5 minutes de questions

QUELQUES CONSIGNES

- Battre le benchmark n'est pas l'objectif principal : il faut comprendre, analyser et justifier votre algorithme
- Vous pouvez bien sûr vous inspirer de méthodes trouvées sur internet, dans des articles, etc.
- Expliquez les résultats numériques : avis sur les sources d'erreurs, signes de sur-apprentissage (validation vs test public vs test privé), les avantages et inconvénients de votre approche
- Proposez des pistes et idées d'amélioration