

Aprendizado de Máquina

Aula 9: Raciocínio Baseado em Casos e Sistemas de Recomendação (Parte 2)

André C. P. L. F de Carvalho
ICMC/USP

andre@icmc.usp.br



Tópicos do módulo

- Introdução
- Raciocínio baseado em casos
- Sistemas de recomendação
- Principais abordagens
- Filtragem por conteúdo
- Filtragem colaborativa
- Métodos baseados em modelos
- Métodos baseados em memória

Tópicos do módulo

- Introdução
- Raciocínio baseado em casos
- Sistemas de recomendação
- Principais abordagens
- Filtragem por conteúdo
- Filtragem colaborativa
- Métodos baseados em modelos
- Métodos baseados em memória

Principais abordagens de RSs

- Filtragem baseada em conteúdo
 - Usa informações sobre o próprio usuário ativo, ou item desejado, para propor itens
- Sistemas baseados em conhecimento
 - Usados para itens que não são frequentemente adquiridos
 - Ex.: imóveis, veículos e artigos de luxo
- Filtragem colaborativa (mais comum)
 - Usa escolhas de itens feitas por outro(s) usuário(s) para propor itens ao usuário ativo
- Aspectos dessas abordagens podem ser combinados, gerando uma abordagem híbrida

Métodos de RSs

- Podem utilizar os seguintes dados
 - Informações sobre usuários e sobre itens
 - Perfil de usuários e propriedades de itens
 - Palavras chave consideradas importantes
 - **Filtragem baseada em conteúdo**
 - Interações usuário-item
 - O quanto um outro usuário gostou de um dado item
 - **Filtragem colaborativa e filtragem baseada em conteúdo**

Filtragem baseada em conteúdo

- Content-based filtering ou content-based recommendation (CB)
- Conteúdo:
 - Descrição (propriedades) dos itens avaliados ou comprados pelo usuário ativo
- Assume que o usuário ativo vai gostar de itens parecidos com os que ele gostou (avaliou bem)
 - Itens parecidos = itens com conteúdo parecido
- Escolhida quando não temos acesso às avaliações de outros usuários
 - Que impede o uso de filtragem colaborativa

Filtragem baseada em conteúdo

- Maioria dos métodos adotados vem da área de recuperação de informação
 - Primeiros sistemas eram baseados em palavras chave (keywords)
- Passos iniciais para criar um RS baseado em conteúdo
 - Definir as propriedades dos itens (atributos preditivos) a serem consideradas
 - Definir como os itens será avaliados (atributo alvo)
 - Rotular um subconjunto dos itens (conjunto de treinamento)
 - Demais itens farão parte do conjunto de teste

Filtragem baseada em conteúdo

- Métodos baseados em conteúdo podem usar algoritmos de aprendizado de máquina
 - Cria um modelo preditivo específico para cada usuário
 - Fase de treinamento
 - Aplica algoritmo de aprendizado de máquina aos itens do conjunto de treinamento
 - Modelo preditivo induzido representa as preferências do usuário
 - Fase de teste
 - Aplica o modelo preditivo aos itens do conjunto de teste
 - Retorna o item com o maior valor predito (estimativa de avaliação)

Exemplo: conjunto de filmes

- Para cada objeto (filme), temos:
 - Identificação do filme
 - Atributos preditivos (propriedades dos filmes)
 - Atributo alvo (nota de 0 a 10, dada pelo usuário após assistir o filme)

Título	Ano	Duração	Gênero	Avaliação
Dumbo	1941	64	Desenho	6
Bonequinha de luxo	1961	115	Clássico	9
Casablanca	1942	103	Clássico	10
Pinóquio	1940	88	Desenho	7
Robin Hood	1973	83	Desenho	7

Sistemas baseados em conhecimento

- Knowledge based systems (KBS)
- Recomendações são baseadas em requisitos explicitamente especificados pelo usuário ativo
 - Ao invés do histórico de avaliações do usuário ativo, e dados de suas compras
 - Usuários informam seus interesses de forma interativa
 - Que são combinados com conhecimento sobre o domínio da aplicação
 - Dados relacionados ao contexto podem ser usados
 - Ex.: Informação social, temporal, de localização, etc.
 - Considerados por muitos similares a RSs baseados em conteúdo

Sistemas baseados em conhecimento

- Geralmente utilizados para itens raramente comprados
 - Podem não ter o número necessário de avaliações
 - Problema semelhante ao problema de *cold start*
- Usados também quando podem ocorrer alterações nas preferências dos usuários para os itens
 - Ex.: Modelo de carro pode evoluir ao longo dos anos, levando a mudanças nas preferências dos usuários
- De acordo com a interface com usuário, pode ser:
 - Sistema de recomendação baseada em restrições
 - Sistema de recomendação baseada em casos

Sistemas baseados em restrições

- Usuário ativo especifica requisitos ou restrições
 - Limites inferiores ou superiores
 - Ex.: valor máximo que está disposto a pagar, tamanho máximo do item, número máximo de anos do item, número de airbags
- Usam regras para casar restrições do usuário com as propriedades do item
 - Baseadas em conhecimento sobre o domínio da aplicação
 - Ex.: carros fabricados antes de 1980 não têm airbag
- Em geral, de acordo com os resultados da consulta, o usuário pode modificar suas restrições

Sistemas baseados em casos

- Usuário ativo alimenta o sistema com casos para serem usados como alvos ou pontos de referência
- Definem métricas de similaridade relacionadas ao domínio da aplicação para recuperar itens similares aos casos
 - Estas métricas formam o conhecimento de domínio usado pelo sistema de recomendação
- Itens retornados são geralmente alimentados no sistema pelo usuário ativo, como novos casos
 - Com modificações para torna-los mais próximos da preferência (alvo) do usuário ativo
 - Processo interativo que ajuda a chegar no item de interesse

Filtragem colaborativa

- *Collaborative filtering* (CF)
- O nome vem do usuário ativo contar com colaboração implícita de outros usuários para encontrar itens que possam interessá-lo
 - Assume que um usuário ativo deve gostar de itens escolhidos por usuários com gostos parecidos
 - Mas o usuário ativo não precisa conhecer os usuários com gostos parecidos ao dele
- Caso usuário ativo queira adquirir ou avaliar um novo item
 - Um RS deve encontrar usuários com preferências semelhantes às dele em outros itens
 - E recomendar ao usuário ativo um produto que eles gostaram



Filtragem colaborativa

- Baseada em comportamentos passados
 - Representados na matriz de avaliações (RM)
- Avaliações na matriz podem ser entendidas como feedback
 - Explícito
 - Implícito

Feedback explícito

- Usuários atribuem valores para avaliar itens, que indicam suas preferências
- Maior dificuldade para a coleta das primeiras avaliações
 - Problema de cold start
- Com frequência, não é fornecido
- Expressa melhor as preferências do usuário

Formas de feedback explícito

- Como um usuário pode avaliar cada item:
 - Valores binários:
 - Se um item é positivo (bom)  ou negativo (ruim) 
 - Valores distribuídos em uma faixa ou intervalo
 - A nota que o usuário dá a um item (Ex.: de 1 a 5)
 - Valores ordinais:
 - O que o usuário acha de um determinado item (Ex.: bom, médio, ruim)
 - Valor unário
 - Se gostou ou comprou um item (usuário pode deixar o item sem avaliação)

Feedback implícito

- Coletado a partir do comportamento do usuário
 - Ex.: número de clicks em links retornados por uma busca, tempo gasto em um site, ...
- Coleta de dados é não intrusiva
- Permite coleta de um volume de avaliações grande e diverso
- Conhecido como feedback apenas positivo (positive-only)
 - Usuário expressa interesse, nunca a falta de interesse

Filtragem colaborativa

- Utiliza basicamente dois grupos de métodos
 - Métodos baseados em modelos
 - Usam modelos para prever as avaliações de um usuário para novos itens
 - Métodos baseados em memória
 - Não geram modelos
 - Em geral, usam algoritmos baseados em vizinhança
 - Ex.: Algoritmo k-NN

Métodos baseados em modelos

- Modelos podem ser criados utilizando
 - Métodos de fatoração de matrizes
 - **Geram modelos de fatores latentes**
 - Estimam avaliações a partir de características de itens e de usuários
 - Cada fator latente equivale a uma característica
 - Algoritmos de aprendizado de máquina
 - Induzem modelos preditivos
 - Ex.: Árvore de decisão, modelo de regressão, redes neurais

Fatoração de matrizes

- Decomposição de uma matriz em um produto de duas ou mais matrizes
 - Nada mais é que quebrar uma tarefa em subtarefas mais simples

- Facilita o cálculo, ou a computação

Várias ideias chave em álgebra linear, se você olhar com atenção, são na verdade fatoração de matrizes. A matriz original vira um produto de 2 ou 3 matrizes

Gilbert Lang

- Usada desde o 2º grau
 - Para resolver sistemas de lineares do tipo $Ax = b$
 - Para fatorar um polinômio $(x^2 - 1)$ no produto $(x-1)(x+1)$

Métodos de fatoração de matrizes

- Em filtragem colaborativa, geram modelos de fatores latentes
 - Ex.: seja a matriz com avaliações de 0 (não gosta) , 1 (gosta) ou 2 (gosta muito) de 7 usuários (U1 a U7) para 6 itens (F1 a F6)

	F1	F2	F3	F4	F5	F6
U1	1	2	1	2	0	1
U2	1	0	1	2	0	1
U3	0	1	2	1	1	2
U4	1	2	0	2	2	1
U5	0	1	2	1	1	2
U6	0	1	2	1	1	2
U7	1	1	1	1	1	1

Métodos de fatoração de matrizes

- Em filtragem colaborativa, geram modelos de fatores latentes
 - Ex.: seja a matriz com avaliações de 0 (não gosta) , 1 (gosta) ou 2 (gosta muito) de 7 usuários (U1 a U7) para 6 itens (F1 a F6)

	F1	F2	F3	F4	F5	F6
U1	1	2				1
U2	1	0	1		0	1
U3	0		2		1	2
U4	1	2		2		1
U5		1	2	1		2
U6		1	2		1	2
U7	1	1	1	1	1	1

Na prática, matrizes de avaliações têm vários itens sem avaliações

Fatores latentes

- Fatoração de uma RM pode gerar matrizes que explicam as avaliações dadas pelos usuários para os itens avaliados
 - Fatores latentes
 - Resumem em uma única linha informações presentes em mais de uma linha de uma RM (e/ou resumem colunas em uma única coluna)
 - Também chamados de conceitos latentes e aspectos latentes
 - A estatística, de forma similar, define variáveis latentes
 - Variáveis que não são diretamente observadas, mas podem ser inferidas das variáveis observadas

Fatoração de matrizes

- Para isso, extrai fatores latentes de uma matriz

	F1	F2	F3	F4	F5	F6
U1	1	2				1
U2	1	0	1		0	1
U3	0		2		1	2
U4	1	2		2		1
U5		1	2	1		2
U6		1	2		1	2
U7	1	1	1	1	1	1

R

\approx

U

V^T

Fatores latentes em relação
às **propriedades** dos itens

No. de fatores latentes =
No. de colunas em U e de
linhas em V^T

Preferências dos
usuários em relação
aos fatores latentes

Modelos de fatores latentes

- Latent factor models (LFM)
- Estado da arte de sistemas colaborativos
- Formados pelo conjunto de fatores latentes (vetores latentes) decompostos da matriz de avaliações (RM)
 - Exploram redundâncias presentes na RM
 - Tornam mais claras características de usuários (preferências) e de itens (propriedades) não diretamente observadas
 - Conseguem capturar padrões que outros modelos não capturam
 - Bons para análise semântica
 - Geram boa estimativa para os itens não avaliados

Função de custo

- S: conjunto de todos os pares usuário-item (i,j) em R com avaliação
 - $S_{ij} = \{(i,j) \text{ tal que } r_{ij} \text{ tem uma avaliação}\}$
 - Com $i \in \{1, 2, \dots, m\}$ e $j \in \{1, 2, \dots, n\}$
- Ao fatorar a matriz incompleta R por um produto UV^T que aproxima R, os elementos preenchidos de R podem ser preditos por

$$\hat{r}_{ij} = \sum_{l=1}^k u_{il} \cdot v_{lj}$$

- Diferença nas avaliações para cada entrada de R: $e_{ij} = (r_{ij} - \hat{r}_{ij})$
- Função de custo para matriz incompleta:

$$J = \frac{1}{2} \sum_{(i,j) \in S} e_{ij}^2$$

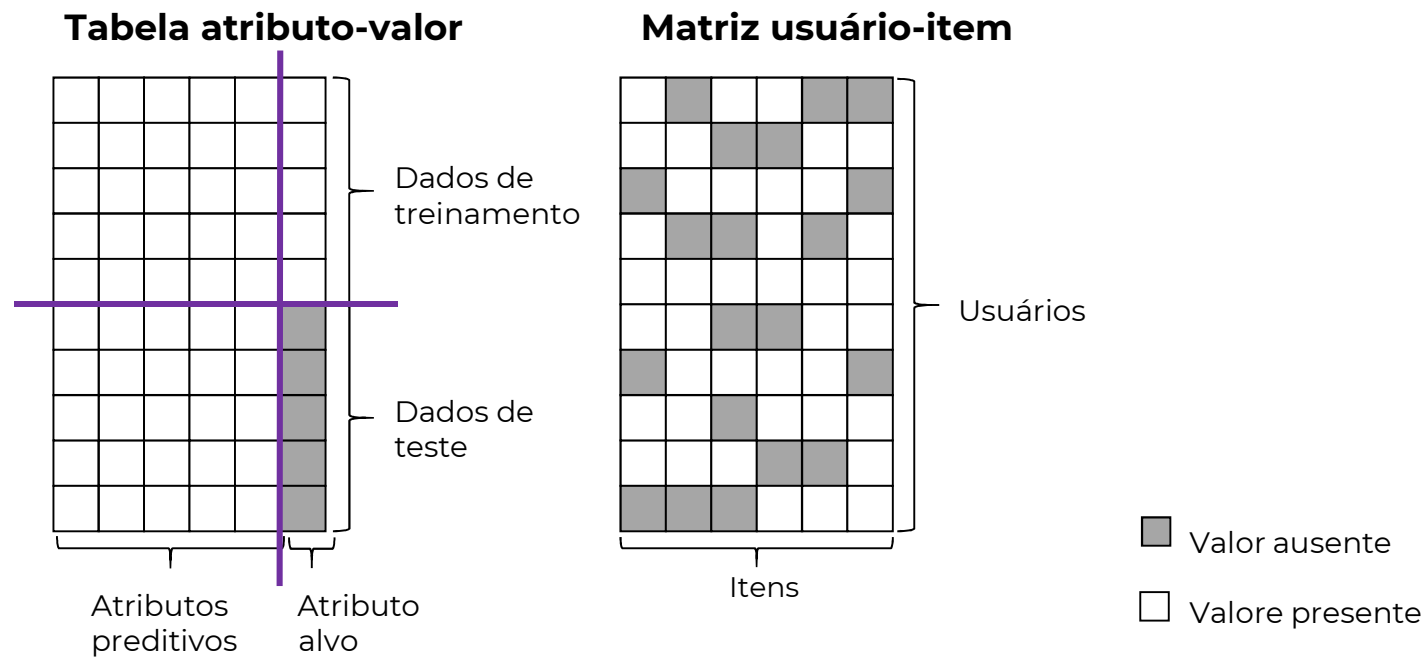
- Computada apenas para os elementos em S

Métodos baseados em modelos

- Algoritmos de aprendizado de máquina e de otimização têm sido usados para induzir modelos para prever avaliações
- Algoritmos mais usados são baseados em:
 - Máquinas de vetores de suporte
 - Particularmente para avaliações binárias
 - Algoritmos para treinamento de redes neurais (incluindo redes profundas)
 - Regressão (incluindo linear e logística)

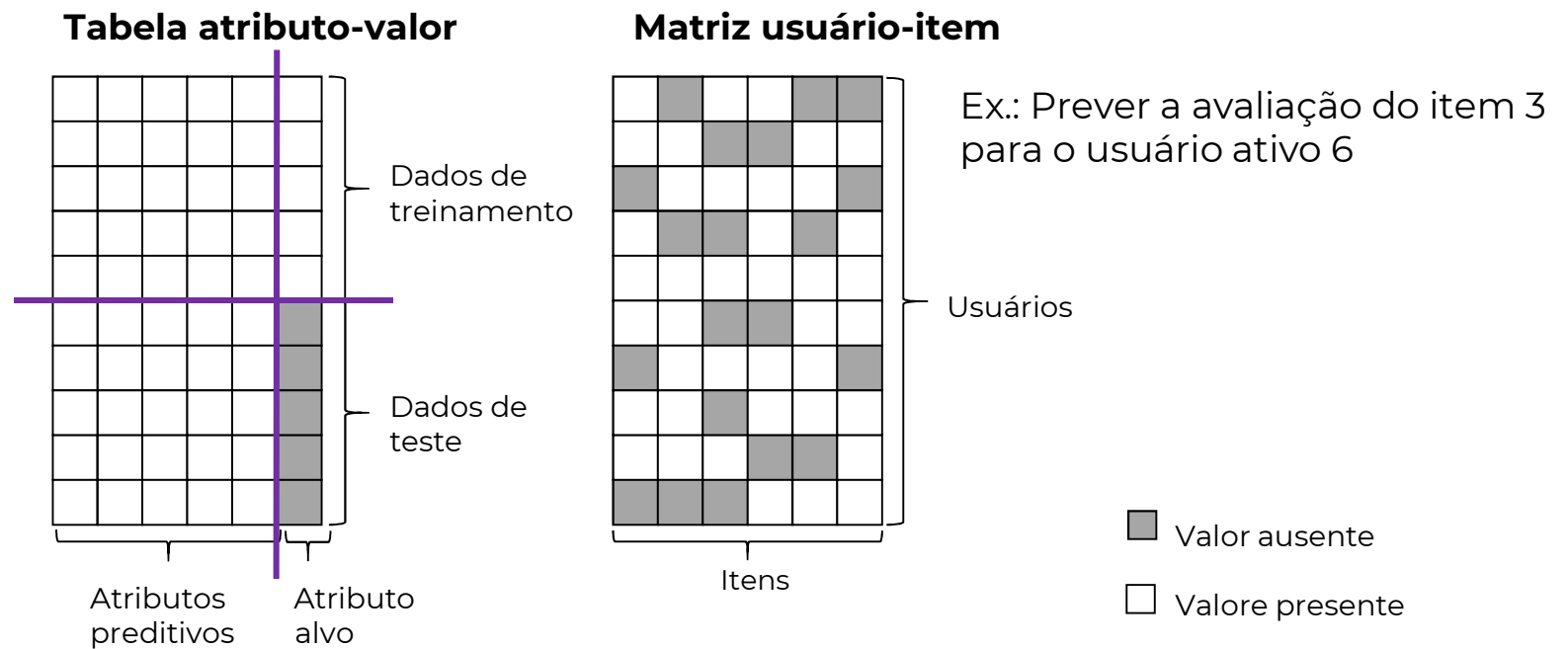
Treinamento

- Nas duas tarefas, dados são divididos em treinamento e teste



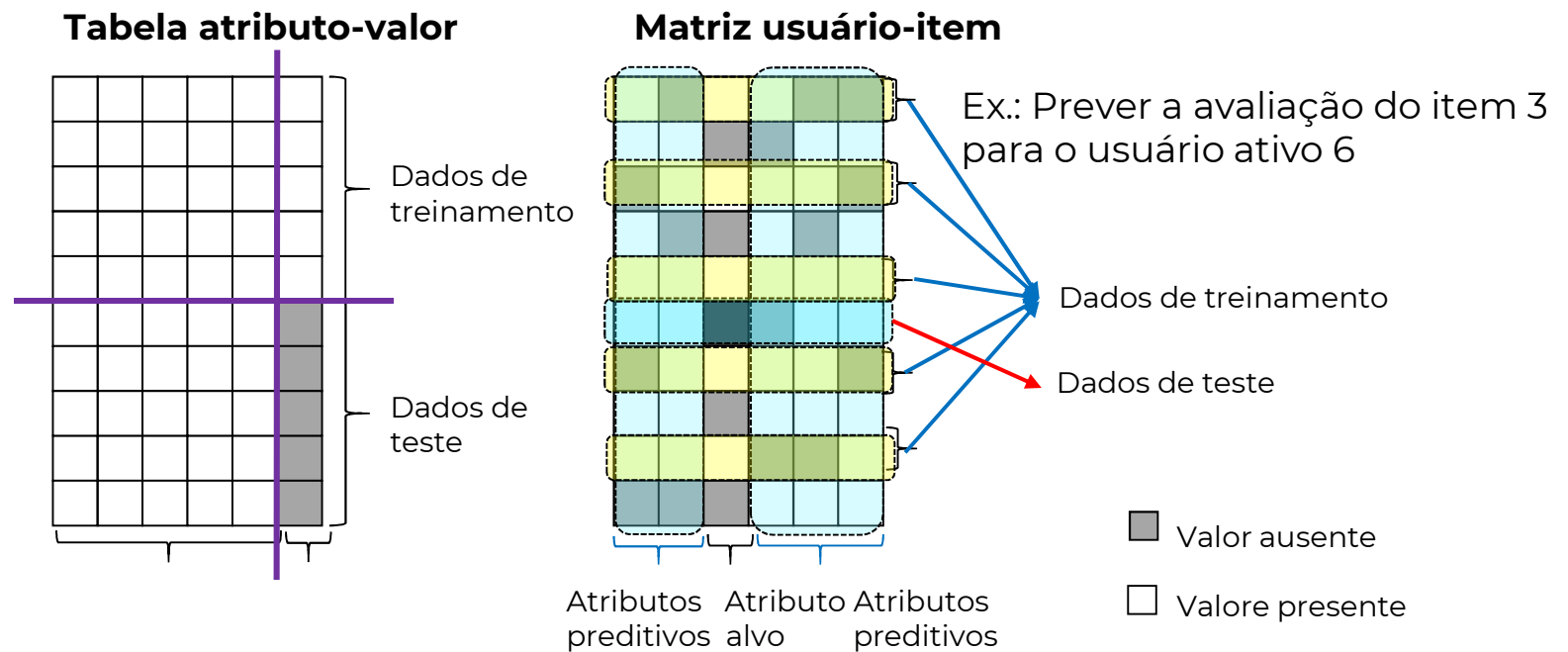
Treinamento

- Nas duas tarefas, dados são divididos em treinamento e teste



Treinamento

- Nas duas tarefas, dados são divididos em treinamento e teste



Contínua na
próxima aula