

RELATÓRIO IMPLEMENTAÇÃO RSVD

Autores: Paulo Roberto Xavier Júnior, Gabriel Segobia, Fellipe Bravo

E-mails: contato.pauloxavier@gmail.com, segobia.gos@gmail.com, fellipe.bravo@gmail.com

1 Sobre

Devido ao “Netflix Prize”, avanços significativos em técnicas de CF ocorreram, dentre elas muitos avanços nos métodos baseados em modelos de *SVD*. O *SVD* é bem estabelecido para identificar fatores semânticos latentes em *Information Retrieval*. O *SVD* em sua forma convencional possui algumas dificuldades devido a esparsidade das matrizes utilizadas nesse tipo de sistema.

Além das matrizes serem esparsas, usar apenas os dados conhecidos observados nas matrizes são altamente propensos a *overfit*. Os trabalhos iniciais dependiam da inserção de dados para preencher as avaliações em branco e fazer a matriz densa. Entretanto, essa inserção é altamente custosa, pois aumenta significativamente a quantidade de dados além de distorcer devido a inserção imprecisa de dados.

Os Trabalhos mais recentes sugerem modelar diretamente as avaliações já existentes, enquanto tenta evitar o *overfitting* através de um modelo adequado regularizado. Os modelos de *SVD* utilizados recentemente são: *SVD*: o *svd* puro, *SVD++*: onde é integrado outras formas de feedback do usuário para aumentar a acurácia a precisão e também *Time SVD++*, onde é levado em consideração também o tempo como um contexto e que a percepção e popularidade de um produto pode mudar com o tempo. Nesse trabalho, será usado o *Regularized SVD*.

2 Definição

O *Regularized SVD* é uma técnica de fatorização de matriz que busca reduzir o erro no treinamento. É uma técnica direta onde o usuário e as matrizes são atualizados com a ajuda de um Gradiente Descendente Estocástico. Essas atualizações penalizam o aprendizado de grandes valores de recursos por meios da regularização.

No *RSVD*, além do parâmetro de regularização λ , são passados também o ranking usuário e as matrizes de item. Esse modelo também é aprimorado se inserindo os bias para cada usuário e itens.

3 Resultados

A base utilizada foi o *Movielens100k*. E esta foi separada em Teste e Treinamento, usando 20 e 80% respectivamente da base. Inicialmente o modelo foi treinado usando a base de treinamento, e partir do resultado obtido, foi feita a predição a partir da base de teste (com 20k de entradas), obtendo-se os valores de predição para as outras entradas restantes (80k).

Table 1: Resultados RSVD

	<i>lr</i>	<i>Lambda</i>	<i>Delta</i>	<i>Acurácia</i>
1	.0003	.0013	.00032	0.8593
2	.0003	.0013	.00032	0.8468
3	.001	.02	.32	1.5285
4	.001	.03	.00032	0.8530

Em seguida, foi feito o MAE para testar a acurácia do método, e a partir do mesmo, encontramos os resultados apresentados na tabela 1, localizada na página 2.

4 Conclusão

O *Regularized SVD* é tido como um dos modelos mais precisos disponíveis para *Collaborative Filtering* em datasets tendo uma densidade na matriz de 2-5%, o que de maneira geral, inclui todos os datasets.

Existem diversos métodos para aumentar a acurácia, especialmente as técnicas de pré-filtragem e melhorias nas funções de Similaridade, que quanto mais complexas e usadas em conjunção com métodos de normalização das avaliações e afins, aumentariam dramaticamente o desempenho do KNN. Entretanto, nesse trabalho, foi utilizado apenas o KNN e uma função de similaridade para fins de aprendizado.

A partir dos resultados encontrados concluímos que o RSVD é definitivamente uma ótima abordagem para a filtragem colaborativa tendo em vista que a sua precisão adapta-se aos diversos datasets disponíveis, mantendo um padrão não importando o tamanho do dataset.

Concluímos também que o seu tempo de execução é mais aceitável do que em outros algoritmos disponíveis, pois tanto em seu treinamento quanto em sua predição, o algoritmo mostrou-se eficiente e rápido.