Progetto Machine Learning

Spotify award

Gianluca Giudice - 830694

Marco Grassi - 829664

Contents

Chapter 1

Introduzione

In questo capitolo viene introdotto il problema, il dominio di riferimento e l'approccio adottato per la risoluzione.

1.1 Descrizione del problema

In questo elaborato consideriamo i singoli musicali. Al giorno d'oggi le canzoni vengono spesso ascltate dagli utenti tramite piattaforme di streming musicale.

L'obbiettivo del lavoro è quello di analizzare le features dei singoli musicali così da prevedere se una canzone diventerà o meno di successo. Nella sezione successiva verrà meglio specificato cosa si intende per **singolo musicale di successo** (subsection 1.1.2).

1.1.1 Spotify

Spotify è un servizio di riproduzione digitale in streaming di musica, podcast e video, con accesso immediato a milioni di brani e altri contenuti di artisti provenienti da tutto il mondo. Questa piattaforma viene utilizzata da milioni di utente per ascoltare canzoni e nello specifico singoli musicali.

Da questo servizio è possibile ottenere migliaia di brani musicali, infatti Spotify mette a disposizione una API da cui è possibile scaricare informazioni su questi brani con assocciate alcune caratteristiche. Pertanto oltre alle classiche informazioni di un brano come "titolo" o "artisti" si avrà a disposizione una serie di caratteristiche specifiche del brano, ad esempio quanto una canzone è "energica" o "ballabile".

1.1.2 Premi per i singoli musicali

Come vedremo in subsection 2.2.1 il dataset ottenuto da Spotify mette a disposizione un campo "popularity" che indica quanto può essere considerata popolare. Tuttavia questa caratteristica non ci sembra adeguata per identificare una canzone come di successo oppure no.

Per questo motivo consideriamo una canzone di successo in base alle certificazioni ottenute, rispettivamente "disco d'oro" o "disco di platino". Queste premi sono dei riconoscimenti vinti da un brano musicale e storicamente fanno riferimento al numero di copie vendute da un singolo. Tuttavia con la costante crescita dell'utilizzo di servizi per lo streming di brani musciali, da qualche anno questi riconoscimenti vengono assegnati anche considerando il numero di streaming sulle diverse piattaforme, tra cui Spotify.

Riteniamo che questo riconosciemnto sia una metrica oggettiva per considerare un singolo come di successo.

1.2 Approccio al problema

Il problema viene approcciato come un **task di classificazione binaria**. Dato un singolo vogliamo prevedere se questo sarà di successo o no. Pertanto sviluppiamo modelli supervisionati di machine learning per classificare i singoli.

1.3 Struttura del codice

Di seguito viene brevemente spiegata la struttura del codice, inoltre viene sotto indicata la working directory e l'entry point del programma.

```
root
├── data
│   ├── raw
│   │   Dataset da integrare
│   └── 'songs.csv'
│       Dataset integrato da dare in input agli algoritmi
├── doc
│   Relazione progetto
├── images
│   Immagini prodotte in output dagli script
└── scraper
    ├── 'combine_data.ipynb'
    │   Notebook per combinare dataframe dei vari dischi d'oro delle nazioni
    ├── 'wikipedia_songs_scraper.py'
    │   Scraper per scaricare da wikipedia le informazioni riguardo le
    │   certificazione dei singoli per le diverse nazioni
    ├── 'import_db.sh'
    │   Import del dataset in mongoDB.
    └── 'record_linkage.js'
        Query in mongoDB per la fase di record linkage.
```

In particlare gli **script in R** si trovano in:

```
root
└── src
    ├── 'main.R'
    │   Entrypoint script in R
    └── functions
        ├── 'preprocessing.R'
        │   Funzioni per il preprocessing e visualizzazione del dataset
        ├── 'training_svm.R'
        │   Training modelli support vector machine
        └── 'training_decisiontree.R'
            Training modelli decision tree
```

N.B.: Per come è stato progettato il codice, la working directory è la `root/` e non la cartella `src/`. L'**entry point del programma** è lo script `src/main.R`

Chapter 2

Dataset

In questo capitolo viene analizzato il dataset. Viene quindi spiegato da dove è stato preso il dataset, descritte le covariate e viene effettuata un'analisi esplorativa.

2.1 Acquisizione del dataset

Il dataset completo contenente informazioni sui singoli e riguardo le varie certificazioni non è disponibile da una sola sorgente. Pertanto abbiamo ottenuto le informazioni necessarie da diversi sorgenti per poi integrare i dati.

2.1.1 Dataset spotify - Kaggle

Per quanto riguarda i brani con le relative informazioni e caratteristiche del brano, Spotify mette a disposizone un'API da cui si possono ottenere questi dati.

Con questa tecnica sono stati ottenuti i dati relativi ai brani, un utente ha quindi caricato il dataset sul sito kaggle.com. Il dataset è disponibile su kaggle all'url indicato. [1].

Il dataset contenente queste informazioni è il file: `data/raw/to_integrate/data.csv`

2.1.2 Premi canzoni - Wikipedia

Le informazioni riguardo i premi delle canzoni, ovvero le varie certificazioni vinte, sono disponibili su wikipedia.org.

Nello specifico siamo interessati a:

- Singoli certificati oro
- Singoli certificati platino
    1. Singoli certificati 1 volta platino
    2. Singoli certificati 2 volte platino
    3. ...
    4. Singoli certificati N volte platino

Inoltre le varie certificazioni dei singoli vengono considerati in base al paese. I paesi da noi presi in considerazione sono:

[1] **Dataset Spotify**: https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks.

1. Italia
2. Australia
3. Stati Uniti d'America
4. Regno Unito
5. Canda
6. Danimarca

Si noti come una certificazione può essere consegnata in diversi paesi alla stessa canzone.

Certificazioni disco d'oro

Per quanto riguarda i dischi d'oro, facciamo riferimento a questa pagina su wikipedia [2]. Fissato quindi uno stato (ad esempio l'Italia) è possibile visualizzare la pagina contenente la lista dei singoli che hanno vinto quel particolare premio. La lista è un elenco di url che puntano alla pagina wikipedia della canzone, un esempio a questo url[3].

Certificazioni disco di platino

Ragionamento analogo viene fatto per i dischi di platino, con l'unica differenza che la pagina è questa[4], inoltre dal momento che i singoli posso vincere più volte un disco di platino, si considerano non solo i singoli che hanno vinto una volta il dischi di platino ma anche quelli che l'hanno vinto N volte.

Scraping

Ottenuti quindi i puntatori alle canzoni certificate, è possibile accedere alla pagina wikipedia del singolo, la quale contiene una tabella riassuntiva della canzone. Un esempio di tabella viene mostrato nella Figura 2.1.

[2] **Disco d'oro per stato**:
https://it.wikipedia.org/wiki/Categoria:Singoli_certificati_oro_per_stato.

[3] **Singoli certificati disco d'oro in Italia**:
https://it.wikipedia.org/wiki/Categoria:Singoli_certificati_disco_d%27oro_in_Italia.

[4] **Singoli certificati disco di platino per stato**:
https://it.wikipedia.org/wiki/Categoria:Singoli_certificati_platino_per_stato.

Chico (singolo)

Da Wikipedia, l'enciclopedia libera.

Questa voce sull'argomento singoli hip hop è solo un abbozzo.
Contribuisci a migliorarla secondo le convenzioni di Wikipedia. Segui i suggerimenti del progetto di riferimento.

Chico è un singolo del rapper italiano Gué Pequeno, pubblicato il 31 luglio 2020 come secondo estratto dal sesto album in studio *Mr. Fini*.[2]

Indice [nascondi]
1 Classifiche
1.1 Classifiche settimanali
1.2 Classifiche di fine anno
2 Note
3 Collegamenti esterni

Classifiche [modifica | modifica wikitesto]

Classifiche settimanali
[modifica | modifica wikitesto]

Classifica (2020)	Posizione massima
Italia[3]	5

Classifiche di fine anno
[modifica | modifica wikitesto]

Classifica (2020)	Posizione
Italia[4]	10

Chico	
Artista	Gué Pequeno
Featuring	Rose Villain e Luchè
Tipo album	Singolo
Pubblicazione	31 luglio 2020
Durata	3:33
Album di provenienza	*Mr. Fini*
Genere	Pop rap
Etichetta	Island
Produttore	Sixpm
Formati	Streaming
Certificazioni	
Dischi di platino	Italia (3)[1] (vendite: 210 000+)
Gué Pequeno - cronologia	
Singolo precedente	**Singolo successivo**
Saigon (2020)	*Bla Bla* (2020)
Luchè - cronologia	
Singolo precedente	**Singolo successivo**
Come me (2019)	*Maserati (Reloaded)* (2020)

Figure 2.1: Esempio di tabella per un singolo musicale su wikipedia

Viene quindi creato uno script per effettuare scraping delle tabelle, il codice è nella directory `scraping/wikipedia_songs_scraper.py`.

Successivamente si integrano le informazioni dei diversi stati, e tipi di certificazione vinti. Il risultato di questa operazione è il file `data/raw/to_integrate/awards_cleaned.csv`.

2.2 Descrizione del dataset

In questa sezione vengono elencate e descritte le features del dataset.

2.2.1 Spotify

Di seguito viene descritto il dataset proveniente da kaggle, ovvero quello contenente le informazioni dei brani.

ATTRIBUTO	DESCRIZIONE	TIPO
id	Identificativo della canzone (generato da spotify)	Intero
name	Titolo della canzone	Stringa
artists	Lista degli artisti che compaiono nel brano	Stringa
year	Anno del brano	Intero
duration_ms	Durata della traccia in millisecondi	Intero
acousticness	Metrica riguardante quanto un brano risulta "acustico"	Float [0, 1]
danceability	Metrica riguardante quanto una traccia è ballabile	Float [0, 1]
energy	Energia del brano	Float [0, 1]
explicit	Indica se la traccia è esplicita oppure no (linguaggio volgare)	Binario
instrumentalness	Contenuto relativo di strumenti musicali nella traccia	Float [0, 1]
valence	Metrica riguardante la "positività" della traccia	Intero
key	Chiave musicale utilizzata	Intero [0, 11]
liveness	Durata relativa della traccia suonata in una performance dal vivo	Float [0, 1]
loudness	Rumorosità della traccia in decibel (dB)	Float [-60, 0]
mode	Indica se la traccia parte con una progressione armonica	Booleano
release_date	Anno di rilascio del brano	Intero
speechiness	Contenuto relativo di voce umana nella traccia	Float [0, 1]
tempo	BPM della traccia	Float
popularity	Popolarità della traccia	Float [0, 100]

2.2.2 Premi

Si noti come la distinzione tra tipo di premio vinto e lo stato in cui è stata ottenuta la certificazione per uno specifico brano, viene fatta solo a scopo dello scraping, in quanto i brani sono così rappresentati sul sito di wikipedia. Tuttavia da questo punto in poi non verrà più tenuto conto di questa informazione, infatti un singolo verrà considerato come **vincitore di un premio** (e quindi di successo) oppure come **non vincitore di un premio** (non di successo).

Il risultato dello scraping della tabella di wikipedia è il seguente:

ATTRIBUTO	DESCRIZIONE	TIPO
title	Titolo della canzone	Stringa
artists	Artisti presenti nela traccia	Stringa
date	Data di rilascio della traccia	Data
genre	Genere musicale del brano	Stringa
award	Premio vinto dal singolo	Stringa {Oro, 1_platino, 2_platino, ...}
nation	Paese in cui è stato vinto il premio	Stringa (Sigla del paese)

2.3 Normalizzazione

OpenRefine con regex per le date

2.4 Data integration

Possiamo etichettare i brani come di successo oppure non di successo.

2.4.1 Modellazione struttura documentale

2.4.2 Record linkage con MongoDB

Una roba veloce giusto per spiegare

1. Import dati mongodb
2. Creazione indici

3. Preprocessing lowercase
4. Unfold campo artista
5. Record linkage -¿ Join basandosi su titolo e artisti
6. Lista artisti come stringa
7. Dump del database in un file .csv come risultato del dataset per i modelli

2.5 Analisi esplorativa

2.5.1 Distribuzione dei valori

Variabili continue

Variabili categoriche

2.5.2 Artisti nelle canzoni

Frequenza artisti

Wordcloud

2.5.3 Correlazione features

2.5.4 Principal component analysis

Prima di PCA normalizzazione del dataset

NB: Non facciamo il plot delle prime due componenti principali in quanto varianza spiegata dalle prime due componenti molto bassa

2.5.5 Creazione dataset bilanciato

Undersampling

2.6 Scelta delle features

2.6.1 Normalizzazione e standardizzazione

Conversione in lowecase Converto in variabili numeriche e factor + label corretta

2.6.2 Dataset proiettato componenti principali

2.6.3 Variabili categoriche

Key -¿ Trasformazione a intero

2.6.4 Artisti in una canzone - BOW

Threshold freqeunza

Viene usato un threshold a 2 per non avere delle canzoni con artisti completamente sconosciuti. Assumiamo quindi di fare previsioni su canzoni cantante da artisti un minimo conosciuti.

Questa assunzione non è molto restrittiva, ci immaginiamo infatti che per vincere un premio una canzone deve essere cantata da artisti non completamente sconosciuti

Inoltre a causa dell'undersampling è possibile avere diverse canzoni cantate da artisti "sconosciuti" e non avendo scelto una particolare strategia per l'undersampling adottiamo a questo punto l'utilizzo di un threshold

2.6.5 Popularity

Chiaramente questa feature viene scartata per non creare modelli biased (Non si conosce questa feature quando una canzone esce)

Chapter 3

Campagna sperimentale

3.1 Approccio

3.1.1 10-folds cross validation

3.1.2 Training set e test set

3.1.3 Model selection

Ottimizzazione iperparametri

Grid search

3.2 Misure di performance

3.2.1 Accuracy

3.2.2 Precision, Recall e F-measure

3.2.3 Curve ROC

3.3 Support Vector Machine

3.3.1 Kernel

3.4 Decision Tree

3.5 Esperimenti

3.5.1 Performance

3.5.2 Modelli a confronto

Chapter 4

Conclusioni

Perchè le performance sono basee?

Spotify genera le caratteristiche con un algorimto quindi sicuramente un po' approssimato

E' effittivamente difficile capire se una canzone vincerà un premio, dipende da molti fattori