Università degli Studi di Padova

Dipartimento di Matematica "Tullio Levi Civita"

Corso di Laurea Magistrale in Informatica

Applicazione di un sistema di raccomandazione in ambito BTB

Relatore
Fabio Aiolli
Università di Padova

Laureando
Daniel Rossi

Ringraziamenti

Da fare

Abstract

La seguente tesi è il risultato della collaborazione svolta nel periodo intercorso tra marzo e agosto 2021 tra il sottoscritto, l'Università di Padova, nella persona del Professor Aiolli, e l'azienda Estilos.
Nel mondo della vendita online, ma non solo, sentiamo parlare ormai sempre più spesso di sistemi di raccomandazione, in questa tesi andremo ad applicare diversi suoi approcci in un contesto non del tutto usuale, ossia quello di un e-commerce BTB di un'azienda Cliente di Estilos. Si è dovuto lavorare sullo storico vendite relativo il canale online in quanto l'e-commerce non prevede la raccolta di valutazioni da parte degli utenti sui prodotti.
Inserito all'interno di un quadro più ampio gli obiettivi del progetto prevedono la rielaborazione dello storico vendite in modo d'avere i dati in forme più classiche, ossia rating discreti su una scala comune, a cui poi applicare gli approcci più popolari al momento nell'ambito dei sistemi di raccomandazione, come il collaborative filtering e il content based filtering.
Più nello specifico il task che si persegue è quello di raccomandare a ciascun cliente una lista di prodotti che si ritiene possano interessarlo.
Ci si proponeva inoltre di trovare prodotti simili e correlati dato uno di partenza, anche utilizzando informazioni esterne. Si voleva poi vagliare approcci ibridi che permettessero di combinare informazioni che descrivono l'interesse di un cliente verso un prodotto rispetto diverse prospettive, quali per esempio la quantità totale acquistata e la recentezza dell'acquisto. Data la natura dei dati si voleva infine provare ad affrontare il problema come un next basket recommendation, andando a considerare le fatture come sessioni d'acquisto e studiando se questo approccio funzionasse meglio dei precedenti.

Indice

Lista delle figure

Lista delle tabelle

1 Introduzione

1.1 Contesto progetto

Nel mondo dei software ERP *(Enterprise Resource Planning)*, ossia prodotti software pensati per le aziende che permettono la gestione e il controllo dei processi e delle funzioni aziendali, uno dei più famosi è di certo il gestionale SAP, il quale è sviluppato in moduli integrabili che, a seconda delle esigenze dell'azienda utilizzatrice, possono essere attivati in qualunque combinazione.
Uno di questi moduli è l'e-commerce hybris, utilizzato dalle aziende come canale di vendita online e alcune delle sue potenzialità sono: l'alto livello di personalizzazione e la possibilità di essere perfettamente integrato con i sistemi SAP, come per esempio con il modulo CRM (*(Customer Relationship Management)*), il quale si occupa di tutte le modalità di gestione delle relazioni con il cliente.

1.2 L'idea di progetto

L'idea nasce, in un'ottica di innovazione del prodotto, all'interno di un progetto aziendale che mira all'ampliamento e miglioramento delle funzionalità di hybris. Uno degli aspetti su cui si vuole lavorare è quello della personalizzazione dei prodotti mostrati agli utenti dell'e-commerce: si vuole quindi sperimentare rac-

comandazioni sui prodotti basate sullo storico vendite e non sui feedback lasciati dall'utente, in quanto la loro raccolta non è prevista dal sistema trattandosi di un e-commerce BTB (dove gli acquirenti sono dealer, ossia aziende che a loro volta rivendono i prodotti). Partendo quindi dallo storico vendite di un'azienda Cliente, con hybris configurato in versione BTB, l'obiettivo era quello di utilizzare i dati disponibili per raccomandare a ciascun cliente una lista di prodotti $TopN$ che gli risultassero interessanti.
Inoltre per ciascun prodotto si vuole presentare una lista di prodotti simili ad esso, sempre interessanti per il cliente a cui viene mostrato quello specifico articolo.
Come detto solitamente si parte da feedback impliciti/espliciti dati dagli utenti ai prodotti, ma non essendo disponibili si cercherà di estrarre informazioni relative l'interesse del cliente rispetto diversi punti di vista, quali può essere la quantità acquistata, la recentezza dell'acquisto, il numero di fatture in cui compare o la spesa totale per quello specifico articolo.
Una volta che le informazioni sono state organizzate in matrici grezze user-item, si voleva eseguire una sorta di preprocessing su di esse, andando a trasformarle in dei rating rispetto una scala comune che fornisse una misura d'interesse del cliente.
Sono state applicate le tecniche più popolari usate nei sistemi di raccomandazione, quali il collaborative filtering alle rating matrix ottenute dal preprocessing descritto precedentemente e il content-based filtering ai dati descrittivi dei prodotti.
Data la non disponibilità di rating si è poi pensato di considerare il problema anche come una next basket recommendation, dove si vanno a considerare le sessioni d'acquisto e in base a queste si predice quella finale, questo approccio potrebbe funzionare nel caso in sui i clienti acquistino spesso gli stessi prodotti.

1.3 Organizzazione del testo

Di seguito viene riportata per ogni capitolo una piccola descrizione delle tematiche trattate:

- **Capitolo 2**: organizzazione dei dati, come sono stati trattati e quali informazioni si sono potute ricavare;
- **Capitolo 3**: breve riepilogo della teoria sui sistemi di raccomandazione, spiegando meglio gli approcci del collaborative filtering e del content based

filtering, oltre che descrivendo il funzionamento degli algoritmi utilizzati e delle metriche;

- **Capitolo 4**: le diverse tecniche di preprocessing utilizzate per trasformare i dati grezzi in valutazioni;
- **Capitolo 5**: le modalità con cui sono stati combinati i dati;
- **Capitolo 6**: una descrizione della libreria Cornac, dove sono implementati modelli e metriche per l'esecuzione di test;
- **Capitolo 7**: i risultati delle metriche rispetto i diversi algoritmi applicati al preprocessing dei dati;
- **Capitolo 8**: i risultati delle metriche rispetto i diversi algoritmi applicati al preprocessing dei dati nella loro versione combinata;
- **Capitolo 9**: i risultati delle metriche considerando il problema come un next basket recommendation;
- **Capitolo 10**: le conclusioni del lavoro svolto, andando a delineare problemi risolti, criticità e sviluppi per il futuro.

1.4 Convenzioni tipografiche

Il testo adotta le seguenti convenzioni tipografiche:

- ogni acronimo, abbreviazione, parola ambigua o tecnica viene spiegata e chiarificata alla fine del testo;
- ogni parola di glossario alla prima apparizione verrà etichetta come segue $parola^{[g]}$.

2
Analisi dei dati

2.1 Principali tabelle

Lo storico vendite dell'azienda Cliente è stato estratto dal modulo hybris, questo è organizzato secondo le tabelle SAP, avremo quindi lo storico delle fatture, composto da una tabella per la testata della fattura, ossia la parte descrittiva dove viene riportato l'acquirente, e una tabella per le posizioni della fattura, ossia la parte dove vengono riportati i materiali acquistati.
Abbiamo inoltre due tabelle che contengono rispettivamente l'anagrafica cliente e materiali, dove possiamo trovare informazioni aggiuntive che li descrivono.
Mi è stato inoltre fornito un glossario che riportava per ciascuna tabella una breve spiegazione di ogni campo.
Tutte queste tabelle sono in formato Excel.
Quindi ricapitolando le principali tabelle disponibili sono le seguenti:

- **VBAK**: testata della fattura;
- **VBAP**: posizioni della fattura;
- **MARA**: anagrafica prodotto;
- **KNA1**: anagrafica materiali.

Andiamo ora a vedere per ciascuna di queste tabelle i campi annessi e alcune informazioni di natura statistica.

2.1.1 Tabella VBAK

La tabella VBAK contiene la testata di circa 35000 fatture, datate dall'anno 2016 fino a maggio 2021.
Ciascuna riga della tabella è la testata di una fattura e ne riporta il suo codice identificativo (**VBELN**) insieme con il codice del cliente a cui è associata (**KUNNR**). Inoltre ciascuna fattura riporta data e ora (**ERDAT**, **ERZET**) in cui è stata emessa, l'importo totale e la valuta corrispondente (**NETWR**, **WAERK**).

2.1.2 Tabella VBAP

La tabella VBAP contiene le posizioni delle fatture (circa 250000), riporta per ognuna di esse la lista di prodotti acquistati indicando diverse informazioni. Ciascuna riga della tabella riporta quindi il codice identificativo (**VBELN**) della fattura e il codice identificativo del prodotto acquistato (**MATNR**), poi vengono riportati per quel prodotto il prezzo unitario (**NETPR**), la quantità acquistata (**KWMENG**), la spesa totale con la valuta (**NETWR**, **WAERK**) e il codice gerarchia prodotto storico (**PRODH**), ossia quello salvato in MARA al momento del'emissione della fattura.

2.1.3 Tabella KNA1

La tabella KNA1 riporta l'anagrafica cliente (circa 3000), per ciascuna riga abbiamo il codice cliente (**KUNNR**), il codice paese d'origine (**LAND1**), il nome dell'azienda (**NAME1**), la località (**ORT01**) e la regione (**REGIO**).

2.1.4 Tabella MARA

Come detto la tabella MARA è quella che riporta l'anagrafica dei materiali, nella nostra futura trattazione considereremo questi materiali come prodotti in quanto sono acquistabili all'interno dell'e-commerce.
Ciascuna riga della tabella riporta quindi un prodotto univoco composto dal

suo usuale codice identificativo (**MATNR**), dal codice della gerarchia prodotto (**PRODH**) e del gruppo merceologico (**MATKL**) e una breve descrizione testuale (**MAKTX**), poi abbiamo delle informazioni su dimensione, volume e peso.
Per le dimensioni abbiamo un campo (**GROES**) che le fornisce nel formato lunghezza X larghezza X altezza, oppure altri (**LAENG**, **BREIT**, **HOEHE**), che indicano rispettivamente lunghezza, larghezza e altezza e l'unità di misura per entrambi i formati viene riportata nello stesso campo (**MEABM**).
Poi abbiamo due campi per volume e peso (**VOLUM**, **NTGEW**) e altri due per le loro rispettive unità di misura (**VOLEH**, **GEEWI**).

2.2 Prodotti

Da questo momento in poi faremo riferimento ai materiali chiamandoli prodotti come detto in precedenza.
In totale nella tabella anagrafica materiali (MARA) sono presenti circa 75000 prodotti diversi, mentre i prodotti effettivamente venduti risultano essere molti meno attestandosi all'incirca verso gli 8000.
Abbiamo però due campi interessanti che riguardano la gerarchia prodotto (PRODH) e il gruppo merceologico (MATKL), questi due ci permettono di studiare la similarità dei prodotti.

2.2.1 Gerarchia prodotto (PRODH)

Il campo gerarchia prodotto (PRODH) è un campo numerico di 18 cifre utile per separare i prodotti rispetto le diverse categorie su più livelli. Nella tabella secondaria T179 vengono definiti i livelli di gerarchia e le diverse categorie. Vediamoli di seguito:

- **PRODH**: codice gerarchia prodotto;
- **STUFE**: livello gerarchia;
- **VTEXT**: descrizione testuale.

Ciascun codice PRODH contenuto nella tabella T179 avrà rispettivamente il seguente numero di cifre in base al livello di gerarchia (STUFE):

- **STUFE = 1**: 1° livello della gerarchia, il codice sarà di 5 cifre;
- **STUFE = 2**: 2° livello della gerarchia, il codice sarà di 10 cifre, dove le prime 5 identificano la categoria di 1° livello a cui appartengono mentre le restanti 5 indentificano la sotto-categoria di 2° livello;
- **STUFE = 3**: 3° livello della gerarchia, il codice sarà di 18 cifre, dove le prime 10 identificano la categoria di 2° livello a cui appartengono mentre le restanti 8 indentificano la sotto-categoria di 3° livello.

Ciascun prodotto sarà quindi provvisto di un codice di 18 cifre che identificherà una categoria per ogni livello.
Nella tabella VBAP ci sono alcune posizioni dove a parità di codice prodotto (MATNR) si hanno codici PRODH diversi, questo è dovuto al diverso momento temporale in cui sono stati acquistati, infatti nella tabella VBAP il codice PRODH è storico, ho provveduto per semplicità ad aggiornarli tutti al codice PRODH più recente riportato nella tabella anagrafica materiali MARA. Il numero di prodotti interessati sono circa 100 su 10000 posizioni.

Selezione categorie di 1° livello

PRODH	*#tot*	*#sold*	titolo
00010	28	0	categoria1
00020	0	0	categoria2
00030	0	0	categoria3
00040	5	0	categoria4
00050	2	0	categoria5
00090	2	0	categoria6
00100	1117	173	CATEGORIA1
00200	645	130	CATEGORIA2
00250	31	11	CATEGORIA3
00300	405	92	CATEGORIA4
00400	525	36	CATEGORIA5
00500	1715	70	CATEGORIA6
00600	334	6	CATEGORIA7
00700	1	0	CATEGORIA8
00900	70441	7702	CATEGORIA9
00950	28	1	CATEGORIA10
09999	215	0	ALTRO

Nella tabella vengono mostrati i codici PRODH delle categorie di 1° livello, nella colonna *#tot* il numero di prodotti diversi per quella categoria, nella colonna *#sold* il numero di prodotti diversi acquistati almeno una volta appartententi a quella categoria ed infine il titolo della categoria. Come possiamo vedere le prime sei categorie con titolo in minuscolo hanno pochi prodotti catalogati in MARA e nessun prodotto venduto.

Chiaramento il fatto che siano tutti a zero è dovuto all'aggiornamento dei codici PRODH di cui abbiamo parlato precedentemente, a prescindere da ciò il numero di posizioni che prima riportavano codici appartenenti alle categorie prese in considerazione non superava la decina, quindi non considerare queste categorie in quanto si è smesso di usarle sembra la scelta più logica.
La categoria ALTRO (09999) non è stata considerata in quanto riporta prodotti

che non sono disponibili sull'e-commerce. Inoltre le categorie CATEGORIA8 (00700) e CATEGORIA10 (00950), dato il basso numero di prodotti presenti in MARA e le basse vendite, si è preferito non considerarle.

Overview categorie di 1° livello

PRODH	*#posizioni*	$\sum_{KWMENG}$	$\mathbb{E}_{KWMENG}$	$\mathbb{E}_{NETPR}$ (€)	$\sum_{NETWR}$ (€)	$\mathbb{E}_{NETWR}$ (€)	*titolo*
00100	5908	15461	2.61	1613.44	3865.97	22840167.61	CATEGORIA1
00200	1936	3219	1.66	5898.09	8640.59	16745496.87	CATEGORIA2
00250	333	2949	8.85	552.99	3772.04	1377422.72	CATEGORIA3
00300	745	1390	1.86	4353.24	5364.34	3996430.94	CATEGORIA4
00400	389	1651	4.24	708.17	2404.47	940777.84	CATEGORIA5
00500	1133	12984	11.46	175.75	1034.63	1172236.58	CATEGORIA6
00600	153	494	3.23	448.52	1338.80	83501.04	CATEGORIA7
00900	239740	1070334	4.46	26.67	66.61	15968039.56	CATEGORIA9
	250339	1108493.51	4.72	1706.86	3225.75	63124073.16	valori riassuntivi

Nella tabella per ogni categoria $PRODH$ di 1° livello possiamo vedere:

- $\#posizioni$: numero di posizioni in cui compaiono prodotti di quella categoria in fattura;
- $\sum_{KWMENG}$: quantità totale di prodotti acquistati appartenenti a quella categoria;
- $\mathbb{E}_{KWMENG}$: quantità media per fattura di prodotti acquistati appartenti a quella categoria;
- $\mathbb{E}_{NETPR}$: prezzo medio per fattura di prodotti acquistati di quella categoria;
- $\sum_{NETWR}$: spesa totale per prodotti di quella categoria;
- $\mathbb{E}_{NETWR}$: spesa totale media per fattura di prodotti di quella categoria.

Dalla tabella possiamo vedere come la categoria RICAMBI & ACCESSORI riporti un prezzo medio per fattura molto più basso rispetto alle altre categorie, questo è dovuto al fatto che i pezzi di ricambio ed accessori non sono macchine o sistemi da usare per fornire un servizio quanto un prodotto per riparare ciò che già si possiede. Possiamo vedere che in termini di posizioni l'acquisto di pezzi di ricambio copra una cospicua parte delle posizioni in fattura, oltre che costituire un'importante parte del fatturato per l'azienda. Le altre categorie indicano macchine e sistemi per la pulizia quindi il prezzo medio per prodotto è molto maggiori

e per i clienti finali questi rappresentano un investimento.
Da quanto detto finora si vengono a creare due macro categorie di prodotti:

- **Macchine**: questa macro categoria racchiude sette categorie di 1° livello (00100, 00200, 00250, 00300, 00400, 00500, 00600);
- **Ricambi**: questa macro categoria invece racchiude la sola categoria 00900.

Dobbiamo dare un'ultima precisazione, la maggior parte delle categorie di 2° e 3° livello risultano essere di prodotti appartenenti alla macro categoria delle macchine, quindi la gerarchia è molto più densa orizzontalmente per le macchine rispetto che i pezzi di ricambio, infatti per le macchine le categorie risultano essere i diversi modelli di macchinari disponibili e i prodotti a catalogo di quella categoria sono le varianti dello stesso modello di macchinario. Per i pezzi di ricambio abbiamo solo poche categorie contenitore che li raggruppano tutti insieme.

2.2.2 Gruppo merceologico (MATKL)

Il gruppo merceologico (MATKL) non è organizzato come una gerarchia, come lo è invece la gerarchia prodotto (PRODH), bensì come un insieme di prodotti, in totale abbiamo circa 160 gruppi, dove uno di questi contiene tutti i prodotti che prima abbiamo classificato come macchine. Rispetto il codice PRODH, il gruppo merceologico è più divisivo rispetto i ricambi, questo ci può aiutare in quanto ora siamo in grado di categorizzare anche i ricambi.

2.2.3 Dimensione, volume e peso

I campi rigurdanti dimensione, volume e peso potrebbero essere utili per ricercare una similarità tra i prodotti.
Le informazioni sulle dimensioni, come lunghezza, larghezza e altezza sono praticamente ridondanti nei campi GROES e (LAENG, BREIT, HOEHE) se non per alcuni prodotti dove le informazioni sono esclusive di uno dei due formati.
Per peso e volume abbiamo i rispettivi campi numerici e altri due campi che riportano le unità di misura, per il volume possono essere i metri cubi o i millimetri cubi, per il peso i kilogrammi o i grammi. La criticità di queste misure riguarda la loro scarsità, infatti su 75000 prodotti abbiamo informazioni su volume e peso

rispettivamente solo sul 20% e 39%, mentre sui prodotti acquistati almeno una volta sul 19% e 5%.

3

Sistemi di raccomandazione

3.1 Introduzione

Uno dei campi più popolari al momento verso cui si rivolge una particolare attenzione è quello dei sistemi di raccomandazione, da ora in poi RS, in quanto l'attività online sta aumentando sempre più e nascono sempre più spesso nuovi servizi che permettono di scegliere oggetti, siano questi prodotti, video, musica, film o molto altro, da cataloghi vastissimi. I sistemi di raccomandazione permettono di navigare questi cataloghi andando a cercare gli oggetti che risultino più interessanti per l'utente.

3.2 Preliminari

In generale possiamo dire che un RS si compone di diversi elementi, in primo luogo abbiamo i cosidetti "attori" del problema, gli user e gli item, rispettivamente gli utenti del sistema e gli oggetti che si vuole consigliare. Abbiamo a disposizione inoltre informazioni riguardo l'interazione tra user e item solitamente sotto forma di feedback implicito o esplicito, questa misura viene definita rating. Questi vengono utilizzati dal RS, insieme con eventuali dati legati al contesto di user e item, per effettuare raccomandazioni.

3.2.1 Feedback impliciti / espliciti

Solitamente le informazioni che legano user e item, ossia i rating, possono essere di due tipi:

- Implicito: 1 se c'è stata interazione tra lo user e l'item, 0 se non c'è stata;
- Esplicito: valutazione numerica intera in una scala da 1 a N, 0 se non c'è stata interazione.

Nel nostro caso di studio però ci ritroviamo a metà strada in quanto, se per esempio considerassimo la quantità come un dato esplicito ci troveremmo così ad avere un dato su una scala non continua, mentre se lo facessimo come se fosse implicito trascureremmo delle informazioni che possono in qualche modo fornire una misura di interesse.

3.2.2 User-item interaction matrix

		Items					
		1	*2*	...	*i*	...	*m*
Users	*1*	5	3		1	2	
	2		2				4
	:			5			
	u	3	4		2	1	
	:					4	
	n			3	2		

I rating sono organizzati in matrici, dette appunto user-item interaction matrix o semplicemente matrici dei rating (R), dove sulle righe abbiamo gli user mentre sulle colonne abbiamo gli item, nell'incrocio abbiamo riportato il rating. La matrice come detto può essere implicita o esplicita e le celle vuote corrispondono allo 0. Quando scriviamo r_{ui} intendiamo che u è lo user e i è l'item.

3.2.3 Task

L'obiettivo del sistema può essere quello di consigliare ad uno user una lista di N item, detta *TopN* che si ritiene possano interessargli, oppure dato un item si può trovare una lista di item che si considerino simili allo stesso in acccordo con i 'gusti' dello user.

3.3 Approcci

Definito quindi il task abbiamo diversi modi per poter soddisfare il nostro obiettivo, in generale abbiamo due principali categorie di RS:

- **Non Personalizzato**: andiamo a consigliare i prodotti che globalmente risultano più popolari, ossia che abbiano complessivamente ricevuto più valutazioni, o quelli con rating più alto. Questo approccio non va a considerare le informazioni relative il singolo user;
- **Personalizzato**: ci sono diversi approcci che vedremo nelle sezioni successive, in generale si fanno raccomandazioni basate sulla similarità tra user. I due approcci più famosi sono il collaborative filtering, dove si cerca di consigliare item ad uno user basandosi su user simili, e il content-based filtering si cerca di raccomandare item simili a quelli con cui si ha già interagito.

Nelle sezione successive andiamo a spiegare più nel dettaglio il collaborative e conte-based filtering.

3.3.1 Collaborative filtering

Il collaborative filtering è un approccio agli RS basato sulla similarità, raccomandiamo ad uno user item interessanti per altri user simili ad esso, e viceversa item simili ad altri item per cui ha dimostrato interesse. La similarità può essere quindi di due tipi: item-based, basata quindi sulla similarità tra prodotti o user-based ossia su quella tra user. Ci sono due approcci possibili al collaborative filtering:

- Memory-based: utilizziamo la matrice dei rating per calcolare la similarità tra user e item, metodi basati sull'algoritmo K nearest neighbour;
- Model-based: utilizziamo dei modelli che attraverso degli algoritmi permettono di predire il rating su item non valutati.

UserKnn

UserKnn è un metodo memory-based che fa uso della matrice dei rating, ogni user avrà quindi un proprio "profilo", ossia la propria riga nella matrice dei rating.

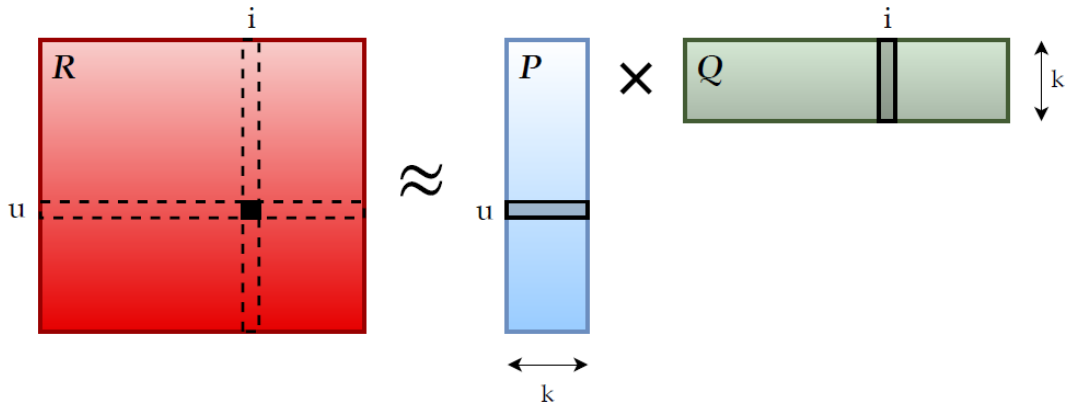
L'idea è quella di calcolare la similarità tra tutti gli user e fatta questa operazione è possibile calcolare il rating previsto per ogni item non valutato rispetto ad uno user. Per fare ciò andiamo a selezionare i k user con similarità più alta con il nostro user target e calcoliamo la media pesata dei loro rating usando come pesi la similarità.
Fatto questo si procede ad ordinare per ciascuno user tutti i prodotti secondo i rating ottenuti e si ottiene così la lista $TopN$ degli item più interessanti.
Questo metodo funziona senza informazioni relative alle caratteristiche degli user o item e può gestire rating espliciti o impliciti con formule leggermente diverse.

ItemKnn

ItemKnn è un metodo memory-based molto simile al precedente, qui si va a considerare però gli item da raccomandare e il loro "profilo" è la propria colonna della rating matrix. Si calcola la similarità tra tutti gli item e dato uno user si procede a calcolare il rating stimato sugli item che non ha valutato andando a trovare per ciascuno di essi la lista di K item che ha valutato più simili ad esso, poi calcola la media pesata dei rating dei K item selezionati usando come peso la similarità.

Matrix Factorization (MF)

Nell'approccio model-based, Matrix Factorization (MF) è uno dei modelli più famosi, questo si basa sul concetto che si possa mappare user e item verso uno spazio delle feature comune di una certa dimensionalità K, possiamo quindi creare due matrici P e Q, dove P è una matrice avente sulle righe gli user e come colonne le K feature, mentre Q è una matrice avente sulle righe le k feature e sulle colonne gli item, quello che vogliamo ottenere è una approsimazione della rating matrix attraverso la moltiplicazione di P e Q, ossia $R \approx P \cdot Q^T = \hat{R}$.



Quello che otteniamo è quindi un profilo sia per gli user che per gli item rispetto lo stesso spazio delle feature. Il problema maggiore risulta però essere quello di ottenere queste due matrici, possiamo farlo creando una funzione di loss apposita e andando ad allenare in modo alternato le matrici P e Q, carcando quindi di ridurre dopo ciascuna iterazione la differenza tra R e $\hat{R}$.
Una volta che il modello è allenato possiamo predire il rating dato da uno user u su un item i moltiplicando i profili corrispondenti $\hat{r}_{ui} = P_u \cdot Q_i^T$.

Variational Auto-Encoder for CF (VAECF)

Rispetto ai modelli precedenti che potevano funzionare sia con rating espliciti o impliciti, il VAECF accetta solo quest'ultimi. Il modello prende in input la matrice dei rating impliciti andando a considerare ogni riga come una distribuzione associata allo user. **Da completare**.

3.3.2 Content-based filtering

Il content-based filtering è un approccio che si basa sull'idea di consigliare item simili a quelli con cui si è già interagito.
Ciascun item possiede delle feature, per esempio nel caso si abbia come item dei film queste potrebbero essere i generi, l'insieme delle feature può essere mappata su di un vettore delle feature, per ogni item quindi si riporta nel vettore 1 se possiede quella feature, 0 altrimenti, questo permette di definire un profilo per l'item. Una volta fatto ciò si procede a creare anche un profilo per lo user, ci sono diversi modi per farlo ma per esempio si può considerare tutti i profili degli item con cui si è interagito e farne quindi la media pesata basata sui rating

corrispondenti. Ottenuto un profilo anche per lo user si può calcolare la similarità tra di esso e quello degli item per poi ordinarli secondo similarità ottenendo così la lista $TopN$.

3.4 Valutazione

Quanto abbiamo visto finora sono metodi che ci permettono di effettuare le raccomandazioni, vogliamo trovare però anche il modo per poterle valutare. Per prima cosa dobbiamo dividere le matrici dei rating in training e test set.
Per fare ciò andiamo ad eseguire uno shuffle delle coppie (user, item) e si va ad assegnare al training set l'80% delle coppie e le restanti al test set, questo sistema non ci assicura che uno user sia presente nel training set. Date le raccomandazioni fornite dal RS vogliamo valutarle rispetto due aspetti principali: il rating e il ranking. Il primo aspetto riguarda semplicemente la diversità tra i rating stimati da quelli reali delle coppie (user, item) del test set, queste metriche non vengono solitamente usate in quanto non è un buon modo per valutare un RS perché non ci permette di capire se un'item consigliato sia rilevante. Andando invece a considerare il concetto di ranking ci rendiamo conto che sia più legato a ciò che vogliamo andare a valutare, le metriche annesse considerano rilevanti gli item presenti nel test set e vanno a verificarne la posizione nella lista $TopN$. In generale le metriche si applicano a ciascuno user e il risultato finale è la media dei singoli risultati.

3.4.1 AUC

L'AUC (Area Under the Curve) è una metrica che permette di valutare un RS basandosi sul numero di coppie di item rivelanti e non, presenti nella $TopN$ in ordine, dove un item non rilevante ha uno score minore rispetto a quello di uno rilevante, vediamo di seguito la formula:

$$AUC = \frac{1}{|N^+| \cdot |N^-|} \sum_{i \in N^+} \sum_{j \in N^-} [s(i) > s(j)]$$

Il termine N^+ è l'insieme degli item presenti nel test set, mentre N^- sono tutti gli item rimanenti. $s(i)$ è la valutazione data dal RS sull'item i, quello che si fa

è andare a contare il numero di coppie di item in ordine nella $TopN$ andando a vedere gli score assegnati loro. Utilizziamo la funzione $[s(i) > s(j)]$ che restituisce 1 se lo score dell'item rilevante è maggiore, 0 altrimenti. Infine dividiamo il numero di coppie ordinate in modo correto per il numero di coppie totali.
Il valore dell'$AUC \in [0, 1]$, dove più il valore si avvicina ad 1, minori saranno le coppie in ordine sbagliato.

3.4.2 nDCG@k

La metrica nDCG@k (normalized Discount Cumulative Gain) ci permette di calcolare una misura basata sulla posizione degli item rilevanti nella lista $TopN$, ossia quelli del test set.

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{log_2(i+1)} \qquad IDCG@k = \sum_{i=1}^{|REL_i|} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

Ed infine per cacolare la $nDCG@k$ usiamo la seguente formula:

$$nDCG@k = \frac{DCG@k}{IDCG@k}$$

Per prima cosa dobbiamo dire che la metrica lavora su una parte della lista $TopN$, ossia la parte contentente i primi K item che definiamo come $TopK$.
Gli item più rilevanti nel nostro caso sono gli item presenti nel test set.
Andiamo ora ad analizzare numeratore e denomicatore per capire meglio la loro funzione.

- La $DCG@k$ si basa sull'idea che gli item più rilevanti debbano trovarsi il più possibile in testa alla lista $TopK$, quindi si vuole penalizzare un item sempre di più via via questo si trovi nella coda della lista, per far ciò il denominatore aumenta all'aumentare della posizione seguendo una scala logaritmica.
- La $IDCG@k$ è equivalente a calcolare la $DCG@k$ sulla lista $TopK$ ideale, ossia la lista riportante tutti gli item ordinati in modo ideale secondo score nella posizione corretta, questo equivale al valore massimo ottenibile dalla metrica.

Quindi andando a dividere il numeratore con il denominatore si attua una normalizzazione, il valore finale di $nDCG@k \in [0, 1]$, dove più si avvicina ad 1 più

la lista $TopK$ assomiglia a quella ideale.

4

Preprocessing storico vendite

In questo capitolo andremo a vedere diverse tecniche che sono state utilizzate per trasformare le matrici grezze user-item in matrici dei rating.
Nei capitoli legati ai risultati andremo a vedere pro e contro di queste tecniche.

4.1 Preliminari

Definiamo l'insieme degli user U, l'insieme degli item I e le matrici grezze user-item RG. Ciascuna tecnica lavora andando a considerare le matrici RG come un vettore di triplette $V = [(u, i, RG_{(u,i)} \neq 0)|\forall(u \in U, i \in I)]$ con $RG_{(u,i)} \in \mathbb{R}$.
Facciamo inoltre riferimento a V_c come il vettore delle coppie (user,item), $V_{(u,i)}$ come il valore della tripletta di user u e item i, V_u il vettore dei valori delle triplette con user u e V_i il vettore dei valori delle triplette con item i.
Ciascuna tecnica implementa una diversa funzione f biettiva di trasformazione che possiamo riassumere come segue:

$$f : [(u, i, V_{(u,i)})|\forall(u, i) \in V_c] \rightarrow [(u, i, r \in [1, scale]|\forall(u, i) \in V_c]$$

Queste tecniche si propongono di trasformare il valore $V_{(u,i)}$ di ciascuna tripletta in un rating $r \in [1, scale]$, con $scale$ sempre dispari.

Alcune tecniche faranno riferimento ad una distribuzione dei rating uniforme discreta o gaussian-like su gruppi di elementi. Quando ci si troverà ad applicare queste distribuzioni avremo un vettore di elementi ordinati secondo un certo criterio. Vediamo come vengono assegnati i rating secondo queste due distribuzioni:

- uniforme discreta: divide il vettore in modo tale che ogni valore nella scala dei rating compaia lo stesso numero di volte, assegnandoli in modo crescente, dal capo alla coda del vettore;
- gaussian-like: si va a definire una distribuzione normale $N(0, scale/3)$, poi si generano una quantità sufficiente di numeri secondo la suddetta distribuzione. Fatto questo si convertono tutti i numeri decimali in interi, si selezionano solo gli interi nell'intervallo $[-scale/2, scale/2]$ e si traslano nell'intervallo $[1, scale]$.
  Infine calcoliamo la probabilità per ciascun numero intero nella scala.
  Per assegnare i rating al vettore non si fa altro che iterare sugli interi dell'intervallo $[1, scale]$, andando ad eseguire in sequenza le seguenti operazioni:

  1. moltiplico la probabilità di quell'intero per la lunghezza del vettore;
  2. converto il valore risultante ad intero, ottenendo quindi il numero di elementi che dovranno avere quel rating;
  3. partendo dall'inizio del vettore assegno quel rating a quello specifico numero di elementi e poi una volta raggiunto l'ultimo procedo col successivo intero della scala a partire dall'elemento seguente.

L'assegnazione dei rating secondo tali distribuzioni è implementato da due funzioni che restituiscono un vettore di coppie, formate dall'elemento e dal rating corrispondente.

4.2 Tecnica product-based

La tecnica *prodotto globale* prevede di andare a considerare gli item da un pusto di vista globale. Si procede andando a considerare gli item in termini assoluti, vediamo di seguito le operazioni per applicarlo:

1. otteniamo il seguente vettore $[(i, \sum V_i) \forall i \in I]$;

2. ordiniamo il vettore ottenuto basandoci sul secondo termine e conserviamo solo il vettore degli item ordinati;

3. andiamo ad applicare la funzione uniforme discreta / gaussian-like a tale vettore, ottenendo per ogni item un rating;

4. per ogni tripletta di partenza (user,item,_) andiamo ad assegnare il rating usato per quello specifico item.

Questa tecnica porta ad avere a dispetto dello user la stessa valutazione per ogni item ed è quindi molto sensibile alla popolarità di un'item nello storico vendite.

4.3 Tecniche group-based

Le tecniche presenti in questa sezione permettono di dividere il vettore delle triplette V in diversi gruppi, applicare separatamente a ciascuno di essi il metodo ed infine unire insieme i vettori risultati. Deve essere rispettata la condizione che l'intersezione tra tutti i gruppi deve essere nulla.
Vediamo le possibili divisioni in gruppi delle triplette di volta in volta:

- un unico gruppo con tutte le triplette;
- un gruppo per ogni user contenente solo le sue triplette;
- per ogni user e per ogni categoria un gruppo contente tutte le triplette di quello user con l'item che appartiene a quella categoria;

Vediamo ora i diversi metodi applicati ad un singolo gruppo.

4.3.1 Normalizzazione Min-Max

Una delle tecniche che viene proposta nella letteratura è quella della normalizzazione min-max, per applicarla andiamo a considerare un gruppo $G \subseteq V$ e applichiamo a ciascuna tripletta la seguente funzione:

$$[(u, i, \frac{G_{(u,i)} - min(G_r)}{max(G_r) - min(G_r)} \in [0, 1])|\forall(u, i) \in G_{(u,i)})]$$

Ora tutti i valori delle triplette di G si troveranno in un intervallo $[0, 1]$, per portarlo invece nell'intervallo $[1, scale]$ dobbiamo applicare la seguente formula:

$$[(u, i, (scale - 1) \cdot \frac{G_{(u,i)} - min(G_r)}{max(G_r) - min(G_r)} + 1 \in [1, scale])|\forall(u, i) \in G_{(u,i)})]$$

Inoltre una volta applicata la formula, oltre che tenere i rating così come sono nel dominio dei numeri reali, si è provato anche a convertirli in numeri interi, verranno chiamate rispettivamente *continous* e *rint*.
Si voleva provare in questo modo a capire intanto se gli user avessero volumi d'acquisto diversi e se prodotti delle stesse coppie avessero logiche d'acquisto simili. Inoltre dobbiamo puntualizzare che se guardiamo per esempio la distribuzione della quantità totale rispetto i prodotti, noteremo che risulta assumere il comportamento di una curva discendente, quindi ci sono molti prodotti acquistati in bassa quantità e pochi in grande quantità. Applicando questo metodo, che non va a cambiare la distribuzione iniziale dei valori ma va solo a scalarli, otterremo quindi molti rating bassi.

4.3.2 Tecnica ordered-based

Il seguente metodo prevede di lavorare su un gruppo di triplette $G \subseteq V$ e di eseguire le seguenti operazioni:

1. ordiniamo il vettore G secondo valore;
2. andiamo ad applicare la funzione uniforme discreta / gaussian-like a tale vettore;
3. andiamo a sostituire al valore della tripletta quello del rating assegnatogli.

Questa tecnica permette di andare a confrontare le triplette attraverso l'ordinamento, permette una migliore distribuzione dei rating rispetto la normalizzazione min-max, ma è da verificare se questa ci fornisca risultati sperimentalmente migliori.

4.4 Approccio implicito

Tutti gli approcci che abbiamo visto producono matrici dei rating esplicite, chiaramente un tentativo sarà quello di usare una versione della matrice grezza implicita.

5

Tecniche combinate

Nel capitolo precedente abbiamo visto diverse tecniche di preprocessing per ottenere dei rating dalle matrici grezze. In questo capitolo andremo invece a vedere due approcci che si sono tentati per cercare di combinare insieme rating provenienti da fonti diverse.

5.1 Premesse

Come riportato nel capito dell'analisi dei dati, le informazioni disponibili sugli item ci permettono di valutare l'interesse dello user verso gli item secondo diversi *aspetti*, quali la quantità acquistata, la spesa totale, il numero di fatture in cui compaiono e la recentezza dell'ultimo acquisto, definiremo questi aspetti da ora in poi come *espressioni di interesse*. Questi *aspetti* sono organizzati in matrici grezze a cui nel capitolo precedente abbiamo applicato diverse tecniche di preprocessing andando a trasformali in rating, da qui cercare di unire insieme queste espressioni di interesse sembra essere un buon modo per migliorare la qualità delle raccomandazioni finali. I metodi combinati prendono in input le matrici grezze e vi applicano una tecnica di preprocessing del capitolo precedente.
Fatto questo ci sono due modi per combinarle insieme, vediamoli di seguito:

5.2 Combinazione liste *TopN*

Il primo metodo si propone di ottenere per ogni user una lista $TopN$ di item per ciascuna espressione di interesse, queste poi andranno combinate insieme attraverso l'uso del borda count, un sistema di voting basato sulla posizione. Vediamo ora quali sono le operazioni da attuare:

1. applicare la stessa tecnica di preprocessing a tutte le matrici grezze delle espressioni di interesse ottenendo le corrispettive matrici dei rating;
2. applicare uno degli approcci del collaborating filtering alle matrici dei rating ottenendo così le liste $TopN$;
3. combinare insieme le liste $TopN$ secondo un sistema di voting, quale il borda count, ogni item nella lista riceve uno score in base alla posizione, si sommano gli score di ciascun item e li si riordina in base a questi.

5.3 Media matrici dei rating

Mentre il precedente metodo prevedeva di applicare il collaborative filtering separatamente a ciascuna matrice dei rating, in questo andiamo ad effettuare una loro media ottenendo così una sola matrice dei rating.
A questa andiamo poi ad applicare uno degli approcci del collaborative filtering e otteniamo così la lista $TopN$.

6

LIbreria cornac e esperimenti

In questo capitolo andremo a vedere le modalità con cui sono stati svolti gli esperimenti e la libreria utilizzata per gli stessi.

6.1 Dataset

Le informazioni dello storico vendite sono state organizzate basandosi sulle due macrocategorie individuate nell'analisi: macchine e ricambi. Gli item considerati sono quelli acquistati almeno una volta mentre per i clienti quelli che hanno effettuato almeno un'acquisto. Abbiamo quindi tre tipi matrici grezze user-item contenenti solo item appartenenti alle macrocategoria delle macchine, solo dei ricambi ed infine una con tutti gli item detta totale. In generale le divisione rispetto i tipi di dataset sono le seguenti:

- **Macchine**: 254 user e 518 item;
- **Ricambi**: 319 user e 7699 item;
- **Totale**: 322 user e 8217 item;

Per ognuno di questi tipi di matrice abbiamo le quattro versioni delle espressioni di interesse, che sono state calcolate tra uno user u e un item i come segue:

- **Quantità**: somma dei campi quantità (*KWMENG*) di tutte le posizione delle fatture di u in cui è presente i.
- **Spesa totale**: somma dei campi spesa totale (*NETWR*) di tutte le posizione delle fatture di u in cui è presente i.
- **Numero di fatture**: conta del numero di fatture di u in cui compare i.
- **Recentezza**: ricerca della posizione di i nelle fatture di u, riportante la data più recente di acquisto. Se l'item è stato acquistato avremo quindi una data a cui andremo a sottrarre la data della fattura più vecchia, il delta temporale viene trasformato in giorni. Quindi più è alto il delta in giorni più recente sarà l'acquisto.

6.2 Libreria Cornac

La libreria cornac gestisce completamente gli esperimenti, dall'acquisizione dei dati fino alla verifica dei risultati. Nello specifico è stata scelta per la presenza di molti *modelli* e metriche per la loro valutazione.
Nello specifico i modelli utilizzati sono stati:

- MostPop: modello basato sulla popolarità, dove un item è più popolare in base al numero di user che lo hanno valutato, usato per ottenere un risultato di base;
- UserKnn: implementazione dell'approccio del collaborative filtering memory-based, che di default prende in considerazione i 20 user più simili ad uno target.
- ItemKnn: come il precedente, ma basato sugli item. Non è stato infine utilizzato in quanto durante le fasi iniziali di test non ha mai riportato risultati superiori a quelli di base e richiedeva troppo tempo per la valutazione;
- MF: implementazione del matrix factorization, metodo model-based del collaborative filtering, di default considera 10 user più simili ad uno target;
- VAECF: modello per la versione implicita, usato per avere un risultato di base.

I dati presi in input sono nel formato *user, item, rating* e volendo potevano essere divisi in training, validation e test set direttamente dalla libreria. Si è preferito però dividerli esternamente con le rispettive percentuali: 70% al training set, 15% per validation e test set. Nella fase di valutazione dei risultati dei *modelli* veniva richiesto un parametro detto *rating_threshold*, il quale serviva a binarizzare gli item rilevanti e irrilevanti del test set basandosi sul rating corrispondente. Non andava in alcun modo ad intaccare i rating nella fare di training.

6.3 Esperimenti sulle singole matrici dei rating

Come riportato nel capitolo relativo il preprocessing delle matrici grezze, non è stata definita una scala dei rating in quanto se ne volevano provare diverse: 3, 5, 7, 9, 13, 19, 25. Inoltre per ogni tecnica dove fosse previsto l'utilizzo di una delle distribuzioni disponibili, si è proceduto a testarle entrambe. Dalle combinazioni di diversa scala e distribuzione dei rating, unito con la presenza di ben dodici matrici grezze (3 tipi di dataset per 4 espressioni di interesse) si sono ottenute circa 1500 matrici grezze. Per eseguire i test si è proceduti come segue:

1. Ottengo un risultato di base con il modello MostPop e con il VAECF;
2. Poi per ogni combinazione di matrice dei rating, scala e funzione di distribuzione se usata, si eseguono le seguenti operazioni:
   (a) *Alleno* il modello con valori di deafult con il training set;
   (b) valuto i risultati del modello sul validation set, se questi valori risultano migliori di quelli di base (dati dal MostPop e VAECF), allora si procede ad un tuning dei parametri;
   (c) Una volta concluso il tuning, si sceglie il modello migliore in accordo al validation set;
   (d) Si restituisce la valutazione finale basata su test set.

6.4 Esperimenti sulle matrici dei rating combinate

Questi esperimenti sono stati condotti andando a combinare insieme secondo le modalità indicate solo le matrici dei rating generate con lo stessa tecnica e sulla stessa scala.

7

Risultati esperimenti singole matrici dei rating

Glossario

Lista degli acronimi