Immagine/Titolo

Christian Gambardella, Vincenzo Offertucci

March 12, 2022

Link GitHub

Index

1 Business Understanding

1.1 Introduzione al problema

Uno dei fondamenti dell'astronomia è la classificazione degli astri celesti in stelle, galassie e quasar(anche detta classificazione stellare). In particolare i quasar sono stati argomento di dibattito all'interno della comunità scientifica per tutta la seconda metà del XX secolo: questi astri, che sembravano stelle, erano troppo luminosi per essere così lontani dal nostro pianeta.

1.2 Obiettivi di business

L'obiettivo del nostro progetto è realizzare un modello di machine learning che sia capace di classificare gli astri celesti, in particolare i quasar, sulla base di dati spettroscopici.

1.3 Descrizione dell'ambiente

PEAS	
Performance	La misura di performance del modello è la sua capacità di avvicinarsi il più possibile alla corretta classificazione dei tre astri celesti
Environment	L'ambiente di riferimento del nostro modello è l'astronomina, inoltre è: • **completamente osservabile** in quanto l'agente in ogni momento ha accesso allo stato completo dell'ambiente; • **episodico** in quanto le azioni del modello in un dato istante non sono influenzate dalle precedenti; • **statico** in quanto l'ambiente rimane invariato mentre l'agenta sta deliberando; • **discreto** in quanto l'agente può ricevere un numero ben definito di percezioni ed effettuare un numero ben definito di azioni.
Actuators	L'agente agisce sull'ambiente tramite lo stream di output del nostro computer fornendo così la tipologia di astro celeste che stiamo valutando
Sensors	L'agente percepirà l'ambiente tramite uno stream di input del nostro computer

1.4 Business success criteria

Per validare il nostro modello adotteremo i seguenti criteri: puntiamo innanzitutto ad avere un accuracy almeno del 90% in quanto i dati a nostra disposizione sono sufficientemente numerosi e molto precisi, si parla comunque di misurazioni effettuate con appositi strumenti. Vogliamo inoltre massimizzare i valori di precision e recall per quanto riguarda l'individuazione dei quasar, che sono l'astro più interessante del nostro problema, in particolare puntiamo a raggiungere l'80& in entrambi i casi.

1.5 Tool da utilizzare

I tool che utilizzeremo per realizzare il nostro modello sono i seguenti:

- **Python**
- **Anaconda**
- **ScikitLearn**
- **Pandas**
- **Kaggle**
- **JupiterLab**
- **Mathplot**
- **TeXStudio**
- **MikTeX**

2 Data Understanding

2.1 Scelta del dataset

Per la realizzazione del nostro progetto, dopo svariate ricerche in rete, abbiamo deciso di adottare questo dataset per la realizzazione del nostro modello di machine learning.

2.2 Analisi del dataset

Nel dataset in questione i dati sono stati collezionati nell'ultimo trentennio da parte della SDSS (Sloan Digital Sky Survey) che si è occupata di processare le foto degli astri celesti in dati, in particolare noi stiamo usando il data release 17 della SDSS-IV. Da notare che il dataset usato da noi non contiene tutte le colonne dell'originale bensì è stata fatta una selezione di 18 (a partire dai 153 iniziali). Nel dataset sono presenti 17 colonne:

- **obj_ID**: è un valore unico che identifica l'oggetto all'interno del catalogo di immagini processate da SDSS.
- **alpha**: ascensione retta, una misura analoga alla longitudine ma proiettata sulla sfera celeste anzichè sulla superficie terrestre.
- **delta**: angolo di declinazione, rappresenta una delle coordinate equatoriale per determinare l'altezza di un astro della sfera celeste (analogo alla latitudine).
- **u**: filtro ultravolietto del sistema fotometrico.
- **g**: filtro verde del sistema fotometrico.
- **r**: filtro rosso del sistema fotometrico.
- **i**: filtro vicino all'infrarosso del sistema fotometrico.

- **z**: filtro infrarosso del sistema fotometrico.
- **run_ID**: è un valore unico che identifica la scansione utilizzata.
- **rerun_ID**: è un valore unico che identifica la modalità con cui l'immagine è stata processata.
- **cam_col**: è un valore che indica quale colonna della camera è stata utilizzata nella scansione.
- **field_ID**: è un valore unico che identifica ogni campo.
- **spec_obj_ID**: è un valore unico che identifica l'astro all'interno del catalogo di immagini processato da SDSS(nel dataset originale erano presente più oggetti relativi allo stesso astro).
- **class**: è la nostra variabile target/dipendente, può assumere i valori "STAR", "GALAXY" o "QUASAR".
- **redshift**: è il valore assunto dal redshift dell'astro basato sull'incremento della lunghezza d'onda (lo spostamento di un astro è da noi percepito come una variazione dello spettro elettromagnetico tendente verso il rosso).
- **plate**: è un valore unico usato come identificatore all'interno dei sistemi SDSS.
- **MJD**: è una versione modificata della data giuliana, in particolare corrisponde a 2400000.5 dopo il giorno 0 del calendario giuliano.
- **fiber_ID**: è un valore unico che identifica la fibra ottica che ha puntato la luce all'interno del piano focale.

//inserire immagine 16.76 Mb 100k righe

2.3 Data quality

Nel dataset non sono presenti dati mancanti, inoltre i dati presentano tutti lo stesso formato (numerico) ma su scale eterogenee, sarà dunque fondamentale prestare attenzione a questo aspetto in fase di data preparation.

2.4 Data exploration

Non sono state individuate relazioni tra i dati del dataset. //immagine di data exploration

3 Data Preparation

3.1 Data cleaning

Dati i risultati ottenuti in fase di data exploration non troviamo necessario l'utilizzo di nessuna tecnica o algoritmo di data imputation.

3.2 Define data and target variables

La variabile target è //corsivaograssa class ed è la quattordicesima colonna del dataset.

3.3 Data balancing