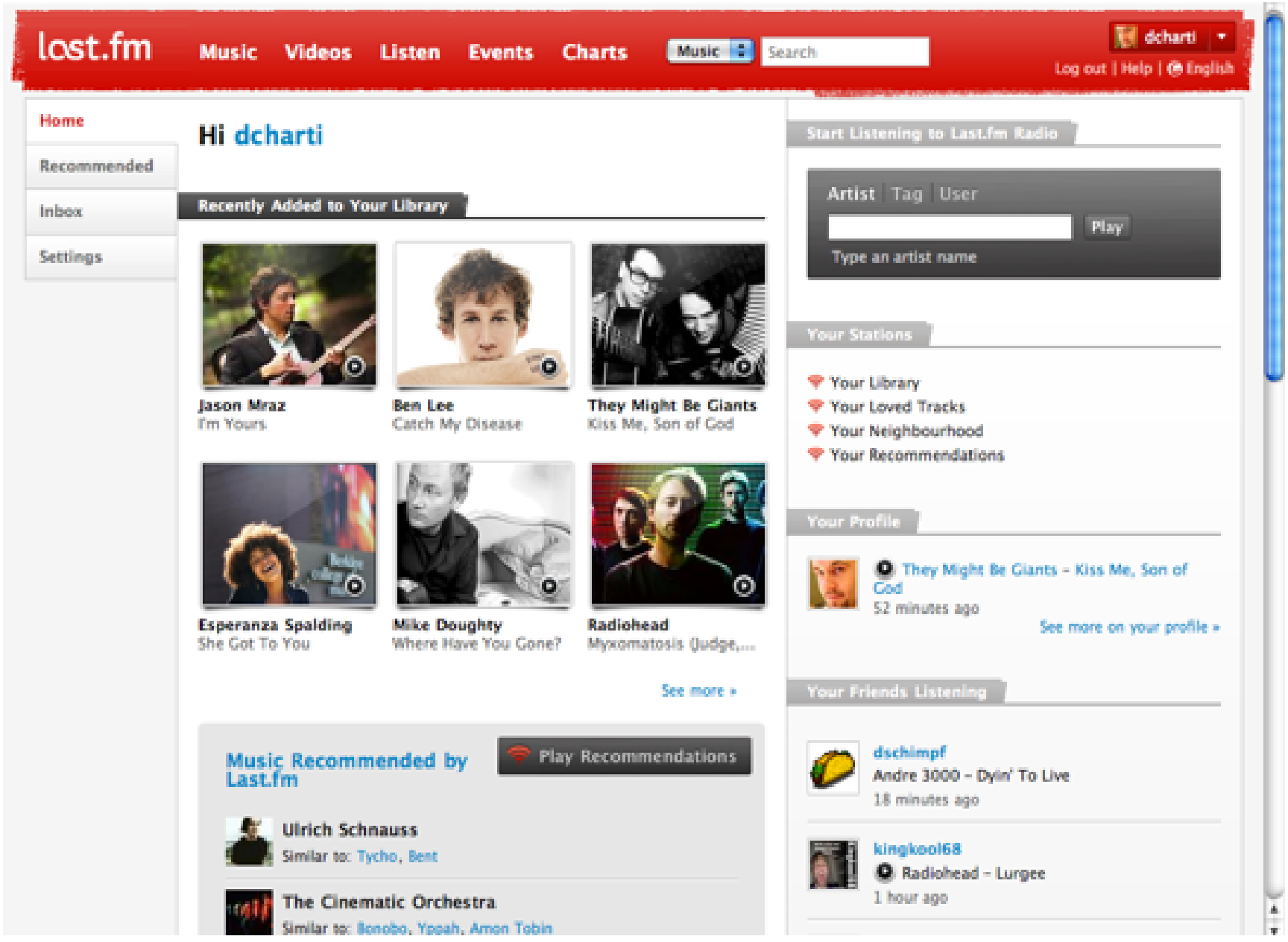
Mineração de Dados

Aula 14: Sistemas de Recomendação

Rafael Izbicki

NETFLIX Navegar Perfil de interesses KiDS Títulos, gente e gêneros Thomas

Principais escolhas para Thomas


last.fm
Music
Videos
Listen
Events
Charts
Music
Search
dcharti
Log out | Help | English
Home
Recommended
Inbox
Settings
Hi dcharti
Recently Added to Your Library
Jason Mraz
I'm Yours
Ben Lee
Catch My Disease
They Might Be Giants
Kiss Me, Son of God
Esperanza Spalding
She Got To You
Mike Doughty
Where Have You Gone?
Radiohead
Myxomatosis (Judge,...
See more »
Music Recommended by Last.fm
Play Recommendations
Ulrich Schnauss
Similar to: Tycho, Bent
The Cinematic Orchestra
Similar to: Bonobo, Ypoah, Amon Tobin
Start Listening to Last.fm Radio
Artist
Tag
User
Play
Type an artist name
Your Stations
Your Library
Your Loved Tracks
Your Neighbourhood
Your Recommendations
Your Profile
They Might Be Giants – Kiss Me, Son of God
52 minutes ago
See more on your profile »
Your Friends Listening
dschimpf
Andre 3000 – Dyin' To Live
18 minutes ago
kingkool68
Radiohead – Lurgee
1 hour ago

Frequently Bought Together

Price For All Three: $72.32

Add all three to Cart | Add all three to Wish List

Show availability and shipping details

- This Item: Punk Love by Susie J. Horgan Hardcover $19.06
- Fucked Up + Photocopied: Instant Art Of The Punk Rock Movement by Bryan Ray Turcotte Hardcover $34.25
- Touch and Go: The Complete Hardcore Punk Zine '79-'83 by Tesco Vee Paperback $24.01

Customers Who Bought This Item Also Bought

Page 1 of 9

Fucked Up + Photocopied: Instant Art Of The... by Bryan Ray Turcotte

(16)

$34.25

Punk Pioneers by Jenny Lens

(13)

Touch and Go: The Complete Hardcore Punk Zine '7... by Tesco Vee

(9)

$24.01

The Encyclopedia of Punk by Brian Cogan

(11)

$14.56

Radio Silence: A Selected Visual History of... by Nathan Nedorostek

(13)

$21.17

Cheetah Chrome: A Dead Boy's Tale: From the... by Cheetah Chrome

(9)

$18.99

American Hardcore - The History of Punk Rock 198... DVD ~ Dez Cadena

(49)

$8.99

Como escolher que produtos recomendar a cada usuário?

Um método simples: regras de associação que já estudamos.

Veremos nesta aula outros métodos.

Como escolher que produtos recomendar a cada usuário?

Um método simples: regras de associação que já estudamos.

Veremos nesta aula outros métodos.

Como escolher que produtos recomendar a cada usuário?

Vamos assumir que cada usuário pode atribuir uma nota (*avaliação*) para cada produto. Por exemplo, no Netflix, podemos avaliar cada filme como 1, 2, 3, 4 ou 5 estrelas.

De um ponto de vista formal, o problema de um sistema de recomendação é o seguinte.

Considere que temos um conjunto de usuários

$$\mathcal{U} = \{u_1, \ldots, u_m\}$$

e um conjunto de produtos (ou *itens*)

$$\mathcal{I} = \{i_1, \ldots, i_n\}.$$

Seja $R_{j,k}$ a avaliação dada pelo usuário j ao produto k.

Como escolher que produtos recomendar a cada usuário?

Vamos assumir que cada usuário pode atribuir uma nota (*avaliação*) para cada produto. Por exemplo, no Netflix, podemos avaliar cada filme como 1, 2, 3, 4 ou 5 estrelas.

De um ponto de vista formal, o problema de um sistema de recomendação é o seguinte.

Considere que temos um conjunto de usuários

$$\mathcal{U} = \{u_1, \ldots, u_m\}$$

e um conjunto de produtos (ou *itens*)

$$\mathcal{I} = \{i_1, \ldots, i_n\}.$$

Seja $R_{j,k}$ a avaliação dada pelo usuário j ao produto k.

Como escolher que produtos recomendar a cada usuário?

Vamos assumir que cada usuário pode atribuir uma nota (*avaliação*) para cada produto. Por exemplo, no Netflix, podemos avaliar cada filme como 1, 2, 3, 4 ou 5 estrelas.

De um ponto de vista formal, o problema de um sistema de recomendação é o seguinte.

Considere que temos um conjunto de usuários

$$\mathcal{U} = \{u_1, \ldots, u_m\}$$

e um conjunto de produtos (ou *itens*)

$$\mathcal{I} = \{i_1, \ldots, i_n\}.$$

Seja $R_{j,k}$ a avaliação dada pelo usuário j ao produto k.

Como escolher que produtos recomendar a cada usuário?

Vamos assumir que cada usuário pode atribuir uma nota (*avaliação*) para cada produto. Por exemplo, no Netflix, podemos avaliar cada filme como 1, 2, 3, 4 ou 5 estrelas.

De um ponto de vista formal, o problema de um sistema de recomendação é o seguinte.

Considere que temos um conjunto de usuários

$$\mathcal{U} = \{u_1, \ldots, u_m\}$$

e um conjunto de produtos (ou *itens*)

$$\mathcal{I} = \{i_1, \ldots, i_n\}.$$

Seja $R_{j,k}$ a avaliação dada pelo usuário j ao produto k.

Usuário/Produto	i_1	i_2	$\ldots$	i_n
u_1	$R_{1,1}$	$R_{1,2}$	$\ldots$	$R_{1,n}$
u_2	$R_{2,1}$	$R_{2,2}$	$\ldots$	$R_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
u_m	$R_{m,1}$	$R_{m,2}$	$\ldots$	$R_{m,n}$

Em geral conhecemos apenas algumas das notas desta matriz.

Nesta aula veremos dois métodos de como imputar essas notas do modo a descobrir, para cada usuário, produtos não avaliados por ele que provavelmente o agradará.

Usuário/Produto	i_1	i_2	$\ldots$	i_n
u_1	$R_{1,1}$	$R_{1,2}$	$\ldots$	$R_{1,n}$
u_2	$R_{2,1}$	$R_{2,2}$	$\ldots$	$R_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
u_m	$R_{m,1}$	$R_{m,2}$	$\ldots$	$R_{m,n}$

Em geral conhecemos apenas algumas das notas desta matriz.

Nesta aula veremos dois métodos de como imputar essas notas do modo a descobrir, para cada usuário, produtos não avaliados por ele que provavelmente o agradará.

Usuário/Produto	i_1	i_2	$\ldots$	i_n
u_1	$R_{1,1}$	$R_{1,2}$	$\ldots$	$R_{1,n}$
u_2	$R_{2,1}$	$R_{2,2}$	$\ldots$	$R_{2,n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
u_m	$R_{m,1}$	$R_{m,2}$	$\ldots$	$R_{m,n}$

Em geral conhecemos apenas algumas das notas desta matriz.

Nesta aula veremos dois métodos de como imputar essas notas do modo a descobrir, para cada usuário, produtos não avaliados por ele que provavelmente o agradará.

Grosso modo, existem dois principais grupos de sistemas de recomendação:

- Sistemas baseado no conteúdo. Estes sistemas olham para características dos produtos que o usuário gosta (ex: diretores dos filmes, gêneros etc), e busca com base nisso descobrir outros produtos com as mesmas características.

- Sistemas baseados em filtros colaborativos. Nestes sistemas se olha apenas para a matriz de notas. Com base nela, busca-se então por padrões (ex: usuários parecidos com o usuário de interesse). O pressuposto básica é que usuários que concordaram no passado irão concordar no futuro.

Nesta aula estudaremos dois métodos baseados em filtros colaborativos.

Grosso modo, existem dois principais grupos de sistemas de recomendação:

- Sistemas baseado no conteúdo. Estes sistemas olham para características dos produtos que o usuário gosta (ex: diretores dos filmes, gêneros etc), e busca com base nisso descobrir outros produtos com as mesmas características.

- Sistemas baseados em filtros colaborativos. Nestes sistemas se olha apenas para a matriz de notas. Com base nela, busca-se então por padrões (ex: usuários parecidos com o usuário de interesse). O pressuposto básica é que usuários que concordaram no passado irão concordar no futuro.

Nesta aula estudaremos dois métodos baseados em filtros colaborativos.

Grosso modo, existem dois principais grupos de sistemas de recomendação:

- Sistemas baseado no conteúdo. Estes sistemas olham para características dos produtos que o usuário gosta (ex: diretores dos filmes, gêneros etc), e busca com base nisso descobrir outros produtos com as mesmas características.

- Sistemas baseados em filtros colaborativos. Nestes sistemas se olha apenas para a matriz de notas. Com base nela, busca-se então por padrões (ex: usuários parecidos com o usuário de interesse). O pressuposto básica é que usuários que concordaram no passado irão concordar no futuro.

Nesta aula estudaremos dois métodos baseados em filtros colaborativos.

Grosso modo, existem dois principais grupos de sistemas de recomendação:

- Sistemas baseado no conteúdo. Estes sistemas olham para características dos produtos que o usuário gosta (ex: diretores dos filmes, gêneros etc), e busca com base nisso descobrir outros produtos com as mesmas características.

- Sistemas baseados em filtros colaborativos. Nestes sistemas se olha apenas para a matriz de notas. Com base nela, busca-se então por padrões (ex: usuários parecidos com o usuário de interesse). O pressuposto básica é que usuários que concordaram no passado irão concordar no futuro.

Nesta aula estudaremos dois métodos baseados em filtros colaborativos.

Método 1: Filtro colaborativo baseado no usuário

Queremos estimar a nota que um usuário j dará a um produto l.

Ideia: primeiro procuramos os k usuários mais parecido a j. Calculamos então as médias das notas dadas por esses k usuários ao produto l (parecido com o KNN).

$$\widehat{R}_{j,l} = \frac{1}{k} \sum_{s \in \mathcal{V}_j} R_{s,l},$$

onde $\mathcal{V}_j$ é o conjunto dos k vizinhos mais próximos ao usuário j.

Método 1: Filtro colaborativo baseado no usuário

Queremos estimar a nota que um usuário j dará a um produto l.

Ideia: primeiro procuramos os k usuários mais parecido a j. Calculamos então as médias das notas dadas por esses k usuários ao produto l (parecido com o KNN).

$$\widehat{R}_{j,l} = \frac{1}{k} \sum_{s \in \mathcal{V}_j} R_{s,l},$$

onde $\mathcal{V}_j$ é o conjunto dos k vizinhos mais próximos ao usuário j.

Como definir quem são os usuários mais próximos a j?

Devemos medir a similaridade entre as notas atribuídas pelo usuário j a cada um dos produtos com as notas atribuídas pelos outros usuários.

Uma medida usual é a correlação de Pearson:

$$sim(u_a, u_b) = cor(R^*_{u_a,.}, R^*_{u_b,.}),$$

onde $R^*_{u_a,.}$ é o vetor de notas atribuídas pelo usuário u_a aos produtos avaliados por ambos u_a e u_b.

Outras medidas também são usuais (ex: medida cosseno).

Como definir quem são os usuários mais próximos a j?

Devemos medir a similaridade entre as notas atribuídas pelo usuário j a cada um dos produtos com as notas atribuídas pelos outros usuários.

Uma medida usual é a correlação de Pearson:

$$sim(u_a, u_b) = cor(R^*_{u_a,.}, R^*_{u_b,.}),$$

onde $R^*_{u_a,.}$ é o vetor de notas atribuídas pelo usuário u_a aos produtos avaliados por ambos u_a e u_b.

Outras medidas também são usuais (ex: medida cosseno).

Como definir quem são os usuários mais próximos a j?

Devemos medir a similaridade entre as notas atribuídas pelo usuário j a cada um dos produtos com as notas atribuídas pelos outros usuários.

Uma medida usual é a correlação de Pearson:

$$sim(u_a, u_b) = cor(R^*_{u_a,.}, R^*_{u_b,.}),$$

onde $R^*_{u_a,.}$ é o vetor de notas atribuídas pelo usuário u_a aos produtos avaliados por ambos u_a e u_b.

Outras medidas também são usuais (ex: medida cosseno).

Como definir quem são os usuários mais próximos a j?

Devemos medir a similaridade entre as notas atribuídas pelo usuário j a cada um dos produtos com as notas atribuídas pelos outros usuários.

Uma medida usual é a correlação de Pearson:

$$sim(u_a, u_b) = cor(R^*_{u_a,.}, R^*_{u_b,.}),$$

onde $R^*_{u_a,.}$ é o vetor de notas atribuídas pelo usuário u_a aos produtos avaliados por ambos u_a e u_b.

Outras medidas também são usuais (ex: medida cosseno).

Método 2: Filtro colaborativo baseado no produto

Queremos estimar a nota que um usuário j dará a um produto l.

Ideia: primeiro procuramos os k produtos mais parecido a l. Calculamos então as médias das notas dadas por esse usuário para cada um desses k produtos (parecido com o KNN também).

$$\widehat{R}_{j,l} = \frac{1}{k} \sum_{s \in \mathcal{V}_l} R_{j,s},$$

onde $\mathcal{V}_l$ é o conjunto dos k vizinhos mais próximos ao produto l.

Método 2: Filtro colaborativo baseado no produto

Queremos estimar a nota que um usuário j dará a um produto l.

Ideia: primeiro procuramos os k produtos mais parecido a l. Calculamos então as médias das notas dadas por esse usuário para cada um desses k produtos (parecido com o KNN também).

$$\widehat{R}_{j,l} = \frac{1}{k} \sum_{s \in \mathcal{V}_l} R_{j,s},$$

onde $\mathcal{V}_l$ é o conjunto dos k vizinhos mais próximos ao produto l.

Como definir quem são os produtos mais próximos a I?

Devemos medir a similaridade entre as notas atribuídas por cada um dos usuários ao produto I com as notas atribuídas por cada um dos usuários aos outros produtos.

Uma medida usual é a correlação de Pearson:

$$sim(i_a, i_b) = cor(R^*_{.,i_a}, R^*_{.,i_b}),$$

onde $R^*_{.,i_a}$ é o vetor de notas atribuídas para o produto i_a por cada um dos usuários que avaliaram i_a e i_b.

Outras medidas também são usuais (ex: medida cosseno).

Como definir quem são os produtos mais próximos a I?

Devemos medir a similaridade entre as notas atribuídas por cada um dos usuários ao produto I com as notas atribuídas por cada um dos usuários aos outros produtos.

Uma medida usual é a correlação de Pearson:

$$sim(i_a, i_b) = cor(R^*_{.,i_a}, R^*_{.,i_b}),$$

onde $R^*_{.,i_a}$ é o vetor de notas atribuídas para o produto i_a por cada um dos usuários que avaliaram i_a e i_b.

Outras medidas também são usuais (ex: medida cosseno).

Como definir quem são os produtos mais próximos a I?

Devemos medir a similaridade entre as notas atribuídas por cada um dos usuários ao produto I com as notas atribuídas por cada um dos usuários aos outros produtos.

Uma medida usual é a correlação de Pearson:

$$sim(i_a, i_b) = cor(R^*_{.,i_a}, R^*_{.,i_b}),$$

onde $R^*_{.,i_a}$ é o vetor de notas atribuídas para o produto i_a por cada um dos usuários que avaliaram i_a e i_b.

Outras medidas também são usuais (ex: medida cosseno).

Como definir quem são os produtos mais próximos a I?

Devemos medir a similaridade entre as notas atribuídas por cada um dos usuários ao produto I com as notas atribuídas por cada um dos usuários aos outros produtos.

Uma medida usual é a correlação de Pearson:

$$sim(i_a, i_b) = cor(R^*_{.,i_a}, R^*_{.,i_b}),$$

onde $R^*_{.,i_a}$ é o vetor de notas atribuídas para o produto i_a por cada um dos usuários que avaliaram i_a e i_b.

Outras medidas também são usuais (ex: medida cosseno).

Normalização das notas

É comum se observar um viés nas notas. Alguns usuários tendem a dar notas muito altas para todos os produtos; outros tendem a dar notas muito baixas.

Para tentar levar isso em conta, é comum renormalizar as notas de cada usuário. Uma forma de se fazer isso é trabalhar com as avaliações dadas por

$$\widetilde{R}_{u,i} = R_{u,i} - \bar{R}_u,$$

onde $\bar{R}_u$ é a média das notas dada pelo usuário u.

Normalização das notas

É comum se observar um viés nas notas. Alguns usuários tendem a dar notas muito altas para todos os produtos; outros tendem a dar notas muito baixas.

Para tentar levar isso em conta, é comum renormalizar as notas de cada usuário. Uma forma de se fazer isso é trabalhar com as avaliações dadas por

$$\widetilde{R}_{u,i} = R_{u,i} - \bar{R}_u,$$

onde $\bar{R}_u$ é a média das notas dada pelo usuário u.

Avaliando a precisão dos métodos

Validação cruzada!! ☺

Permite também escolher parâmetros dos modelos usados (ex: k no filtro colaborativo).

Avaliando a precisão dos métodos

Validação cruzada!! ☺

Permite também escolher parâmetros dos modelos usados (ex: k no filtro colaborativo).

Avaliando a precisão dos métodos

Validação cruzada!! ⌣̈

Permite também escolher parâmetros dos modelos usados (ex: k no filtro colaborativo).

Como medir quão razoável um modelo é?

Separar os usuários em treinamento e validação.

Notação: $\mathcal{K}$ — conjunto todos os pares usuário/produto para o qual a nota real $R_{i,j}$ é conhecida, mas não foi usada para treinar o modelo.

Várias medidas podem ser usadas.

Como medir quão razoável um modelo é?

Separar os usuários em treinamento e validação.

Notação: $\mathcal{K}$ — conjunto todos os pares usuário/produto para o qual a nota real $R_{i,j}$ é conhecida, mas não foi usada para treinar o modelo.

Várias medidas podem ser usadas.

Como medir quão razoável um modelo é?

Separar os usuários em treinamento e validação.

Notação: $\mathcal{K}$ — conjunto todos os pares usuário/produto para o qual a nota real $R_{i,j}$ é conhecida, mas não foi usada para treinar o modelo.

Várias medidas podem ser usadas.

Como medir quão razoável um modelo é?

Separar os usuários em treinamento e validação.

Notação: $\mathcal{K}$ — conjunto todos os pares usuário/produto para o qual a nota real $R_{i,j}$ é conhecida, mas não foi usada para treinar o modelo.

Várias medidas podem ser usadas.

Algumas delas:

- EQM: $\frac{\sum_{(i,j)\in\mathcal{K}}(R_{i,j} - \widehat{R}_{i,j})^2}{|\mathcal{K}|}$
- RMSE: $\sqrt{EQM}$
- MAE: $\frac{\sum_{(i,j)\in\mathcal{K}}|R_{i,j} - \widehat{R}_{i,j}|}{|\mathcal{K}|}$

Estas medidas avaliam se as notas estão bem preditas.

Podemos ter interesse apenas em medir se o método está fornecendo boas recomendações.

Algumas delas:

- EQM: $\frac{\sum_{(i,j)\in\mathcal{K}}(R_{i,j} - \widehat{R}_{i,j})^2}{|\mathcal{K}|}$
- RMSE: $\sqrt{EQM}$
- MAE: $\frac{\sum_{(i,j)\in\mathcal{K}}|R_{i,j} - \widehat{R}_{i,j}|}{|\mathcal{K}|}$

Estas medidas avaliam se as notas estão bem preditas.

Podemos ter interesse apenas em medir se o método está fornecendo boas recomendações.

Algumas delas:

- EQM: $\frac{\sum_{(i,j)\in\mathcal{K}}(R_{i,j} - \widehat{R}_{i,j})^2}{|\mathcal{K}|}$
- RMSE: $\sqrt{EQM}$
- MAE: $\frac{\sum_{(i,j)\in\mathcal{K}}|R_{i,j} - \widehat{R}_{i,j}|}{|\mathcal{K}|}$

Estas medidas avaliam se as notas estão bem preditas.

Podemos ter interesse apenas em medir se o método está fornecendo boas recomendações.

N melhores recomendações

Vamos assumir que recomendamos para um dado usuário os N produtos com maior valor predito para a avaliação.

Definimos também o que é uma nota boa e o que é uma nota ruim (ex: nota boa é uma nota maior que 3).

Com base nisso, podemos montar a seguinte tabela para o conjunto de validação:

Matriz de confusão

	Valor Predito	
Valor verdadeiro	Ruim	Bom
Ruim	a	b
Bom	c	d

N melhores recomendações

Vamos assumir que recomendamos para um dado usuário os N produtos com maior valor predito para a avaliação.

Definimos também o que é uma nota boa e o que é uma nota ruim (ex: nota boa é uma nota maior que 3).

Com base nisso, podemos montar a seguinte tabela para o conjunto de validação:

Matriz de confusão

	Valor Predito	
Valor verdadeiro	Ruim	Bom
Ruim	a	b
Bom	c	d

N melhores recomendações

Vamos assumir que recomendamos para um dado usuário os N produtos com maior valor predito para a avaliação.

Definimos também o que é uma nota boa e o que é uma nota ruim (ex: nota boa é uma nota maior que 3).

Com base nisso, podemos montar a seguinte tabela para o conjunto de validação:

Matriz de confusão

	Valor Predito	
Valor verdadeiro	Ruim	Bom
Ruim	a	b
Bom	c	d

Matriz de confusão

	Valor Predito	
Valor verdadeiro	Ruim	Bom
Ruim	a	b
Bom	c	d

Podemos definir as seguintes medidas:

- Acurácia: $\frac{a+d}{a+b+c+d}$
- Precisão: $\frac{d}{b+d}$
- Lembrança: $\frac{d}{c+d}$

Podemos também usar sensibilidade, especificidade etc.

Outros métodos

Há uma literatura muito grande.

Um classe importante é composta por modelos baseados em fatores latentes (SVD). Eles são muito mais rápidos, mas nem sempre muito precisos.

Demonstração no R.