Slovenská technická univerzita v Bratislave
Fakulta informatiky a informačných technológií
Ilkovičova 2, 842 16 Bratislava 4

Odporúčací systém založený na LSTM neurónových sieťach

Viktória Markovičová, Ondrej Unger

Predmet: Neurónové siete
Cvičiaci: Ing. Michal Farkaš

Obsah

1 Motivácia

Odporúčacie systémy sa v súčasnosti tešia veľkej popularite. S nárastom informácií na internete vznikol problém s vyhľadávaním a vybraním si informácií relevantných pre nás. Práve tento problém nám majú odporúčacie systémy pomôcť riešiť. Ich hlavnou úlohou je odporučiť používateľovi taký produkt, ktorý bude preňho zaujímavý.
V zadaní budeme riešiť úlohu odporúčania produktu pomocou metódy next item prediction. Zo sekvencie hodnotení používateľa sa budeme snažiť predikovať ďalší film, ktorý si pozrie a ten mu odporučíme. Ďalšou úlohou, ktorú budeme riešiť v zadaní je zobrazenie jednotlivých filmov vo vektorovej podobe, tak aby filmy s podobnými vlastnosťami boli blízko seba, teda item-embedding. Následne v sekvencií používateľa zameníme filmy za ich číselnú reprezentáciu. Po zámene očakávame, že sa výsledné odporúčanie zlepší.
Pri riešení zadania budeme skúmať úspešnosť LSTM v oblasti odporúčaní. Taktiež budeme porovnávať použitie dát v textovej a vektorovej podobe a jeho vplyv na výsledné predikcie.

1.1 Podobné práce

Kolaboratívne filtrovanie s využitím rekurentných neurónových sietí - práca sa venuje skúmaniu kolaboratívneho filtrovania z pohľadu sekvenčnej predikcie. Na generovanie odporúčaní využíva LSTM neurónové siete, vďaka ktorým dosahuje lepšie výsledky, ako v prípade metód najbližších susedov alebo maticovej faktorizácie, a to najmä v oblasti krátkodobých odporúčaní.
Dlhodobé a krátkodobé odporúčania s využitím rekurentných neurónových sietí - práca poskytuje vizualizáciu a porovnanie viacerých odporúčacích systémov. Taktiež skúma možné modifikácie rekurentných neurónových sietí za účelom prispôsobenia pre krátkodobé alebo dlhodobé odporúčania.

1.2 Návrh

Ako prvý bod pri riešení nášho zadania namapujeme filmy do vektorového priestoru a to tak aby filmy, ktoré sa zvyknú pozerať spolu boli blízko seba. Tento problém budeme riešiť pomocou známej metódy v spracovaní prirodzeného jazyka word embedding. Ide o metódu, ktorá mapuje slová na číselnú reprezentáciu a to tak, aby sa slová s podobnými vlastnosťami nachádzali blízko seba. Naša metóda bude veľmi podobná, ale namiesto sekvencie slov budeme používať sekvenciu filmov v poradí, v akom boli hodnotené používateľmi. Ako ďalší krok vytvorené mapovanie spojíme spolu s obsahovými vlastnosťami filmov. V poslednom kroku, zoberieme sekvenciu používateľom hodnotených filmov, už namapovaných na číselnú reprezentáciu a využijeme neurónovú sieť LSTM (Long short-term memory) na predikciu ďalšieho filmu.

2 Dátová analýza

Na overenie úspešnosti nášho modelu sme si vybrali známy dataset Movielens. Dataset Movielens obsahuje hodnotenia filmov používateľmi, informácie o hodnotených filmoch a

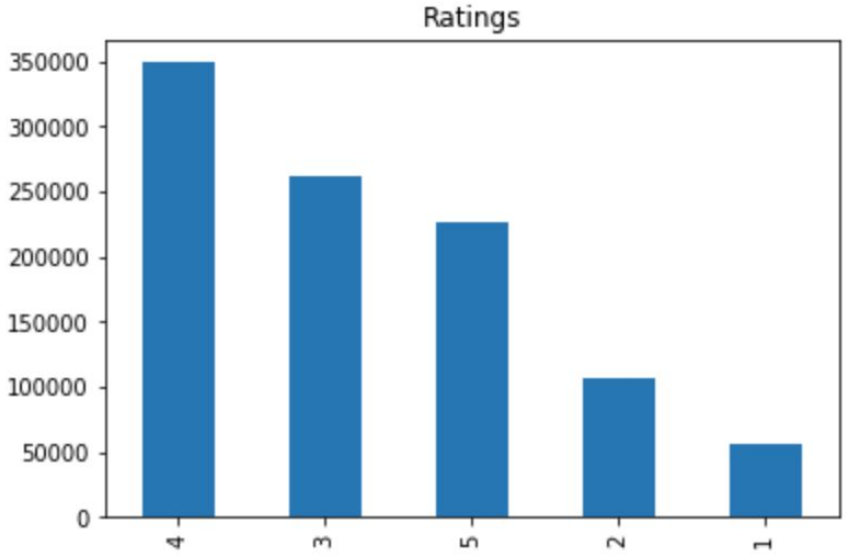
informácie o používateľoch. Movielens poskytuje viacero datasetov s rôznym množstvom dát. My sme si vybrali dataset o veľkosti 1 milióna. Informácie o používateľoch v našom modeli nebudeme potrebovať.

2.1 Hodnotenia filmov

V datasete sa nachádza 1 000 209 záznamov, ktoré sú plnohodnotné a pre nijaký atribút nechýbajú údaje. Dataset hodnotení obsahuje 4 atribúty: id filmu, id používateľa, hodnotenie a čas hodnotenia.

Názov	**Typ**	**Popis**
UserID	celé číslo	identifikačné číslo hodnotiaceho používateľa
MovieID	celé číslo	identifikačné číslo ohodnoteného filmu
Rating	celé číslo	používateľove hodnotenie filmu
Timestamp	desatinné číslo	čas vykonania hodnotenia

Hodnotenia filmov sú definované na škále od 1 do 5. Priemerná hodnotenie je 3,58 so štandardnou odchýlkou 1,12. Najčastejšie udeľované hodnotenie je 4, zatiaľ čo najmenej časté je 1.



V datasete sa nachádza 6 040 Pre riešený problém sú dôležité sekvencie akcií. Preto sme v datasete analyzovali počet akcií vykonaných jednotlivými používateľmi.

- minimálny počet akcií: 20
- priemerný počet akcií: 166
- maximálny počet akcií: 2314

Ďalej sme sa zamerali na počet hodnotení jednotlivých filmov, nakoľko pri malom počte hodnotení, by bolo zložité nájsť medzi filmami podobnosti.

- minimálny počet hodnotení: 1
- priemerný počet hodnotení: 270
- maximálny počet hodnotení: 3428

Hodnotenia filmov pochádzajú z časového obdobia takmer 3 rokov a to od 25.4.2000 do 28.2.2003.

2.2 Informácie o filme

Celkový počet filmov v datasete je 3952. Dataset filmov obsahuje 3 atribúty: id filmu, názov filmu a žánre filmu.

Názov	**Typ**	**Popis**
MovieID	celé číslo	identifikačné číslo filmu
Title	text	názov filmu spolu s rokom vydania
Genres	text	žánre pod ktoré spadá film odedelné |

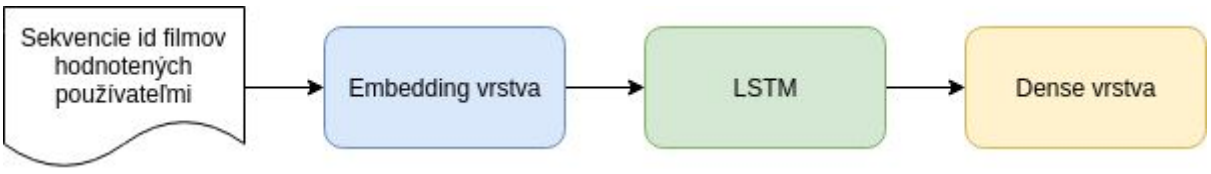
Z tohto datasetu použijeme informácie o žánroch, ktorých sa v ňom nachádza 18 unikátnych.

3 Architektúra neurónovej siete

3.1 Prvotný návrh

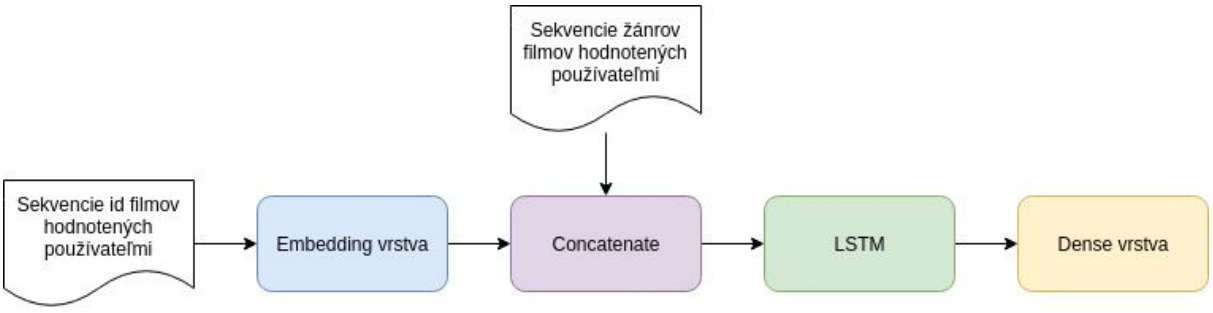
Neurónovú sieť vytvoríme pomocou modelu, skladajúceho sa z 3 vrstiev:

- Embedding - V tejto vrstve vytvoríme vektory filmov na základe sekvencií ich hodnotení používateľmi.
- LSTM - Vytvoríme predikcie na základe vstupných sekvencií.
- Dense - Pre každý film obsahuje 1 výstup. Určí najpravdepodobnejší film na základe softmaxu.



3.2 Výsledná architektúra

Od prvotného návrhu sa líši pridaním jedného kroku. Po vytvorení embeddingov filmov k nim pripojíme one-hot enkódované žánre.



Vstupom pre neurónovú sieť je pole s 2 prvkami. Prvým prvkom je matica sekvencií id hodnotených filmov pre jednotlivých používateľov. Druhým prvkom sú sekvencie embeddingov žánrov hodnotených filmov pre jednotlivých používateľov.

Sekvencie filmov sme vytvorili tak, že sme si zoradili dataset hodnotení podľa času. Následne sme zagregovali hodnotenia pre jednotlivých používateľov. Nakoľo používatelia mali rôzne množstvo udelených hodnotení (20 - 2 314 hodnotení na používateľa), rozhodli sme sa pre používateľov s príliš veľkým množstvom akcií rozdeliť ich akcie na sekvencie určitej dĺžky. Vďaka tomu sme získali väčšie množstvo sekvencií pre náš model. Zo všetkých vytvorených sekvencií sme nakoniec vyfiltrovali tie, ktoré boli príliš krátke (mali dĺžku menšiu ako polovica ideálnej dĺžky sekvencie hodnotených filmov pre jednotlivých používateľov). Ostatné sekvencie sme zarovnali na jednotnú dĺžku pomocou paddingu. Embedingy žánrov pre filmy sme získali pomocou one-hot enkódovania žánrov príslušného filmu.

Zo sekvencií sme pomocou embedding vrstvy získali embedingy pre jednotlivé filmy. Embeddingy sme následne poslali do concatenate vrstvy, kde sme k nim pripojili embedingy žánrov získané z druhej časti vstupu neurónovej vrstvy. Výstup z tejto vrstvy pokračoval do LSTM a z nej do výstupnej dense vrstvy s veľkosťou 3 952 (počet hodnotených filmov).

Výsledkom neurónovej siete sú predikované pravdepodobnosti pre každý film. Z nich vytvoríme rebríček 10 najpravdepodobnejších filmov, ktoré používateľ ohodnotí ako ďalšie.

3.3 Trénovanie modelu neurónovej siete

Časovo usporiadané dáta sme si rozdelili na trénovaciu (60%), validačnú (20%) a testovaciu sadu (20%). Pre každú sadu sme vytvorili sekvencie hodnotení. Model sme trénovali s variabilnou veľkosťou batchu a počtom epoch.

Úspešnosť natrénovaného modelu sme vyhodnocovali pomocou 2 metrík:

- sps (short term prediction success) - zachytáva schopnosť modelu predikovať ďalší film v poradí. Pre každý film je hodnota buď 1 (ak sa ďalší film nachádza medzi 10 predikovanými) alebo 0 (ak sa ďalší film nenachádza medzi 10 predikovanými). Súčet týchto hodnôt je vydelený počtom predikcií.
- item coverage - zachytáva schopnosť modelu robiť rozmanité úspešné predikcie. Metriku vypočítame ako súčet úspešne predikovaných unikátnych filmov predelený súčtom všetkých unikátnych filmov nachádzajúcich sa v sade ďalších filmov pre používateľov.

4 Experimenty

Ako baseline sme použili odporúčanie 10 najpopulárnejších produktov pre všetkých používateľov. S týmto prístupom sme dosiahli nasledovné hodnoty:

- **sps:** 0.0236
- **item coverage:** 0.0040

Experimenty nachádzajúce sa v spoločnej tabuľke sú vykonávané pri rovnakých podmienkach (rovnako nastavených parametroch) s výnimkou testovaného parametra.

opis experimentu	**výsledné metriky**	
veľkosť LSTM	sps	item coverage
32	0.0325	0.0565
64	**0.0363**	**0.0659**
256	0.0313	0.0632
1024	0.0275	0.0471

opis experimentu	**výsledné metriky**	
veľkosť batchu	sps	item coverage
64	**0.0388**	**0.0686**
32	0.0301	0.0565

opis experimentu	**výsledné metriky**	
dĺžka sekvencie	sps	item coverage
50	0.0388	0.0686
20	0.0641	0.1498
10	**0.0906**	**0.2294**

opis experimentu	**výsledné metriky**	
veľkosť embeddingu	sps	item coverage

400	0.0906	0.2294
200	0.0927	0.2342
100	0.0967	0.2421
50	**0.1013**	**0.2401**
25	0.1009	0.2274

opis experimentu	**výsledné metriky**	
optimizer	sps	item coverage
adam	**0.0967**	**0.2421**
sgd	0.0246	0.0396

opis experimentu	**výsledné metriky**	
loss funkcia	sps	item coverage
sparse cross categorical entropy	**0.0967**	**0.2421**
cross categorical entropy	0.0034	0.0032
categorical hinge	0.0028	0.0040

opis experimentu	**výsledné metriky**	
embeding žánrov	sps	item coverage
áno	**0.0967**	**0.2421**
nie	0.0870	0.2017

opis experimentu	**výsledné metriky**	
váhovanie tried	sps	item coverage
žiadne	**0.0984**	**0.2433**
balanced	0.0967	0.2421

opis experimentu	**výsledné metriky**	
počet epoch	sps	item coverage
32	**0.1013**	**0.2401**
256	0.0678	0.2049

opis experimentu	**výsledné metriky**	
early stopping	sps	item coverage
áno	0.0959	0.1661
nie	**0.1013**	**0.2401**

opis experimentu	**výsledné metriky**	
regularizácia	sps	item coverage
áno	0.0246	0.0396
nie	**0.0959**	**0.1661**

4.1 Výsledný model

Na základe vykonaných experimentov sme vytvorili finálny model s nasledovnými parametrami:

- veľkosť LSTM: 64
- veľkosť batchu: 64
- počet epoch: 32
- dĺžka sekvencií: 10
- veľkosť embeddingu: 50
- optimizer: adam
- loss funkcia: sparse categorical cross entropy
- bez váhovania tried
- použitie embedingu žánrov

5. Záver

Úspešne sa nám podarilo implementovať úlohu odporúčania pomocou neurónovej siete. Podarilo sa nám tiež úspešne overiť riešenie na reálnej dátovej sade. S pomocou Embeddingu a LSTM siete sme dokázali odporúčať viacero filmov, ktoré boli pre používateľa relevantné. Pomocou pridania obsahových prvkov sa nám podarilo zlepšiť presnosť odporúčania Pri určovaní finálneho modelu sme vyskúšali viacero experimentov.

V budúcnosti je možné zlepšenie v pridávaní viacerých obsahových prvkov k embeddingu. Ďalšou pridanou hodnotou môže byť overenie v inej doméne alebo nad dátovou sadou s implicitnou spätnou väzbou.