INDICE

Link Repo GitHub: https://github.com/TraneLoneWolf/SG-Library

Esame Fondamenti di Intelligenza Artificiale

Gallo Luigi 0512105704

Nocerino Gaetano Giovanni 0512105974

INTRODUZIONE

SCOPO DEL SISTEMA

L'obiettivo è quello di realizzare un algoritmo suggeritore per gli utenti di una piattaforma e-commerce di libri. L'algoritmo si occupa di realizzare una lista di suggerimenti per gli acquisti futuri, attraverso un modulo di intelligenza artificiale, capace di apprendere le preferenze degli utenti in funzione delle loro interazioni con i prodotti della piattaforma.

DEFINIZIONE DEL PROBLEMA

Lo scopo del sistema è quello di comprendere come determinare l'insieme di libri da consigliare all'utente, elaborando le informazioni degli ordini effettuati dallo stesso. Volendo generalizzare: se un utente è solito acquistare libri di un certo autore/genere/categoria, l'algoritmo dovrà apprendere le sue preferenze e suggerirgli i libri che più si avvicinano ad esse.

SPECIFICHE AMBIENTALI

SPECIFICA PEAS

PEAS	
PERFORMANCE	Percentuale di consigli pertinenti
AMBIENTE	E-commerce
ATTUATORI	Visualizzazione dei libri consigliati
SENSORI	Collegamento al database

Link Repo GitHub: https://github.com/TraneLoneWolf/SG-Library
Esame Fondamenti di Intelligenza Artificiale

Gallo Luigi 0512105704
Nocerino Gaetano Giovanni 0512105974

PROPRIETA' AMBIENTE

L'ambiente comprenderà i dati relativi all'e-commerce

AMBIENTE	
COMPLETAMENTE OSSERVABILE	I sensori possono accedere in qualsiasi momento a tutte le informazioni dell'ambiente
NON DETERMINISTICO	Lo stato successivo dell'ambiente non è determinato solamente dallo stato corrente e dall'azione dell'agente
SEQUENZIALE	Ogni azione eseguita dall'agente avrà un impatto sulle azioni successive ad essa
DINAMICO	I libri e gli utenti presenti nel bookstore online saranno soggetti a continui cambiamenti
DISCRETO	L'agente viene chiamato solo un numero determinato di volte
AGENTE SINGOLO	Nel sistema opererà un singolo agente.

SPECIFICHE PROGETTUALI

LINGUAGGIO E TECNOLOGIE

Per la realizzazione di questo progetto si è optato per il linguaggio Python, particolarmente indicato nella realizzazione di software Data Science oriented, nello specifico avvalendosi del framework Anaconda, che fornisce tutti gli strumenti per effettuare l'analisi (pandas, numpy), la manipolazione, la rappresentazione dei dati (matplotlib, seaborn) e l'implementazione degli algoritmi di apprendimento (scikit-learn).

Link Repo GitHub: https://github.com/TraneLoneWolf/SG-Library
Esame Fondamenti di Intelligenza Artificiale

Gallo Luigi 0512105704
Nocerino Gaetano Giovanni 0512105974

DEFINIZIONE DATASET

Considerando che gli algoritmi di apprendimento si basano sui dati, ci siamo concentrati su di essi sin dalle prime fasi dello sviluppo, inizialmente individuando due possibili approcci alla realizzazione:

· utilizzare dataset di grandi dimensioni reperiti online;

· utilizzare un dataset ridotto creato su misura per l'e-commerce.

Analizzando l'eventualità del primo approccio ne abbiamo subito notato l'estrema inefficacia, in quanto analizzando il dataset vi erano molti valori nulli, e l'eliminazione dei datapoint incompleti ne riduceva le dimensioni di circa il 10%, caratterizzando dunque una pessima qualità dei dati. Altra questione da prendere in seria considerazione, è quella della suddivisione dei libri in categorie non sempre rilevanti, in quanto la maggior parte delle categorie individuate avevano cardinalità compresa tra 1 e 5 libri ciascuna; è dunque subito evidente come su un dataset di considerevoli dimensioni, una informazione come la categoria non sia di fatto utilizzabile data la sua scarsa efficacia.

In seguito a questa attenta analisi si è deciso di procedere con il secondo approccio, e dunque di realizzare da zero un dataset ad-hoc per la piattaforma, composto da:

· ISBN (intero)

· Titolo (stringa)

· Autore (stringa)

· Editore (stringa)

· Prezzo (float)

· Anno di pubblicazione (intero)

· Categoria (stringa)

· Basato su storia vera (booleano)

· Edizione illustrata (booleano)

· Appartenente a una saga (booleano)

· Numero pagine (intero)

Per la realizzazione del dataset abbiamo utilizzato uno script Python "create_dataset.py" che, dopo aver importato il file .csv (libri) frutto di un dump del database di sistema (query SQL) è stato ottimizzato, eliminando le colonne non rilevati (titolo, descrizione) al fine dell'operazione. Successivamente, applicando una tecnica di scaling (scaler), abbiamo sostituito tutti i dati testuali in numeri interi al fine di migliorare le prestazioni di processabilità dell'algoritmo.

Link Repo GitHub: https://github.com/TraneLoneWolf/SG-Library

Gallo Luigi 0512105704
Nocerino Gaetano Giovanni 0512105974

Conclusivamente per dettagliare al meglio i suggerimenti per l'utente, si è scelto di aggiungere le seguenti informazioni relative al:

· numero di recensioni positive ricevute dal sistema (intero),

· numero di acquisti avvenuti nel sistema (intero);

(Informazioni contenute nel file user*_dataset.txt)

Nell'ambito dei dati utenti è stato eseguito un dump dal DB (query SQL), degli ISBN dei libri con i quali l'utente ha interagito, per scopi dimostrativi ne è stato simulato il funzionamento con l'inserimento di quattro utenti fittizi, creati ad-hoc per testare ogni funzionalità dell'algoritmo. I dati dei singoli utenti sono contenuti nei file del tipo *user_X.txt*. Nello script create_dataset.py si è inserita una colonna al dataset precedentemente ottenuto contenente un booleano, vero se l'utente inserito in input ha interagito con un determinato titolo, falso altrimenti, i risultati sono invece contenuti nel file user_X_dataset.txt.
Lo script ha inoltre il compito di generare il file ISBN_index.txt, che associa ogni ISBN all'indice del dataset risultante, e verrà utilizzato per ricostruire le informazioni relative al titolo e l'autore.

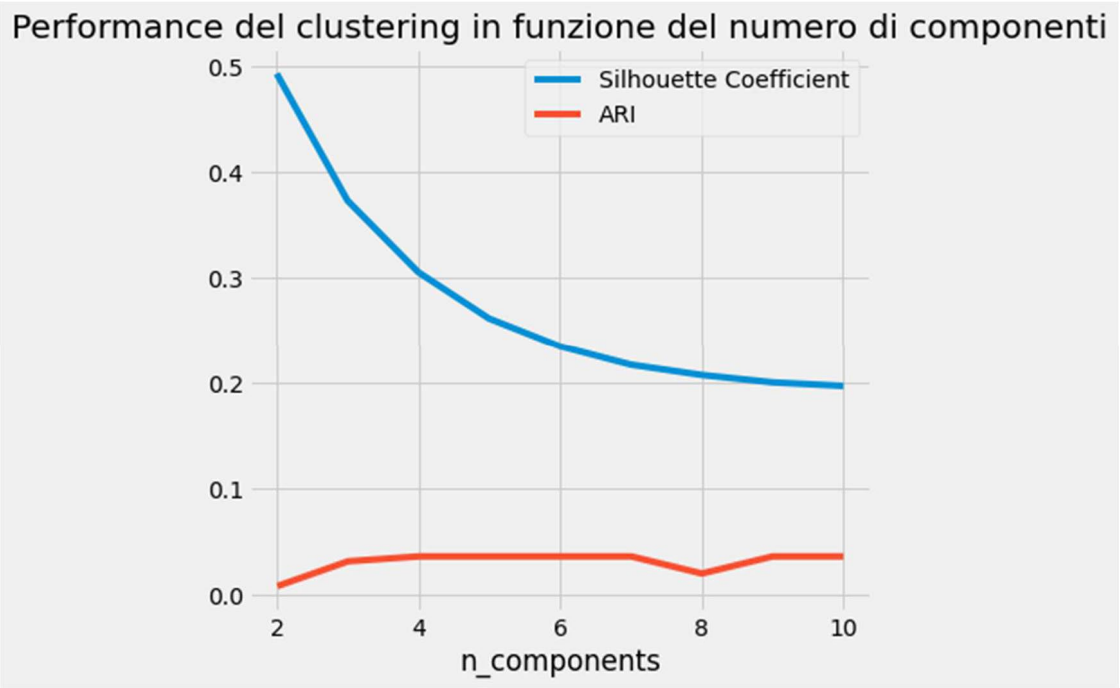
SCELTA DELL'ALGORITMO

Per quanto riguarda la scelta dell'algoritmo di apprendimento sono stati presi in considerazione sia gli algoritmi di apprendimento supervisionato(classificazione), che gli algoritmi di apprendimento non supervisionato(clustering). Alla fine si è scelto di utilizzare un algoritmo di clustering, dal momento che la struttura dei dati sarebbe stata difficilmente gestibile da un algoritmo di classificazione per via di una mancata definizione chiara delle etichette. Il clustering permette dunque di raggruppare i singoli datapoint in insiemi di elementi simili in base a dei pattern ritrovati nei dati. Un'ipotesi di come consigliare una lista di libri in output all'utente è quello di selezionare il cluster contenente il maggior numero di libri con cui l'utente ha già interagito. L'algoritmo di base da cui si è pensato di iniziare è il k-means. Per poter operare su un qualsiasi algoritmo di clustering bisogna scalare le features più grandi che possono avere più peso rispetto ad altre più piccole ma più rilevanti. Per scalare i dati è stato utilizzato il package sklearn.preprocessing. Il metodo di scaling utilizzato è lo StandardScaler(). Esso è un metodo di standardizzazione delle feature che si ottiene rimuovendo la media e dividendo per la deviazione standard. Questo serve per ridistribuire i dati in una distribuzione normale standardizzata. Si è utilizzata la PCA, ossia l'analisi delle componenti principali, che permette di ridurre le dimensioni del dataset, cercando un nuovo insieme di variabili con dimensione minore ma senza perdere il contenuto informativo del dataset iniziale. L'implementazione di questo metodo è stata effettuata utilizzando il package sklearn.decomposition. Il k-means, dal package sklearn.cluster, ha come problema la scelta del numero iniziale dei cluster. Per questo motivo sono state effettuate varie prove per trovare il k più adatto, e si è cercato di analizzare quanto una diversa inizializzazione dei centroidi potesse influire sulla scelta dei dati. Per stimare la qualità dei cluster risultanti abbiamo utilizzato le seguenti metriche di valutazione:

• silhouette coefficient

• ARI (Adjusted Rand Index)

Il numero di cluster risultante il migliore nella maggior parte dei casi è compreso tra 2 e 3, quindi la versione finale ha k = 3.

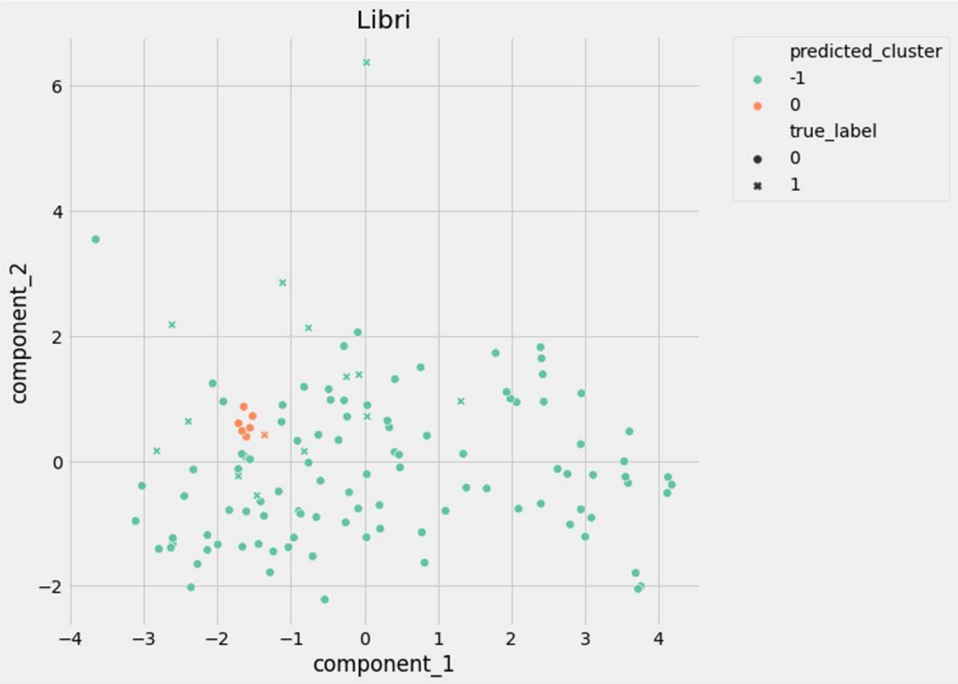


Questo tipo di grafico mostra l'andamento delle due metriche scelte precedentemente in base al numero di cluster. Il primo, che rappresenta il Silhouette Coefficient, decresce linearmente poiché dipende dalla distanza tra i datapoint. La seconda, ossia la ARI, non migliora significativamente al crescere dei cluster; ciò significa che dal primo risulta essere il k migliore il numero 2 mentre il secondo dà un piccolo miglioramento (seppur impercettibile) con il valore 3. Come si evince anche dal grafico, la purezza dei cluster non è eccellente, dal momento che questi numeri dovrebbero tendere a 1. Purtroppo, questo è dovuto alla piccola quantità di dati su cui è stato testato l'algoritmo, il quale con altre accortezze potrebbe essere sicuramente migliore. Il progetto, infatti, si presta molto all'espansione per via della struttura tramite pipeline e della semplicità di utilizzo degli algoritmi già implementati in sklearn.

Gallo Luigi 0512105704
Nocerino Gaetano Giovanni 0512105974

CONFRONTO CON DBSCAN

Il confronto con altro algoritmo di clustering come il DBSCAN ci fa subito notare la sua evidente inadeguatezza nel manipolare dati che non hanno una forma definita:




```
silhouette_score(preprocessed_data, predicted_labels_DB)
```

```
-0.11964472337335649
```

```
adjusted_rand_score(labels, predicted_labels)
```

```
0.008709306040733993
```

Come è visibile dal grafico, questo algoritmo dividerebbe in maniera casuale i dati senza apportare alcuna forma di raggruppamento ad essi, diventando di fatto fortemente sconsigliato all'utilizzo.

SPECIFICA SULL'IMPLEMENTAZIONE

Come documentato nella pagina model_notebook di Jupyter, presente nella repository, il processo ha origine utilizzando i dati prodotti dallo script create_dataset.py (path data/script).
Una volta selezionate le feature e la label (che indica se l'utente ha interagito o meno con un determinato prodotto).

Link Repo GitHub: https://github.com/TraneLoneWolf/SG-Library
Esame Fondamenti di Intelligenza Artificiale

Gallo Luigi 0512105704
Nocerino Gaetano Giovanni 0512105974

Come definito nella work pipeline:

- Definiamo un preprocessore, composto dallo StandardScaler e la PCA, che riduce in due componenti i dati:

```
preprocessor = Pipeline(
    [
        ("scaler", StandardScaler()),
        ("pca", PCA(n_components=2, random_state=0)),
    ]
)
```

- Definiamo l'algoritmo di clustering:

```
clusterer = Pipeline(
   [
       (
           "kmeans",
           KMeans(
               n_clusters=n_clusters,
               init="random",
               n_init=50,
               max_iter=300,
               random_state=0,
           ),
       ),
   ]
)
```

  - n_clusters è una variabile che in questo caso assume valore 3;
  - il metodo di inizializzazione è 'random', dunque si sceglieranno n_clusters righe casualmente dai dati per definire i centroidi iniziali;
  - n_init è il numero di volte che l'algoritmo k-means verrà riprodotto con una scelta diversa dei centroidi. Il risultato finale scelto sarà l'output migliore in base all'inerzia dei clusters, ossia quanto possono essere considerati coerenti internamente i singoli clusters;
  - max_iter è il numero massimo di iterazioni che può eseguire l'algoritmo per una singola esecuzione.

- Dichiariamo la composizione della pipeline

```
pipe = Pipeline(
    [
        ("preprocessor", preprocessor),
        ("clusterer", clusterer)
    ]
)
```

- Eseguiamo il metodo fit sulle features.

Gallo Luigi 0512105704
Nocerino Gaetano Giovanni 0512105974

Una prima analisi ottenuta dal clustering è la seguente, che è composta dai dati pre- processati e le label risultanti dall'esecuzione:

```
preprocessed_data = pipe["preprocessor"].transform(features)
```

```
predicted_labels = pipe["clusterer"]["kmeans"].labels_
```

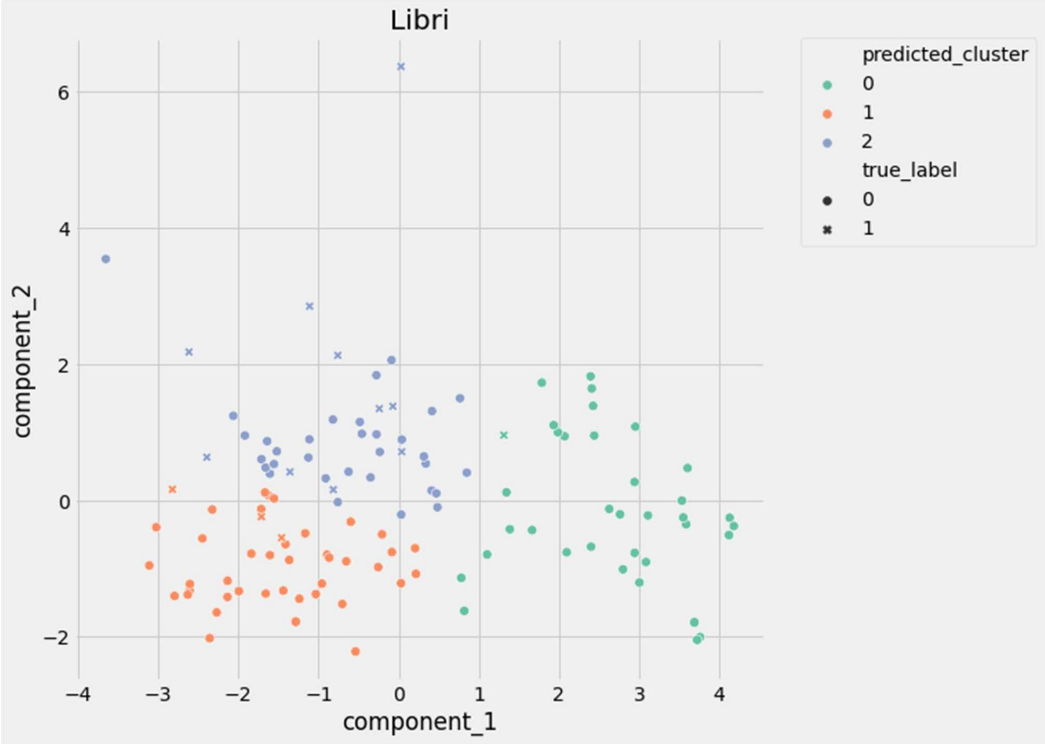
```
silhouette_score(preprocessed_data, predicted_labels)
```

```
0.3978327487551547
```

```
adjusted_rand_score(labels, predicted_labels)
```

```
0.008709306040733993
```

Per avere un'idea migliore dei dati li visualizziamo nei tre cluster e con le label originarie:



Il processo ha ancora dei dati non descritti correttamente, ma la suddivisione sembra comunque accettabile.

I titoli che andranno in output nel sistema finale saranno presi dal cluster con il maggior numero di label 1, quindi ad esempio, in questo caso, il cluster 0 in verde.

Gallo Luigi 0512105704
Nocerino Gaetano Giovanni 0512105974

I libri con cui ha interagito l'utente selezionato sono:

	ISBN	Titolo	Autore	Editore	Prezzo	Anno	Nome_Categoria
15	9788245335872	L isola di Arturo	Elsa Morante	Einaudi	13.0	2019	Narrativa italiana
34	9788804670520	Il mondo nuovo	Aldous Huxley	Mondadori	14.0	2016	Fantascienza
43	9788804726685	Il tempo della clemenza	John Grisham	Mondadori	22.0	2020	Thriller
44	9788804727880	Illuminismo adesso	Steven Pinker	Mondadori	18.0	2020	Saggistica
66	9788820067748	Pet Sematary	Stephen King	Sperling & Kupfer	20.0	2019	Horror
69	9788822719713	La guerra dei mondi	G.W. Wells	Newton Compton Editore	4.9	2018	Fantascienza
72	9788830104716	Il Signore degli Anelli	J.R.R. Tolkien	Bompiani	50.0	2020	Fantasy
76	9788834738955	La svastica sul sole	Philip K. Dick	Fanucci Editore	16.0	2019	Fantascienza
84	9788845298752	Homo Deus	Yuval Noah Harari	Bompiani	16.0	2018	Saggistica
85	9788845932984	Essere una macchina.	Mark O Connell	Adelphi	19.0	2018	Saggistica

E i consigli sono:

Gallo Luigi 0512105704
Nocerino Gaetano Giovanni 0512105974

27	9788804664994	Ciclo delle Fondazioni. Prima Fondazione-Fonda...	Isaac Asimov	Mondadori	16.0	2017	Fantascienza
30	9788804666905	Alexandros. La trilogia	Valerio Massimo Manfredi	Mondadori	17.0	2016	Narrativa storica
36	9788804676379	Storie della buonanotte per bambine ribelli. 1...	Favilli/Cavallo	Mondadori	20.0	2017	Ragazzi
38	9788804688846	Storie della buonanotte per bambine ribelli 2	Favilli/Cavallo	Mondadori	20.0	2018	Ragazzi
41	9788804717638	Gli invisibili	Valerio Varesi	Mondadori	16.0	2019	Thriller
42	9788804721871	La nona casa	Leigh Bardugo	Mondadori	20.0	2020	Narrativa straniera
45	9788804730422	Storie della buonanotte per bambine ribelli. 1...	Favilli	Mondadori	20.0	2020	Ragazzi
48	9788806237530	Peccato mortale. Indagine del commissario De Luca	Carlo Lucarelli	Einaudi	17.0	2018	Gialli
50	9788806242442	L inverno piu nero. Indagine del commissario D...	Carlo Lucarelli	Einaudi	18.0	2020	Gialli
57	9788811608776	A sangue freddo	Truman Capote	Rizzoli	18.0	2020	Narrativa Straniera
59	9788811671572	L estate dell innocenza	Clara Sanchez	Garzanti	15.0	2020	Rosa
61	9788811682691	Il linguaggio segreto dei fiori	Vanessa Diffenbaugh	Rizzoli	10.0	2013	Rosa
64	9788817106825	Dio di illusioni	Donna Tartt	BUR	13.0	2014	Gialli
67	9788820068288	Se scorre il sangue	Stephen King	Sperling & Kupfer	20.0	2019	Horror

CONCLUSIONI

CONSIDERAZIONI FINALI

Allo stato attuale, il livello di sviluppo generale del progetto è ancora in stato embrionale e lascia un ampio margine per futuri miglioramenti ed un considerevole ingrandimento della base dati.

Link Repo GitHub: https://github.com/TraneLoneWolf/SG-Library

Esame Fondamenti di Intelligenza Artificiale

Gallo Luigi 0512105704
Nocerino Gaetano Giovanni 0512105974