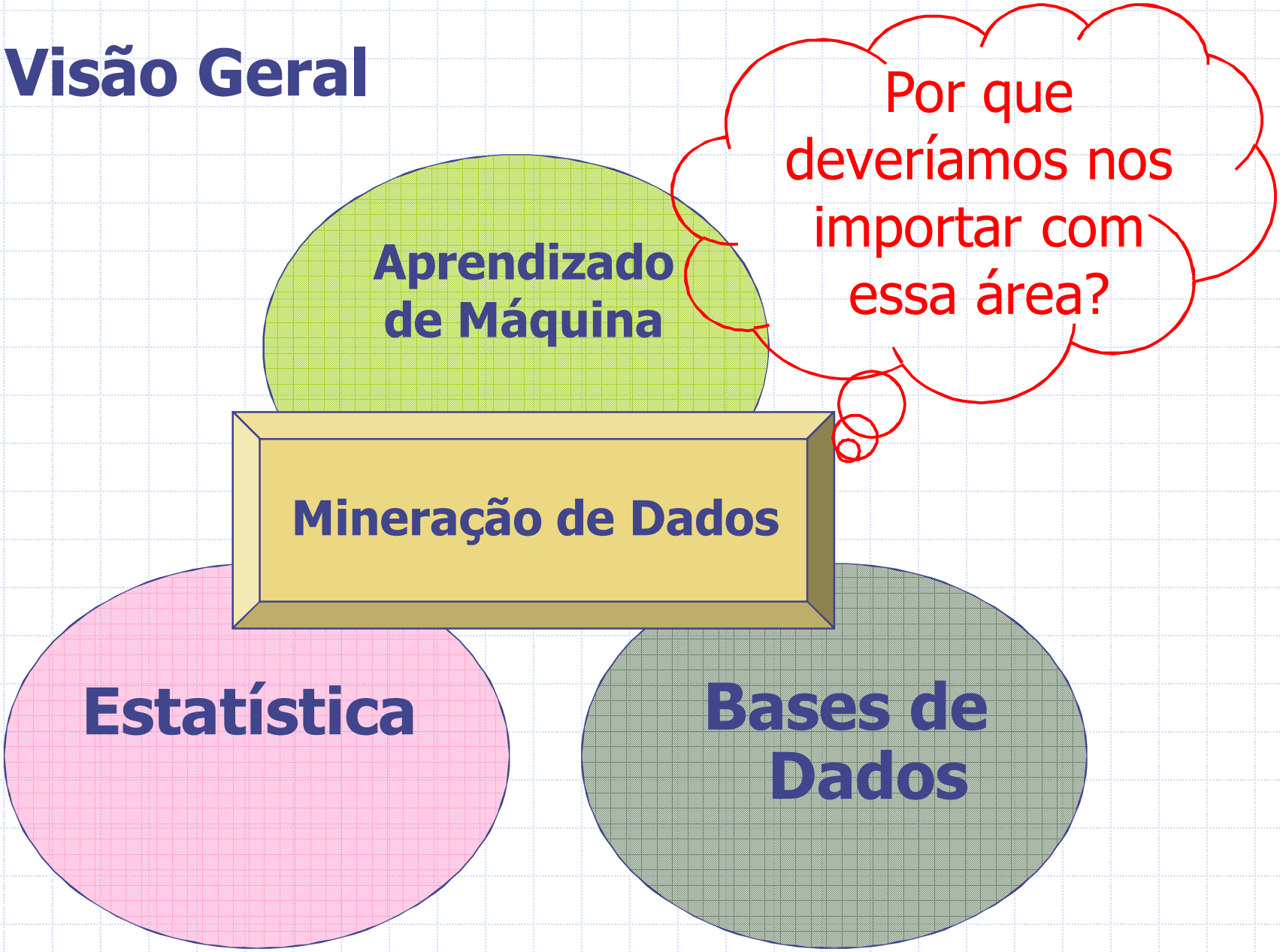
Mineração de Dados: Panorama, Aplicações e Tendências

Eduardo R. Hruschka

Mapa do Território

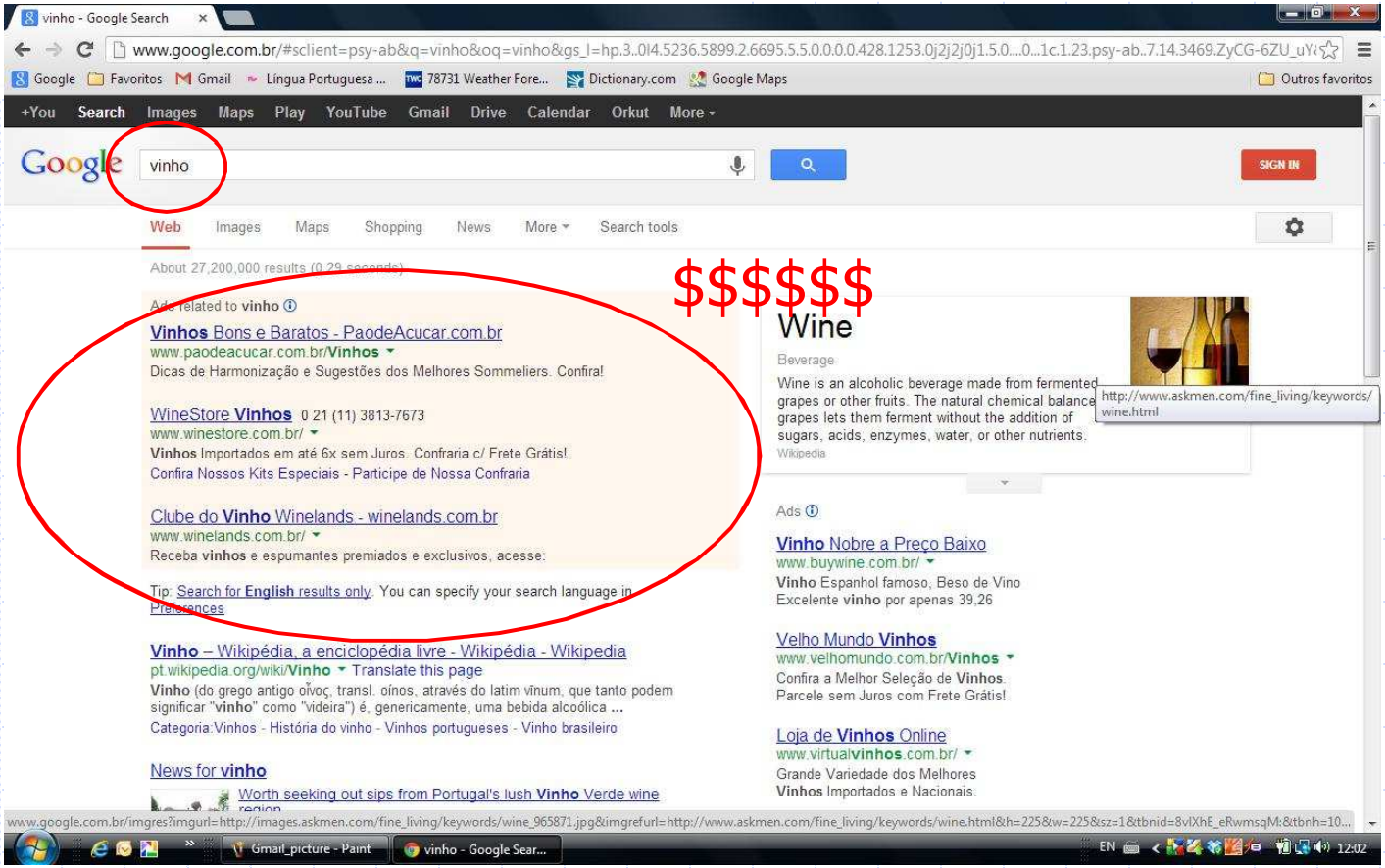
1) Visão Geral sobre Mineração de Dados (*Data Mining, Knowledge Discovery from Databases, Data Science, Predictive Analytics, Big Data*)
2) Agrupamento de Dados ("Clusterização")
3) Classificação
4) Sistemas de Recomendação (Regressão)
5) Tendências Futuras e Impactos na Sociedade

1. Visão Geral



Alguns (poucos) exemplos...

Busca por "a": 25 bilhões de *web pages*



Projeção (2013):

R$ 100 bilhões

Alguns (poucos) exemplos...

- Taxas de *clicks* em torno de 0.05% - R$ 12 bilhões (2013);
- 30% de *smartphones* e *tablets;*
- 1 bilhão de usuários (65 milhões no Brasil), $>10^9$ posts/dia.

Alguns (poucos) exemplos...

Cortesia - Prof. Marko Grobelnik

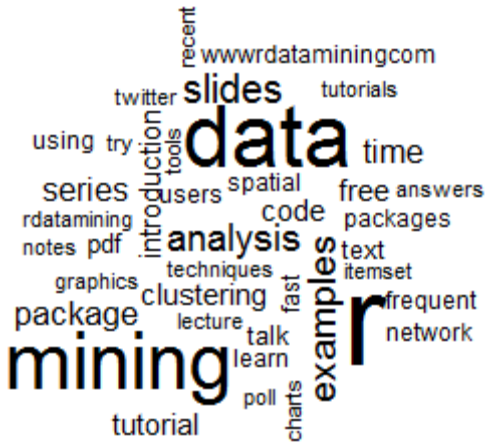
Comunicações entre diferentes usuários:

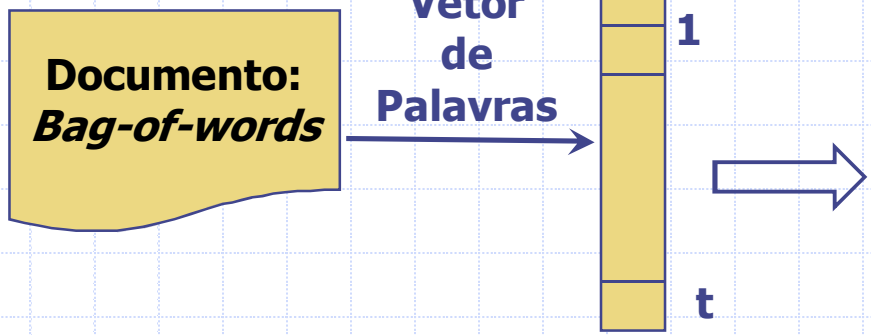
> 100 bilhões de amizades (130 amigos em média);

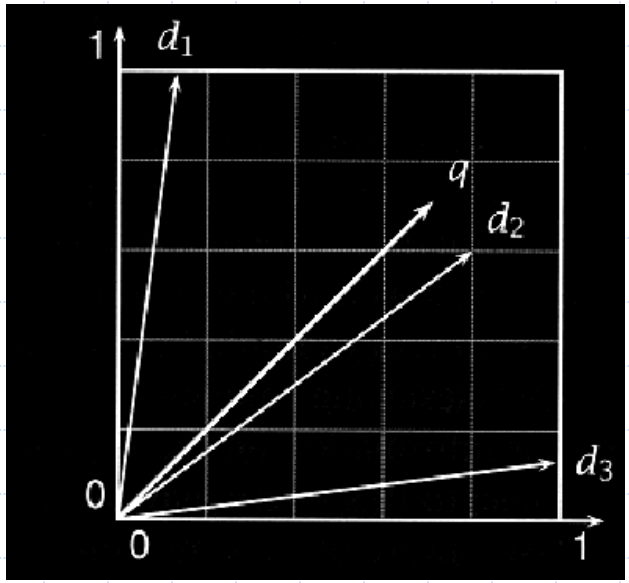
> 2 bilhões de "curtir/comentários" por dia;

Qual é a ideia básica?

- Selecionar palavras de interesse (eliminar preprosições, artigos, etc.);
- Para cada palavra, computar sua frequencia por *texto.*







- Cada texto é representado por um vetor;
- Como encontrar *textos* semelhantes?

Como medir similaridades?

- Em geral, trata-se de um problema difícil:

- Abordagens matemáticas (e.g., cosseno entre dois documentos) são comumente adotadas;
- Técnicas de agrupamento de dados são úteis...

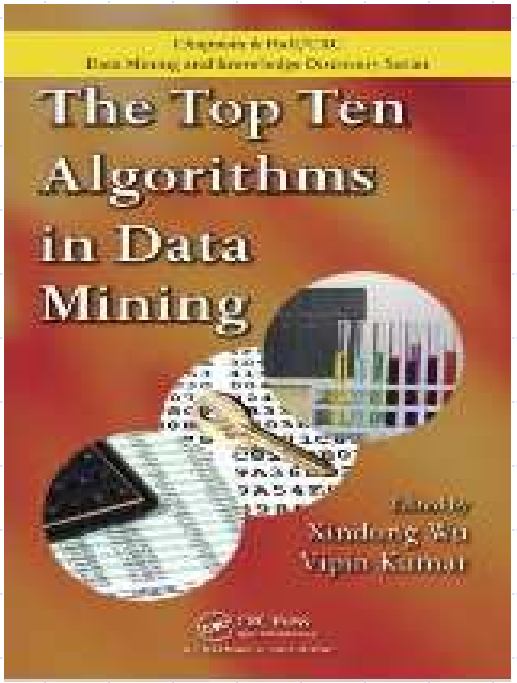
Keogh, E. A Gentle Introduction to Machine Learning and Data Mining for the Database Community, SBBD 2003, Manaus.

2. Agrupamento de Dados

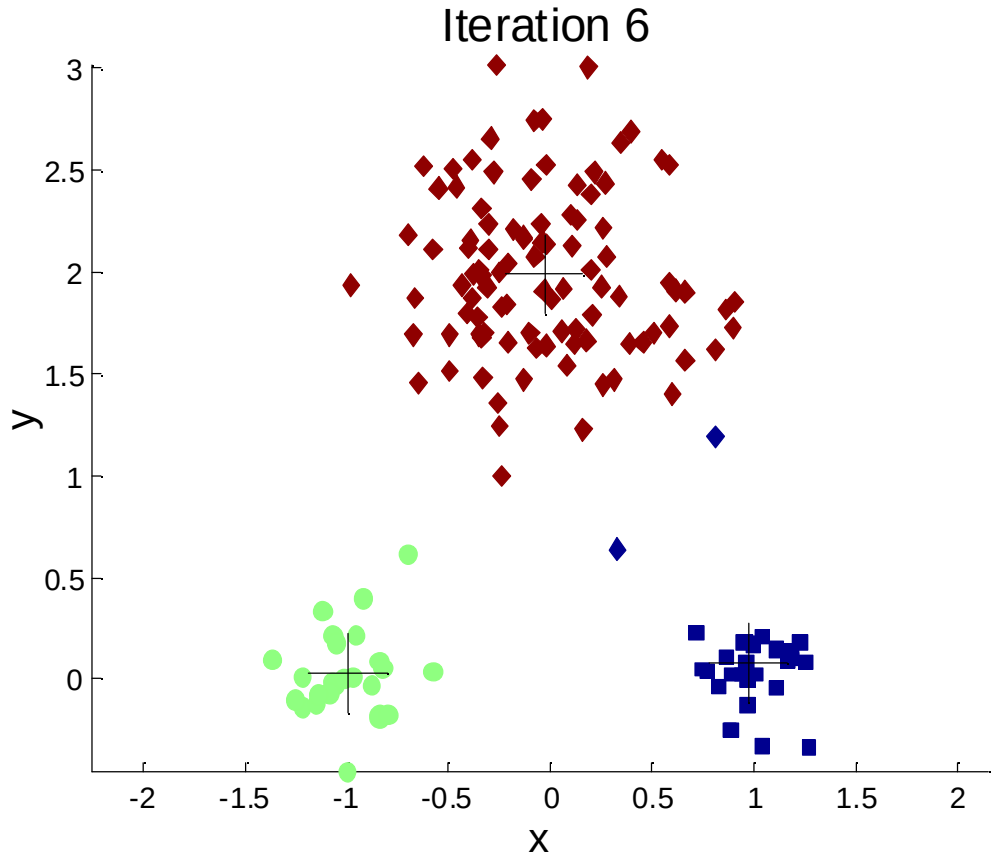
- Encontrar grupos (*clusters*) de dados similares;
- Diversas aplicações reais – análise exploratória de dados: mineração de textos, *marketing* (segmentação de clientes), recuperação de informação, reconhecimento de padrões, ...

Ideia geral e intuitiva por meio de um exemplo pedagógico:

- Algoritmo K-means (MacQueen, 1967; Kulis & Jordan, 2012)

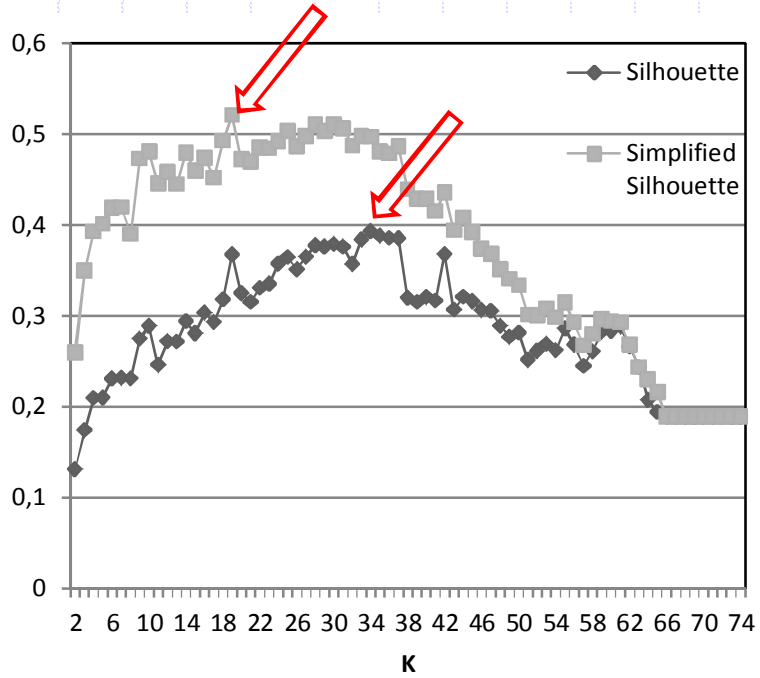


Exemplo para K-means (K=3):



Tan et al., Introduction to Data Mining, 2004.

- Otimização convexa para cada *K* : converge para ótimos locais com diversas medidas de distância...Bacana, mas:
  - Sensível à inicialização;
  - Como estimar *K* a partir dos dados?
- Rodar K-means várias vezes para diferentes valores de K;
- Problema de otimização multi-modal:



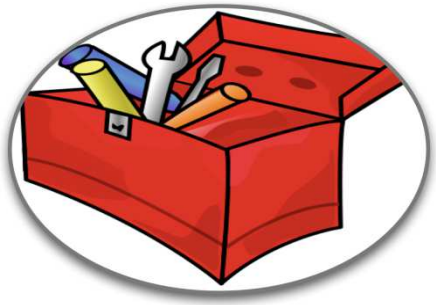
Número de grupos em torno de 20-30;

Agrupamento de documentos para análise forense (Nassif & Hruschka, 2012):

- Computadores apreendidos pela PF;
- Facilitar análise do perito – inspecionar grupos de documentos.

- Tarefa computacionalmente custosa;
- Algoritmos de busca probabilística (EAs, PSO, SA, ...);
- Abordagens que combinam diferentes algoritmos de otimização (Hruschka et al., 2009): busca local + busca global:
  - Idealmente dependentes de poucos parâmetros *não críticos* definidos pelo usuário.
- Extensões para Modelos de Misturas de Gaussianas:
  - *Expectation-Maximization* (EM) + complexidade do modelo.
- Modelos Gráficos Probabilísticos (Blei, 2012).

- Muitos algoritmos disponíveis;
- Caixa de ferramentas (compacta)?



- K-means, Bisecting K-Means, K-medoids;
- Índices para estimar K (e.g., silhueta);
- *Cluster ensembles*;
- Expectation-Maximization (EM) para modelos de misturas de Gaussianas.

3. Classificação

- Fraude: financeira, comércio eletrônico, seguros, ...
- Resulta em perdas de bilhões de Reais por ano;
- Como detectar automaticamente?

Prevenção de fraude em tempo real: esta transação é fraudulenta?

a) Senha (ajuda em alguns casos)

b) Sistema gera um escore baseado em fatores que qualificam fraude

c) Poucos segundos para tomar decisão

→ Erros de classificação custam caro

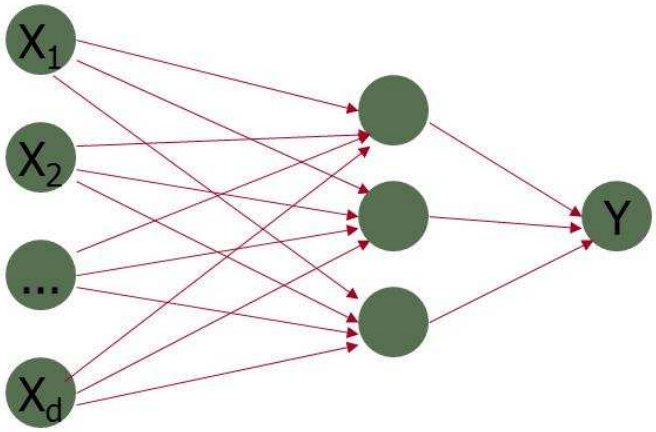
→ Requer modelos estatísticos

- Construindo classificadores automáticos $Y=f(X_1, X_2,...,X_d)$:

X_1	X_2	...	X_d	Y (classe)
...	...	...	...	...
...	...	...	...	...

$Y \in \{fraude, normal\}$ (menos do que 1% de transações fraudulentas)
$X = \{X_1, X_2,...,X_d\}$: variáveis descrevendo as transações

- Diversos modelos – *e.g.*, redes neurais:



Passos principais:

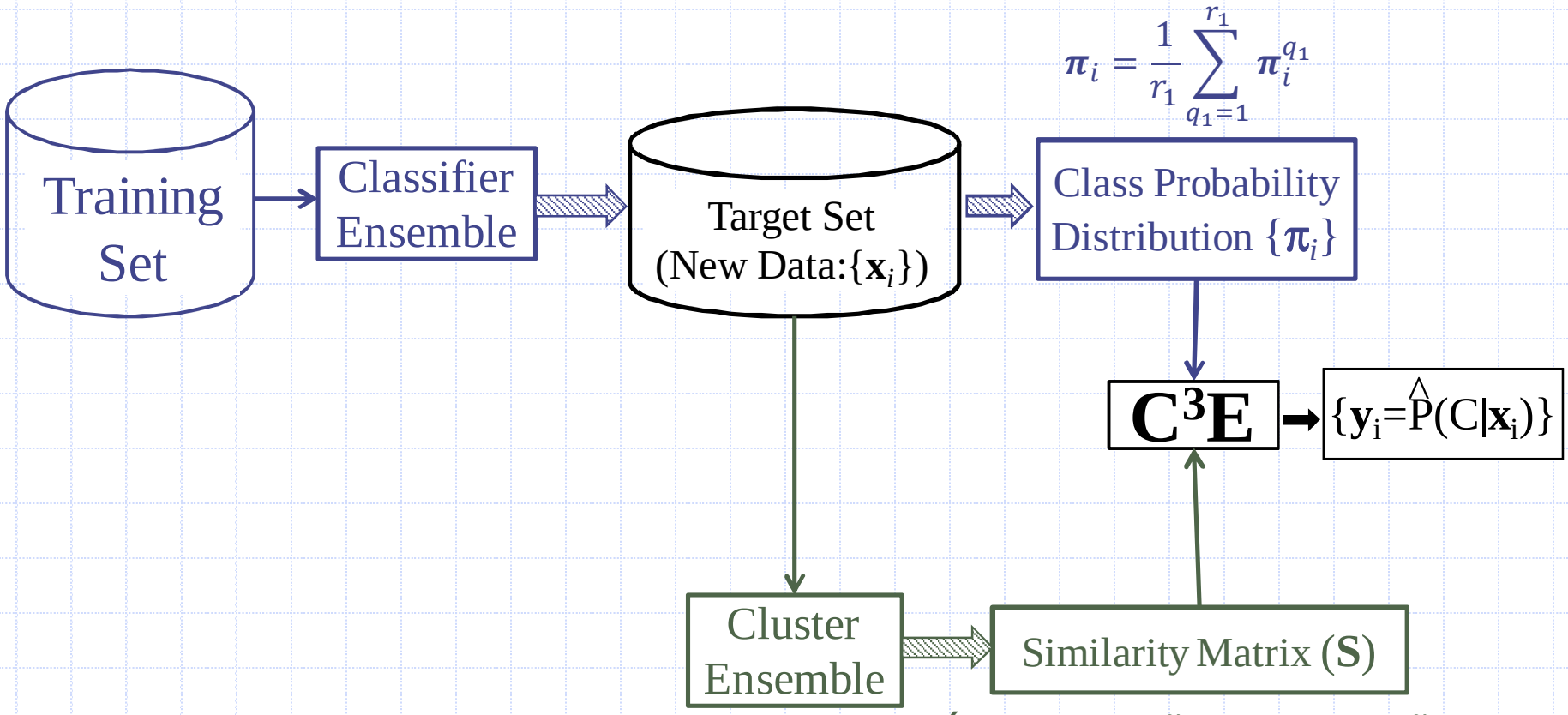
1) Aprender/ajustar os parâmetros do modelo a partir de uma amostra de transações (algoritmos de otimização);
2) Predição de classes (Y) para novas transações baseando-se em $\{X_1, X_2,...,X_d\}$

- Na prática, usualmente vários modelos diferentes são combinados (*classifier ensembles*) tal como se formassem um comitê de especialistas;
- Premissa: dados são *independent and identically distributed* (i.i.d.);
- Violada no mundo real: amostras são polarizadas (formadas apenas por aqueles que foram *pegos*);
- Distribuições de probabilidades de classe podem mudar (entre treinamento e uso dos modelos);

➢ Visão geral de um método que permite refinar distribuições de probabilidade de classe.

- Combinar agregadores de classificadores e agrupadores:
  - Modelos não supervisionados podem fornecer restrições para classificação de novos dados:
    - ✓ Objetos semelhantes no conjunto alvo provavelmente são da mesma classe.
  - Melhorar acurácia preditiva, especialmente quando dados rotulados para treinamento são escassos.
  - Projetar métodos de aprendizado de máquina que sejam conscientes das possíveis diferenças entre distribuições de treinamento e do conjunto alvo.

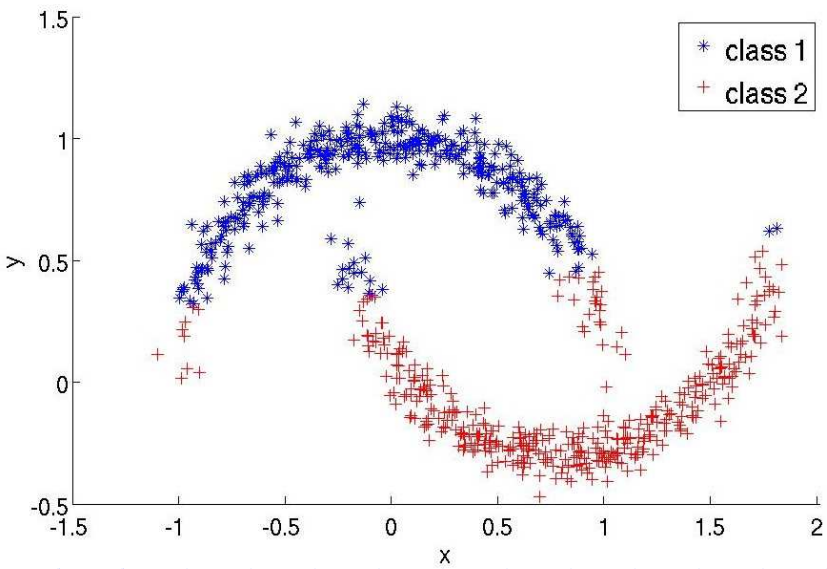
Combining Classifier and Cluster Ensembles (**C³E**):



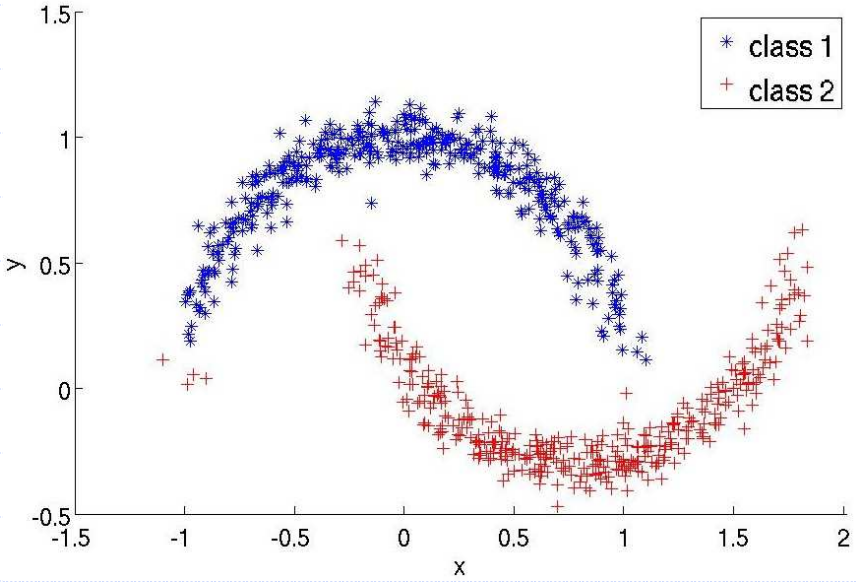
(e.g., proporção das r_2 partições nas quais 2 objetos (transações) pertencem ao mesmo grupo (cluster)

Algoritmo de otimização re-estima probabilidades de classes via similaridades entre novos objetos.

Exemplo pedagógico (somente 2% de objetos rotulados):



Classifier Ensemble



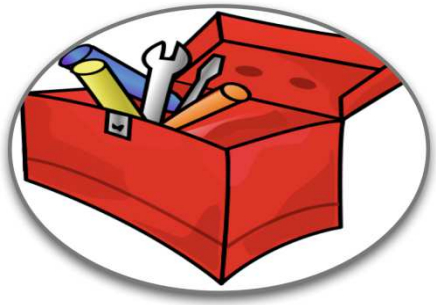
C³E

- Aplicações de sucesso em categorização de textos e classificação de imagens;
- Outras potenciais aplicações ?

- Outras aplicações de classificação incluem:

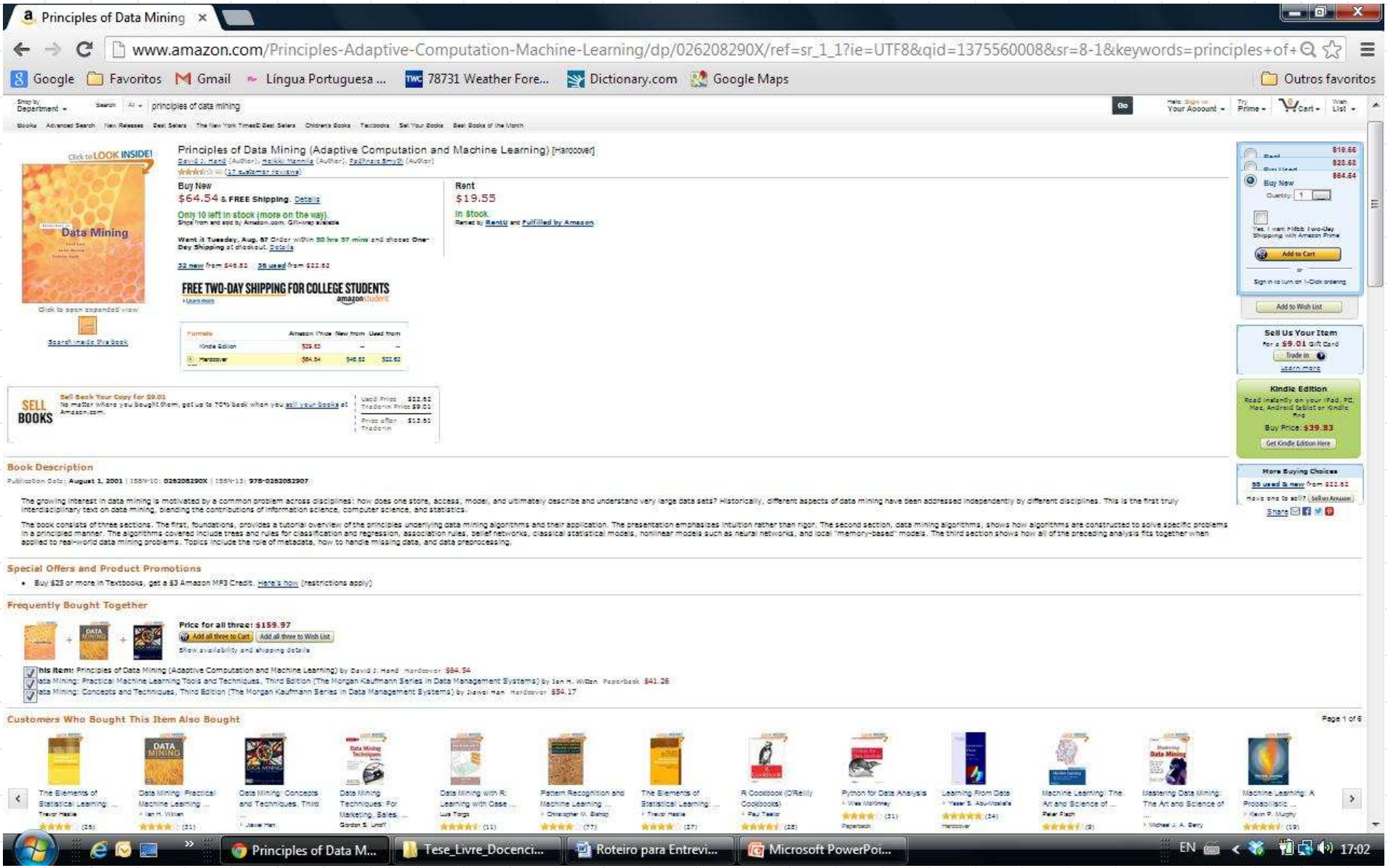
  - *Churn prediction* (cliente abandona serviço/produto): telefonia, tv a cabo, assinatura de revista, cartão de crédito, etc.
  - Classificação de spam;
  - Recrutamento de profissionais (RH);
  - Abandono de emprego;
  - Análise de sentimentos (redes sociais);
  - Finanças (cliente irá cumprir contrato de financiamento?, cliente irá atrasar o pagamento do cartão de crédito?)
  - Bioinformática, Medicina, etc...
    - Requer dados de boa qualidade;
    - Diferentemente do trabalho típico de um estatístico mais tradicional, dados não foram especificamente coletados com o propósito de modelagem.

- Muitos algoritmos disponíveis;
- Caixa de ferramentas (compacta)?



- Naïve Bayes (*wrapper*);
- Árvores de Decisão e *Random Forests*;
- Regressão Logística;
- *Classifier Ensembles*;
- Engenharia de atributos (feature selection);
  - SVMs, redes neurais, etc.

4. Sistemas de Recomendação



4. Sistemas de Recomendação ...

Suposto usuário

➢ Agrupamento e Regressão Simultâneos

- Em problemas difíceis de classificação/regressão, frequentemente se segmenta a base de dados em grupos relativamente homogêneos;
- Posteriormente, constrói-se um modelo por grupo;
- Tal procedimento usualmente proporciona bons resultados com modelos mais simples e interpretáveis;
- SCOAL: *Simultaneous CO-clustering and Learning* (Deodhar and Ghosh, 2010);
- Permite modelagem preditiva de dados em grande escala;
- Aprender modelos locais a partir dos grupos:

4. Sistemas de Recomendação ...

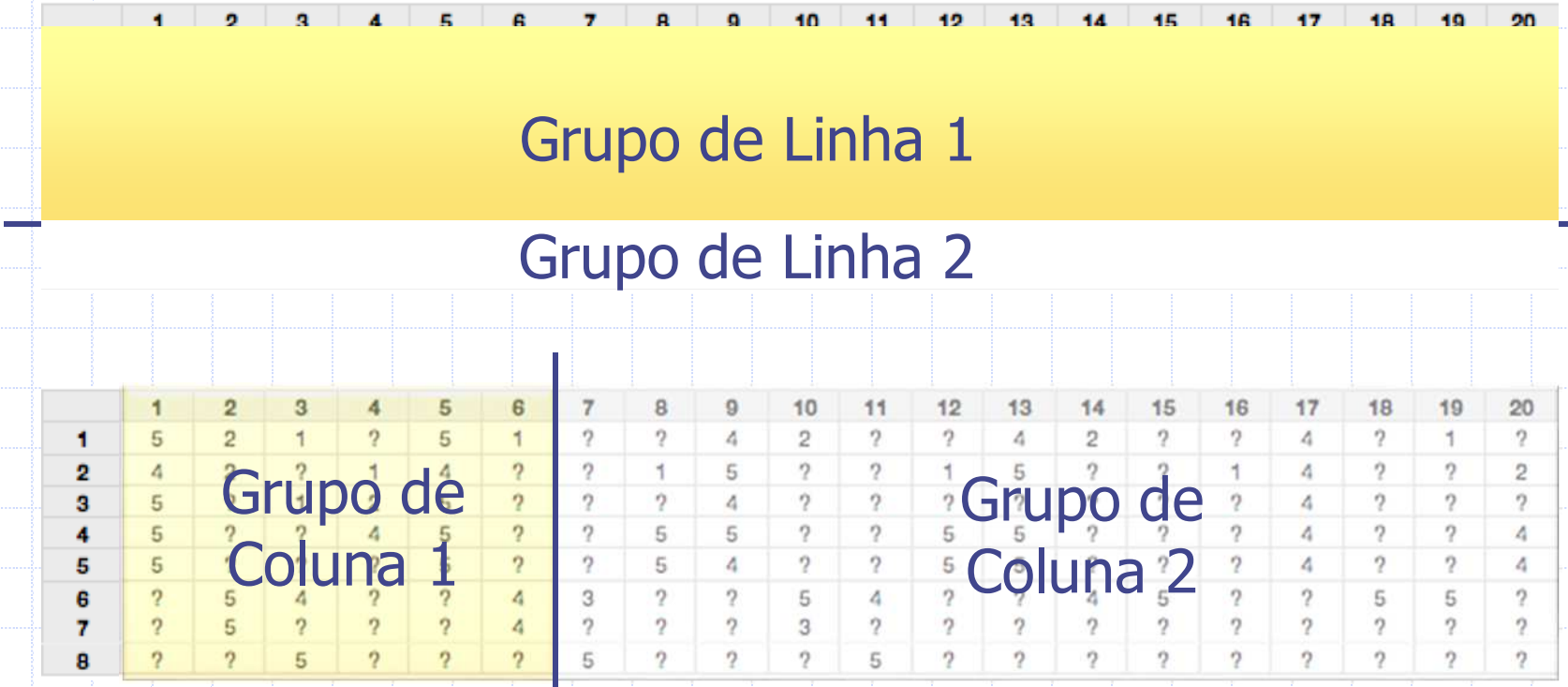
Dados de filmes

Dados de usuários

		95- ação/aventura/thriller	94- comédia	95- policial/drama/thriller	89- romance\comédia	95- ação	94- comédia	95- policial/thriller	97- romance/drama	94- ação/thriller	94- comédia/romance	95- policial/thriller	95- romance/drama	95- ação/drama\thriller	88- comédia	94- policial/drama/rom./thriller	95- comédia/romance	96- ação/thriller	79- comédia	95- policial	96- romance/drama
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
24 anos - M - técnico	**1**	5	2	1	?	5	1	?	?	4	2	?	?	4	2	?	?	4	?	1	?
53 anos - F - outros	**2**	4	2	?	1	4	?	?	1	5	?	?	1	5	?	?	1	4	?	?	2
23 anos - M - escritor	**3**	5	?	1	2	5	?	?	?	4	?	?	?	?	?	?	?	4	?	?	?
24 anos - M - técnico	**4**	5	?	?	4	5	?	?	5	5	?	?	5	5	?	?	?	4	?	?	4
33 anos - F - outros	**5**	5	?	?	?	5	?	?	5	4	?	?	5	5	?	?	?	4	?	?	4
42 anos - M - executivo	**6**	?	5	4	?	?	4	3	?	?	5	4	?	?	4	5	?	?	5	5	?
57 anos - M - admin	**7**	?	5	?	?	?	4	?	?	?	3	?	?	?	?	?	?	?	?	?	?
36 anos - M - admin	**8**	?	?	5	?	?	?	5	?	?	?	5	?	?	?	?	?	?	?	?	?
29 anos - M - estudante	**9**	?	?	?	5	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?	5
53 anos - M - advogado	**10**	5	?	?	?	5	?	?	?	?	?	?	?	?	?	?	?	?	?	?	?

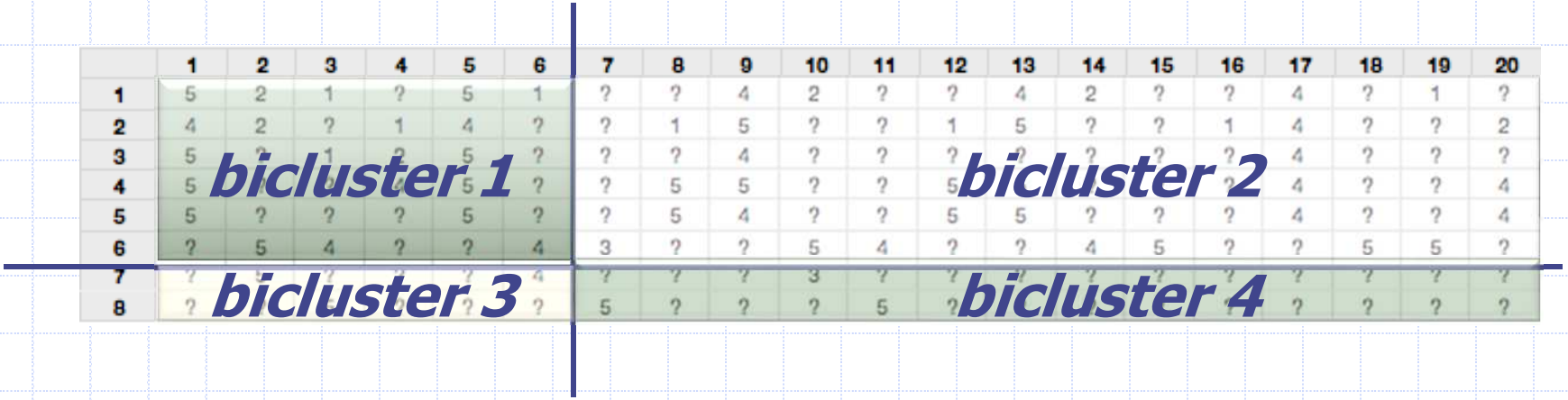
Nota/preferência do usuário pelo item (filme)

Escolher aleatoriamente alguns grupos de linhas e colunas



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	5	2	1	?	5	1	?	?	4	2	?	?	4	2	?	?	4	?	1	?
2	4	?	?	1	4	?	?	1	5	?	?	1	5	?	?	1	4	?	?	2
3	5	?	?	?	?	?	?	?	4	?	?	?	?	?	?	?	4	?	?	?
4	5	?	?	4	5	?	?	5	5	?	?	5	5	?	?	?	4	?	?	4
5	5	?	?	?	5	?	?	5	4	?	?	5	?	?	?	?	4	?	?	4
6	?	5	4	?	?	4	3	?	?	5	4	?	?	4	5	?	?	5	5	?
7	?	5	?	?	?	4	?	?	?	3	?	?	?	?	?	?	?	?	?	?
8	?	?	5	?	?	?	5	?	?	?	5	?	?	?	?	?	?	?	?	?

- Treinar 4 modelos de regressão (um por grupo).



- Um modelo por bicluster:

$$x_{ij} = [1, u, v]$$

$$\beta^T = [\beta^0, \beta_i^T \beta_j^T]$$

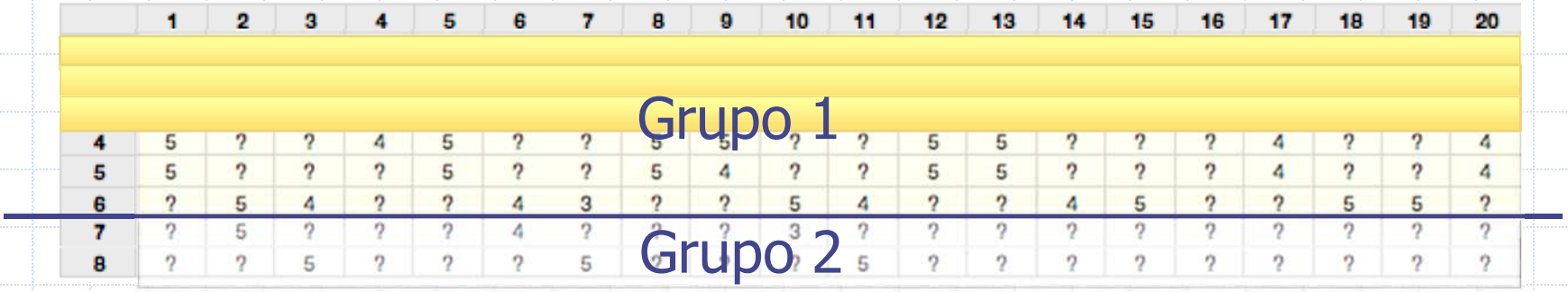
$$Compute: \hat{z}_{ij} = \beta^T x_{ij}$$

$$MSE = \sum_{ij} w_{ij} (z_{ij} - \hat{z}_{ij})^2$$

MSE: 0.2887

Atualizar grupos de linhas:

Mover linhas para grupos que minimizam o erro de predição



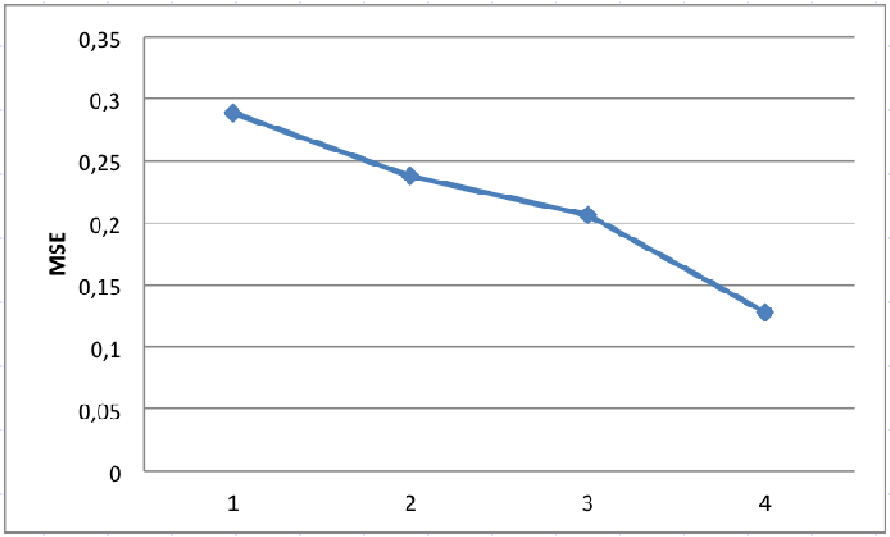
- Linha 3 obtém menor MSE ao ser predita pelos modelos do grupo 2:

- Repetir o processo para cada linha/coluna...

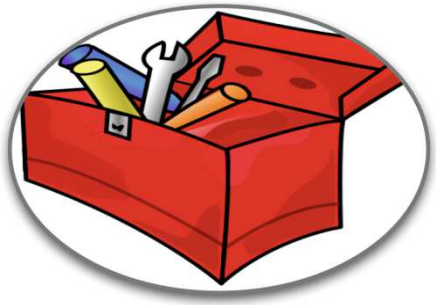
Após o grande laço, re-estimar os modelos:

MSE: 0.23758

- Mover linhas/colunas para os grupos que minimizam o erro;
  - MSE global é (garantidamente) minimizado nas iterações:



- Muitos algoritmos disponíveis;
- Caixa de ferramentas (compacta)?



- Regressão linear;
- Árvores de regressão / *random forests* ;
- Modelos lineares generalizados;
- Outros modelos não lineares (e.g., redes neurais).

5. Tendências e Impactos

- Tendências no projeto de algoritmos para *big data*:
  - Combinar diferentes algoritmos de otimização;
  - Diminuir o número de parâmetros críticos definidos pelo usuário via ajuste automático (a partir dos dados);

  "Essentially, all models are wrong, but some are useful."

  (George E. P. Box, Professor Emeritus, University of Wisconsin)

  - Crescente número de novas aplicações;
  - Gartner (líder em TI) prevê que em 2015 a demanda por profissionais de *big data* será de 4.4 milhões.
  - Questões éticas do uso de modelos automáticos.

Agradecimentos:

USP

Thiago F. Covões, Luiz F. S. Coletta, André P. Vizine

University of Texas at Austin

Joydeep Ghosh, Ayan Acharya, Sreangsu Acharyya

Referências

- Kulis & Jordan, Revisiting k-means: New Algorithms via Bayesian Nonparametrics, Proc. of ICML 2012, Edinburgh, Scotland.
- Nassif & Hruschka, Document Clustering for Forensic Analysis: An Approach for Improving Computer Inspection. IEEE Transactions on Information Forensics and Security, to appear.
- Hruschka et al., A Survey of Evolutionary Algorithms for Clustering. IEEE Transactions on Systems, Man and Cybernetics - Part C: Applications and Reviews, 2009.
- Blei, Probabilistic topic models, Communications of the ACM, 2012.
- Oza & Tumer, Classifier ensembles: Select real-world applications, 2008.
- Wang et al., Bayesian cluster ensembles. Statistical Analysis and Data Mining, 2011.
- Acharya et al., C^3E: A Framework for Combining Ensembles of Classifiers and Clusterers. Proc. of 10th Int. Workshop on Multiple Classifier Systems, 2011.
- Deodhar and Ghosh, SCOAL: A Framework for Simultaneous Co-Clustering and Learning from Complex Data, ACM Transactions on Knowledge Discovery from Data, 2010.