Документация к диплому

Quasar NN GMM - модель оценки photo-z рентгеновских квазаров.

- data - каталог с данными для обучения и проведения экспериментов. Содержит наборы данных:
  - train20 (DR14Q + VHzQ) - для обучения модели оценки photo-z рентгеновских квазаров.
  - train15 - для обучения модели оценки photo-z галактик.
  - S82X.
  - DR16Q (Z > 5).
  - HELP.
  - Bandos.
  - Yang.
- feature_lists - каталог с наборами признаков для различных конфигураций модели.
- trained_models - каталог с обученными моделями. {#1}_{#2}_g{#3}_m{#4}_model_{#5}.pkl
  - 1 - название обучающего набора данных.
  - 2 - набор признаков (18, 20, 21, 22, 35).
  - 3 - число гауссиан в смеси.
  - 4 - число моделей в ансамбле.
  - 5 - full/01/02 - обучена на полном наборе данных или с помощью двойной кросс-валидации (номер фолда).
- NN_GMM_photoz.py - скрипт для применения обученной модели.
  - model_file - файл с модель.
  - features_file - файл с признаками.
  - features_no - набор признаков.

    - out_file - файл с результатами.
    - Пример запуска: NN_GMM_photoz.py trained_models/train20_35_g5_m5_model_full.pkl data/Yang/part-0.features.gz_pkl 35 out.pkl_gz
    - Файл с признаками должен представлять из себя pickle файл объекта pandas.DataFrame, сжатый алгоритмом gzip, и содержать признаки указанного набора.
- DeepEnsemble.py - содержит классы:
    - HZ_dataloader_new - класс загрузчика данных для обучения и применения модели на PyTorch. Дублирование объектов n_dup раз, каждый раз с вероятностью, определяемой функцией p_func(z). Батч формируется из объектов с z < z_thr и z > z_thr пропорционально распределению объектов в выборке.
    - MLP_GMM - класс многослойного полносвязного перцептрона для оценки смеси нормальных распределений.
    - DeepEnsemble_GMM - Класс модели глубокого ансамбля для оценки смеси нормальных распределений.
- metric.py - функции для расчета нормализованного медианного нормального отклонения.
- train_predict_35_5models_05.ipynb - jupyter notebook с примером обучения и первичного анализа модели на наборе данных train20.
- train_all.ipynb - jupyter notebook, в котором осуществляется обучение на наборах данных train20 и 15 и получение предсказаний на всех наборах данных из каталога data.
- requirements.txt - необходимый набор зависимостей.
- Запуск осуществлялся на ОС Ubuntu 20.04.5 LTS. Необходимо наличие GPU.

Clusters&Filaments - модель построения карты волокон и модель оценки вероятности скопления галактик в заданном направлении и расстояния до него.

- disperse.py - модель на языке python3 для построения карт филаментов. Использует пакет disperse, инструкция по установке - http://www2.iap.fr/users/sousbie/web/html/index888d.html?archive.
- _disperse_03 - собранный пакет disperse.
- ACT_02.ipynb - jupyter notebook для обучения и оценки моделей отождествления галактических скоплений.
- ACT_01_galaxies.csv - каталог галактик (подвыборка SDSS 16) для обучения моделей.
- ACT_01_clusters.csv - каталог скоплений (подвыборка ACT) для обучения и оценки моделей.
- Файлы с промежуточными данными (признаки):
  - ACT_01_dists_ext_train.npy
  - ACT_01_dists_train.npy
  - ACT_02_clusters_ext.csv
  - ACT_02_feas.npy
  - ACT_02_feas_ext.npy
- boosting_ACT_02_{0|1}.txt - файлы с сохраненными моделями.
- requirements.txt - необходимый набор зависимостей.
- Запуск осуществлялся на ОС Ubuntu 20.04.4 LTS.
- Для запуска ACT_02.ipynb необходимо указать путь к папке bin установленного пакета disperse (по инструкции, указанной выше) при инициализации класса алготирма disperse в ячейке 10.