

Lancement projet

Alice HELIOU and Vincent THOUVENOT
Laboratoire de Data Science de ThereSIS



Content

- 1. Bilan mi-projet**
- 2. Déroulement du projet**
- 3. Rappel sur le cas d'usage**
- 4. Ce qui est fourni**
- 5. Ce qui est attendu**

Bilan mi-projet

Paris-Saclay University - Lancement projet

Bilan mi-projet

Dans l'ensemble vous avez bien compris l'analyse demandée.

La partie biais a été plus hétérogène.

Nous accordons de l'importance à la réflexion et aux explications.

Bilan mi-projet

Dans l'ensemble vous avez bien compris l'analyse demandée.

La partie biais a été plus hétérogène.

Nous accordons de l'importance à la réflexion et aux explications.

Rappels

- L'analyse univariée peut révéler des déséquilibres, l'analyse bivariée peut révéler des biais.
- Les biais peuvent avoir une explication (voire le Paradoxe de Simpson), mais ils restent factuels.
- Un modèle de machine learning (tout comme nous) a tendance à amplifier les biais présents.

Déroulement du projet

Déroulement

- Sessions dédiées: semaines du 10/03, 17/03 et /03
- Rapport final à rendre pour la semaine du 31 mars (une semaine avant votre soutenance, retard non accepté, nous aurons besoin d'avoir vu les rapports avant la soutenance)
- Soutenance: semaine du 7 avril

Déroulement

- Sessions dédiées: semaines du 10/03, 17/03 et /03
- Rapport final à rendre pour la semaine du 31 mars (une semaine avant votre soutenance, retard non accepté, nous aurons besoin d'avoir vu les rapports avant la soutenance)
- Soutenance: semaine du 7 avril
- note finale = $\frac{mi-parcours+notebookfinal+2xsoutenance}{4}$

Rappel sur le cas d'usage

Cas d'usage

Chest X ray NIH 14 Dataset

<https://www.kaggle.com/datasets/nih-chest-xrays/data>

Dataset d'image de rayons X, fourni avec de nombreuses informations sur les personnes (notamment age, genre, maladies). Pour votre projet les images ont toutes été compressées en 256x256.



Ce qui est fourni

Sur ecampus dans le dossier Projet

Les données

Un zip par personne dans le dossier données_par_eleve , chacun contient un dossier avec cette structure: NOM_PRENOM

- train:
 - malade : avec des images
 - sain: avec des images
- valid:
 - malade : avec des images
 - sain: avec des images
- metadata.csv (fichier similaire à celui que vous avez eu pour le mi-projet)

Apprentissage modèle de classification d'image

Vous trouverez un fichier `train_classifieur.py` et un fichier `requirements.txt`. Le premier contient le code pour apprendre et utiliser un classifieur d'image, le second, les requirements pour installer votre environnement python.

Installation environnement

```
python3.10 -m venv projet-env
source projet-env/bin/activate
python -m pip install --upgrade pip
python -m pip install -r requirements.txt
```

Le fichier `train_classifieur.py` peut être utilisé à part ou intégré dans votre code si vous l'importez.

Utilisation dans un notebook

```
from train_classifieur import train_classifieur, pred_classifieur
```

```
ckpt_path, ckpt_score = train_classifieur( # prend entre 20 et 60min  
logdir="./expe_log/",  
datadir="NOM_PRENOM/",  
weights_col="WEIGHTS",  
csv="NOM_PRENOM/metadata.csv")
```

```
pred_classifieur( # prend entre 5 et 10min  
datadir="NOM_PRENOM/",  
csv_in="NOM_PRENOM/metadata.csv"  
csv_out="./expe_log/preds.csv",  
ckpt_path = ckpt_path)
```

Utilisation en ligne de commande

```
python train_classifieur.py
--logdir expe_log/
--datadir NOM_PRENOM/
--csv NOM_PRENOM/metadata.csv
--weights_col WEIGHTS
--csv_out expe_log/preds.csv
```

En ajoutant les arguments

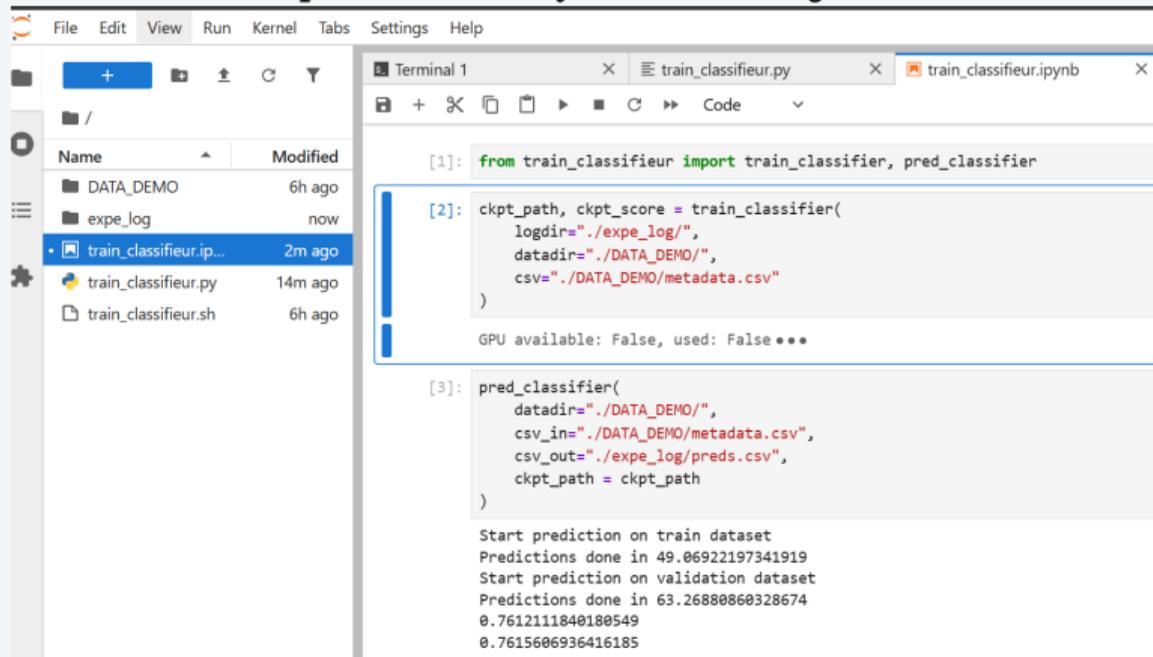
```
"--train False --ckpt /chemin_vers_votre_modele.ckpt"
```

vous pouvez ne faire que les prédictions.

Possibilité d'utiliser Mydocker

https:

[//mydocker.universite-paris-saclay.fr/shell/join/rKSHnVUBveBfHzPjBntC](https://mydocker.universite-paris-saclay.fr/shell/join/rKSHnVUBveBfHzPjBntC)



The screenshot shows a JupyterLab environment. On the left is a file browser with a table of files:

Name	Modified
DATA_DEMO	6h ago
expe_log	now
train_classifieur.ip...	2m ago
train_classifieur.py	14m ago
train_classifieur.sh	6h ago

On the right is a terminal window with the following Python code and output:

```
[1]: from train_classifieur import train_classifier, pred_classifier

[2]: ckpt_path, ckpt_score = train_classifier(
      logdir="./expe_log/",
      datadir="./DATA_DEMO/",
      csv="./DATA_DEMO/metadata.csv"
    )
GPU available: False, used: False ***

[3]: pred_classifier(
      datadir="./DATA_DEMO/",
      csv_in="./DATA_DEMO/metadata.csv",
      csv_out="./expe_log/preds.csv",
      ckpt_path = ckpt_path
    )
Start prediction on train dataset
Predictions done in 49.06922197341919
Start prediction on validation dataset
Predictions done in 63.26880860328674
0.7612111840180549
0.7615606936416185
```

Possibilité d'utiliser Mydocker

- Il faudra cependant y copier vos données, l'environnement n'est pas personnalisé.
- Le dossier /root/fairness_project sera sauvegardé
Vous avez accès à votre répertoire personnel là: /repertoire_personnel/
- d'après quelques tests mydocker est plus lent: il faudrait compter entre le double et le triple (1 à 3h pour un apprentissage, 15 à 30min pour les prédictions)
- Si le temps de calcul est une limitation, n'hésitez pas (en expliquant et justifiant vos choix) à ne considérer qu'une sous-partie de votre jeu de données.

Ce qui est attendu

Ce qui sera évalué

- Analyse rapide des données: En plus court que le mi-projet, étudier le dataset et montrer les déséquilibres et biais.
- Analyse de l'impact de la pondération sur les performances du modèles: Prendre en compte la performance en "balanced accuracy" mais aussi sur les métriques de fairness. Essayer plusieurs méthodes de pondération
- Analyse de l'impact du post-processing: Idem prendre en compte plusieurs critères pour la comparaison et tester plusieurs approches. Regarder les combinaisons pre/post processing
- Soin: Comme toujours le rapport devra comporter une introduction, une conclusion et un soin devra être apporter pour aider la compréhension/lecture

Notation

- 1/4: Rapport final:
 - Introduction (/3)
 - Preparation et analyse des données (/3)
 - Application des méthodes de pre processing (/5)
 - Application des méthodes de post processing (/5)
 - Analyse, compréhension (/3)
 - Conclusion (/1)
- 1/2: Soutenance de 15 minutes, (10' présentation, 5' de questions)
 - Présentation avec illustration du travail effectué
 - Ainsi que de l'organisation choisie au sein du groupe