Sistemas de Recomendação

Métricas de Avaliação

Pós-Graduação Lato Sensu

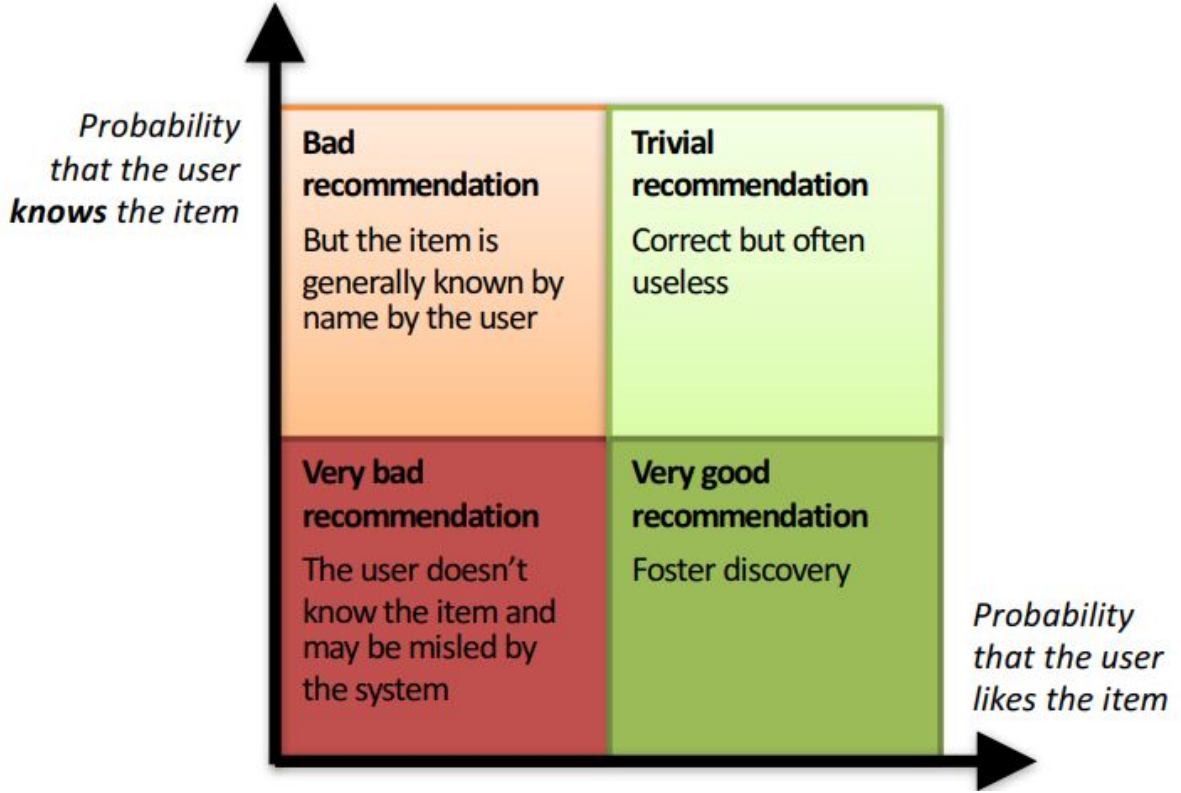
Nícollas Silva

O que é um bom método?

PUC Minas

As melhores recomendações não são as óbvias!

Como avaliar essa qualidade?



Métricas Tradicionais

PUC Minas

- Acurácia e métricas de erro:
  - MAE, MSE, RMSE
- Métricas de Decisão
  - Precision, Recall, MAP
- Acurácia do Ranking
  - Reversões, desempenho inicial
- Métricas voltadas ao usuário
  - Cobertura, retenção de usuários, satisfação

Métricas Comerciais

PUC Minas

- Apenas a acurácia não é suficiente
  - Aumento de vendas, vendas cruzadas, conversões
- O foco não pode estar apenas no usuário
  - Os objetivos do recomendador também importam

Novas Métricas

PUC Minas

- Métricas desenvolvidas para objetivos específicos
  - Métricas de ranking sofisticadas
  - Diversidade e Novidade
  - Serendipidade
- Avaliações holísticas
  - Influência da página do sistema

O que é fundamental para mensurar todas essas métricas?


FEEDBACK

Feedback

PUC Minas

Explícito:

- 1-5 estrelas
- Escasso e difícil de ser coletado
- Possui ruído: itens os usuários **gostam** vs. o que eles **dizem** que gostam.

Implícito:

- Likes, views, compras, etc.
- São abundantes.
- Problemático: como distinguir entre um feedback negativo ou a ausência de feedback?

Feedback

PUC Minas

Offline:

- As avaliações que estão ocultas no conjunto de teste
- Dados que ficam ocultos ao método durante o treinamento.

Online:

- As avaliações que estão acontecendo em tempo real
- Dados fornecidos pelo usuário sobre o que foi apresentado

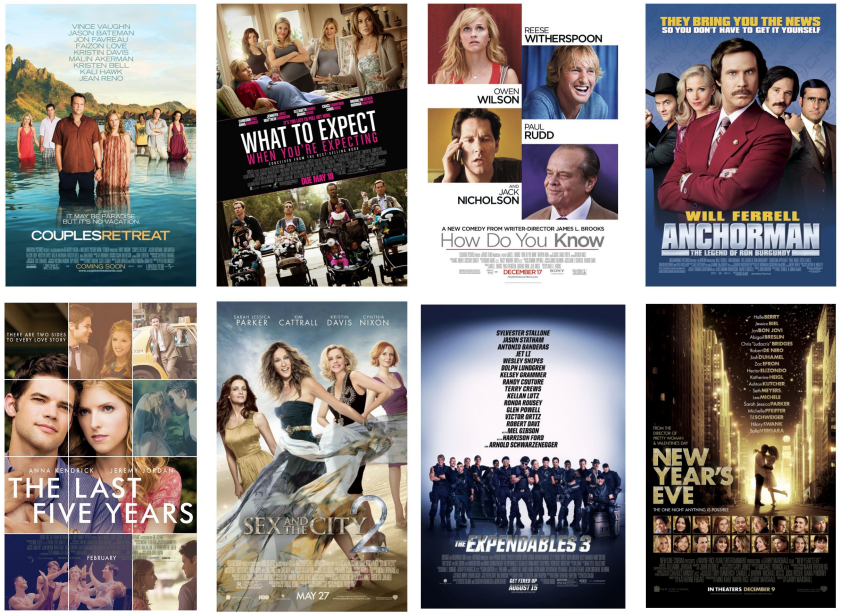
Avaliação

PUC Minas

Predito:

≠

Feedback:



Métricas de Avaliação

PUC Minas

- Acurácia da predição
  - Quão bem o método estima as preferências absolutas?
- Suporte a decisão
  - Quão bem o método retorna as coisas relevantes?
- Acurácia do Ranking
  - Quão bem o método estima as preferências relativas?

Métricas de Acurácia

Objetivo: mensurar a distância dos ratings preditos aos reais

- Ratings reais refletem a preferência desconhecida do sistema
  - *Oculta*: avaliações offline
  - *Realmente desconhecida:* avaliações online
- Tipicamente mensurada por **métricas de erro**
  - e.g., predito = 2.5; real = 3; erro = 0,5

Taxa de Acerto

PUC Minas

Hit-Rate: média do número de itens recomendados que estão no conjunto teste de cada usuário.

- Cada usuário tem uma lista de recomendação R associada
- Para cada usuário, conte quantos itens estão em R e no conjunto teste
- Calcule a média (i.e., some tudo e divida pelo número de usuários)
- Quanto maior o valor, melhor é o recomendador!

MAE: mean absolute error

- Estima o erro como a distância para as preferências atuais

$$\hat{r}_i - r_i$$

- Essa métrica não se importa com a direção do erro

$$\left|\hat{r}_i - r_i\right|$$

- MAE:

$$\frac{1}{n}\sum_{i=1}^{n}\left|\hat{r}_i - r_i\right|$$

MSE: mean squared error

- Similar ao MAE, porém considera o erro quadrado
  - Remove a direção do erro
  - Penaliza erros maiores que os menores
- MSE:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{r}_i - r_i\right)^2$$

- Desvantagem: não tem uma escala intuitiva

RMSE: root mean squared error

PUC Minas

- Consiste na raiz quadrada do MSE
- RMSE:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{r}_i - r_i\right)^2}$$

- Vantagem: mantém a escala dos ratings

Discussões

PUC Minas

- Qual o problema dessas abordagens?
  - Elas consistem de médias sobre todos ratings
- O que aconteceria se um usuário tiver 10k ratings e o outro apenas 10?
  - A avaliação seria enviesada!
  - É preciso analisar também a média por usuário.
- Importante:
  - Não podemos comparar dados com diferentes escalas de ratings
  - Os erros podem ser dominados por partes irrelevantes dos dados

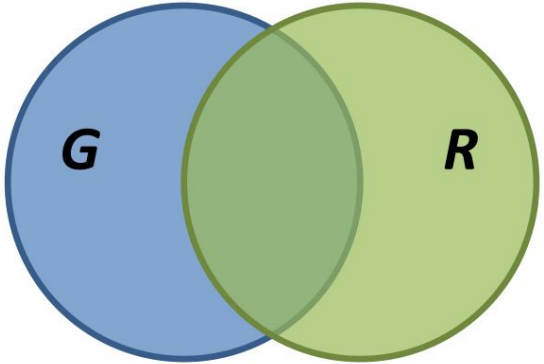
Métricas de Decisão

PUC Minas

Objetivo: medir quão bem o recomendador ajuda o usuário a fazer boas decisões e evitar escolhas ruins.

- O que é bom e o que é ruim?
  - Depende da aplicação
- Em geral:
  - Tarefa de Predição: 4 vs 2.5 é pior que predizer 2.5 vs 1
  - Tarefa de Recomendação: as primeiras posições são fundamentais

Precision & Recall



G: relevant

R: retrieved

Precision:

- Porcentagem de itens retornados que são relevantes

$$Prec = \frac{|G \cap R|}{|R|}$$

Recall:

- Porcentagem de itens relevantes que são retornados

$$Rec = \frac{|G \cap R|}{|G|}$$

Precision & Recall

PUC Minas

Precision:

- Visa ter o maior número de itens relevantes na recomendação.
- Assume que existem mais itens relevantes do que você pode examinar.

Recall:

- Visa não deixar nenhum item relevante fora da recomendação.
- Assume que você tem tempo para filtrar entre as recomendações.

Precision & Recall

PUC Minas

Precision:

- Visa ter o maior número de itens relevantes na recomendação.
- Assume que existem mais itens relevantes do que você pode examinar.

Recall:

- Visa não deixar nenhum item relevante fora da recomendação.
- Assume que você tem tempo para filtrar entre as recomendações.

$$F1 = \frac{2\,Prec\,Rec}{Prec + Rec}$$

Precision & Recall

PUC Minas

Problema 1:

- É difícil cobrir todo o dataset.

Solução:

- Cortes no ranking
  - *Prec@n*
  - *Rec@n*

Problema 2:

- É necessário saber o que é relevante para o usuário.
- Mas, na maioria dos casos nós não sabemos o rating atribuído.

Solução:

- Limitar a análise somente aos itens avaliados.

MAP

- AP (Average Precision): precisão nos itens relevantes

$$\frac{\sum_{i=1}^{n} Prec@k \times rel(i)}{|G|}$$

- MAP (mean AP)

$$\frac{1}{m}\sum_{i=1}^{m} AP(q_i)$$

Discussões

- Todas essas métricas tendem a ser correlacionadas entre si
  - *Precision* e *Recall* são talvez as mais utilizadas (e fáceis de entender)
- Sobretudo, nenhuma dessas métricas superam o problema de serem baseadas apenas nos itens previamente avaliados.
  - E, claro, no ruído que essas avaliações possuem
  - Ainda está em aberto o que fazer com a falta de avaliações

Métricas de Ranking

PUC Minas

Objetivo: medir quão bem o recomendador ajuda o usuário a visualizar as coisas relevantes primeiro que as demais.

- A tarefa de recomendação é uma tarefa de ranking
  - Devemos colocar os itens em ordem de preferência
- Premissa: os usuários vão inspecionar os itens recomendados do topo para o fim da lista.

Cenários Tradicionais

Diferente dos cenários de busca, os rankings estão dispostos na horizontal da interface.

MRR: Mean Reciprocal Rank

- Similar ao *Precision* e *Recall*
  - Prec/Rec mensuram a probabilidade dos itens serem relevantes (precision) e do usuário encontrá-los (recall)
  - RR mensura o 'esforço' do usuário para encontrar os relevantes
- RR = *1/i* , onde *i* é a posição do primeiro item relevante
- MRR: média desse esforço para os usuários

$$\frac{1}{m}\sum_{i=1}^{m} RR(q_i)$$

Coeficientes de Correlação

- Mensuram quão bem as recomendações seguem a ordem correta
- Sp

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

n: número de itens

d: diferença do rank

- Ke

$$\tau = 2\frac{n_c - n_d}{n(n-1)}$$

nc: número de pares concordantes

nd: número de pares discordantes

DCG

- Discounted Cumulative Gain mensura a utilidade do item em cada posição
  - $DCG = \sum_{i=1}^{n} \frac{2^{rel(i)} - 1}{log_2(i+1)}$ sição do item
- Na prática, calcula-se a média do DCG normalizado pelo DCG ordem preferida do usuário)

$$nDCG = \frac{1}{m} \sum_{i=1}^{m} \frac{DCG(q_i)}{iDCG(q_i)}$$

Discussões

- Existem diversas outras métricas para mensurar a habilidade do recomendador em ordenar os itens
  - A maioria vêm do campo de Recuperação de Informação
- nDCG é frequentemente utilizada
  - Assim como o MRR

O foco da Indústria

Na indústria, nós preocupamos em **reter** os usuários no sistema e **fazê-los felizes**; não apenas em melhorar a acurácia das recomendações em 1%.

- Tao Ye, Senior Scientist at Pandora
- RecSys 2015, Industry Panel

Métricas de Business

Grandes empresas estão interessadas em:

- Satisfazer as preferências dos usuários
  - Métricas de Acurácia
  - Métricas de Suporte da Decisão
  - Métricas de Ranking
- Potencializar o interesse do usuário nos seus itens
  - Cobertura
  - Diversidade
  - Serendipidade

Cobertura

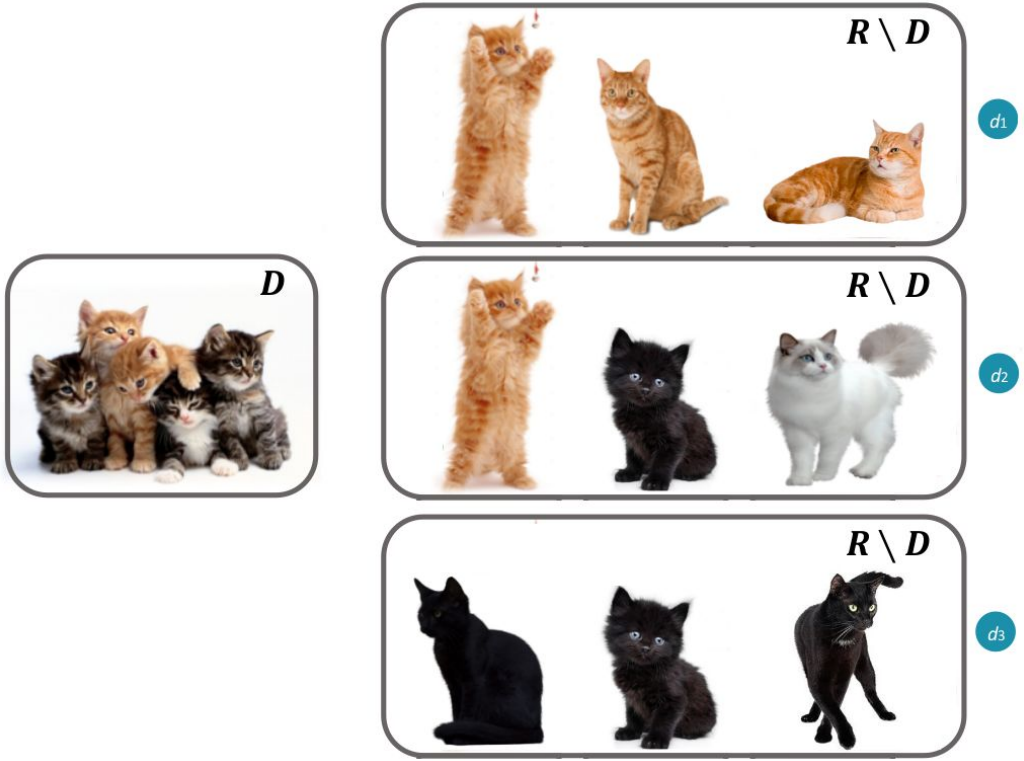
PUC Minas

- Mensura a porcentagem de itens recomendados com relação ao todo
  - Pode ser medida sobre os itens preditos
  - Ou também medida sobre um nível de confiança
    - e.g., quantos filmes 5 estrelas serão recomendados?
- Para o mercado é interessante que a recomendação englobe todo o catálogo de filmes (até mesmo os menos populares)

Diversidade

Mensura o quão diferente os itens recomendados são entre si.

- Aplicado nos top-N itens



Diversidade

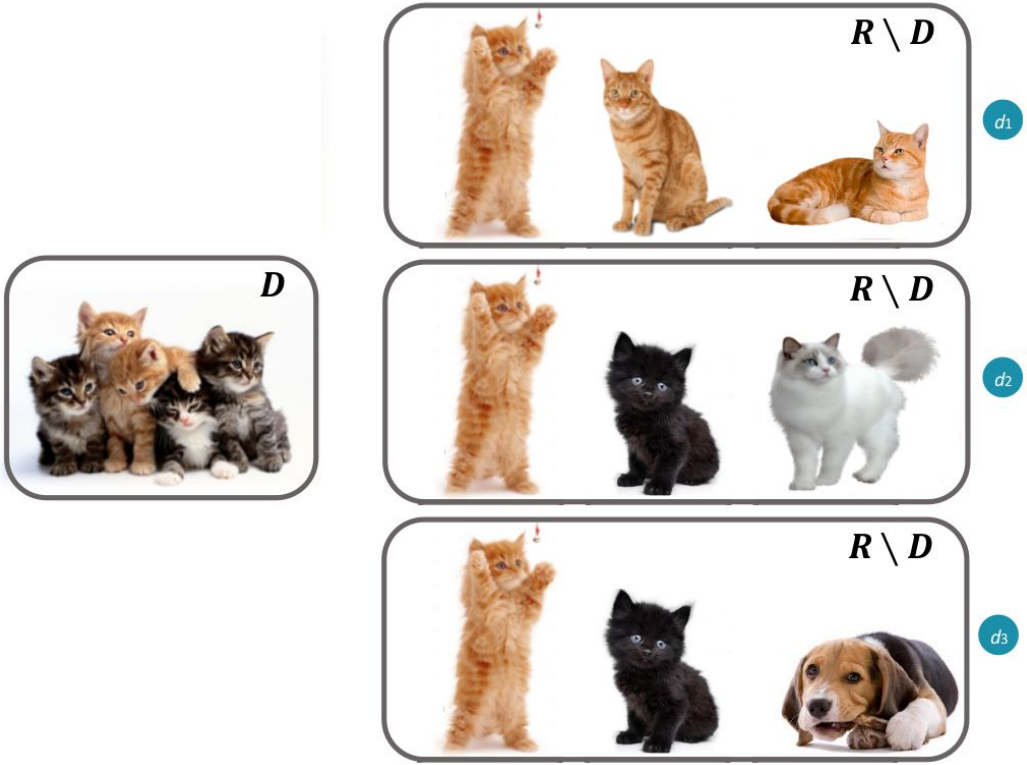
PUC Minas

- Intra-List similarity:
  - Média da similaridade computada par-a-par entre os itens.
  - Define os itens como vetores de usuários e calcula sobre eles.
  - Em alguns casos, baixos valores significam alta diversidade.
- Genre/Category Similarity:
  - Média da similaridade computada par-a-par entre os itens.
  - Define os itens pelas suas categorias (e.g., gênero dos filmes).

Novidade

Mensura o quão diferente os itens recomendados são daqueles que já foram consumidos pelo usuário.

- Aplicado nos top-N itens e nos itens do treinamento.



Novidade

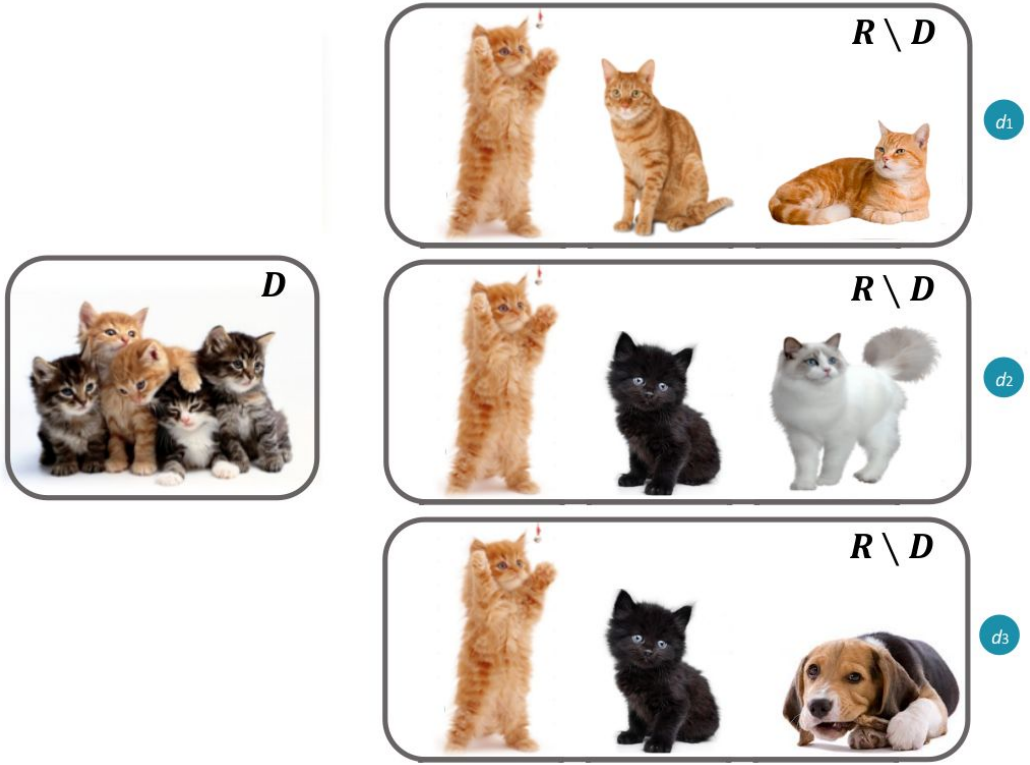
PUC Minas

- EPC:
  - Baseado na similaridade computada par-a-par entre os itens recomendados e os itens avaliados pelo usuário.
  - Define os itens como vetores de usuários e calcula sobre eles.

Serendipidade

Mensura a ocorrência de eventos com chances de surpreender o usuário de uma maneira positiva.

- Surpresa com itens relevantes
- Altamente relacionado a *rarity* e *unexpectedness*



Em suma...

- Existem vários objetivos específicos a serem avaliados.
  - Identifique o propósito da sua recomendação e defina a melhor métrica
- Sobretudo, **não avalie** apenas uma única métrica
  - Interessante verificar o comportamento em distintos pontos de vista