Agrupamento espectral e um experimento educacional

Nicolau L. Werneck

LTI—PCS—USP

Geekie, São Paulo
20 de Novembro de 2012

Resumo e Sumário

O agrupamento espectral, ou *spectral clustering* é uma técnica que permite a classificação não-supervisionada.

Discutiremos a técnica, e um experimento com dados de uma simulação de testes respondidos por alunos.

Referência: von Luxburg [2007].

Sumário:

Algumas definições

Cada amostra $d_i \in \mathbf{R}^c$ possui um conjunto de características, e existe uma função de *similaridade*

$$s_{ij} = f(d_i, d_j), \qquad s_{ij} \in \mathbf{R}.$$

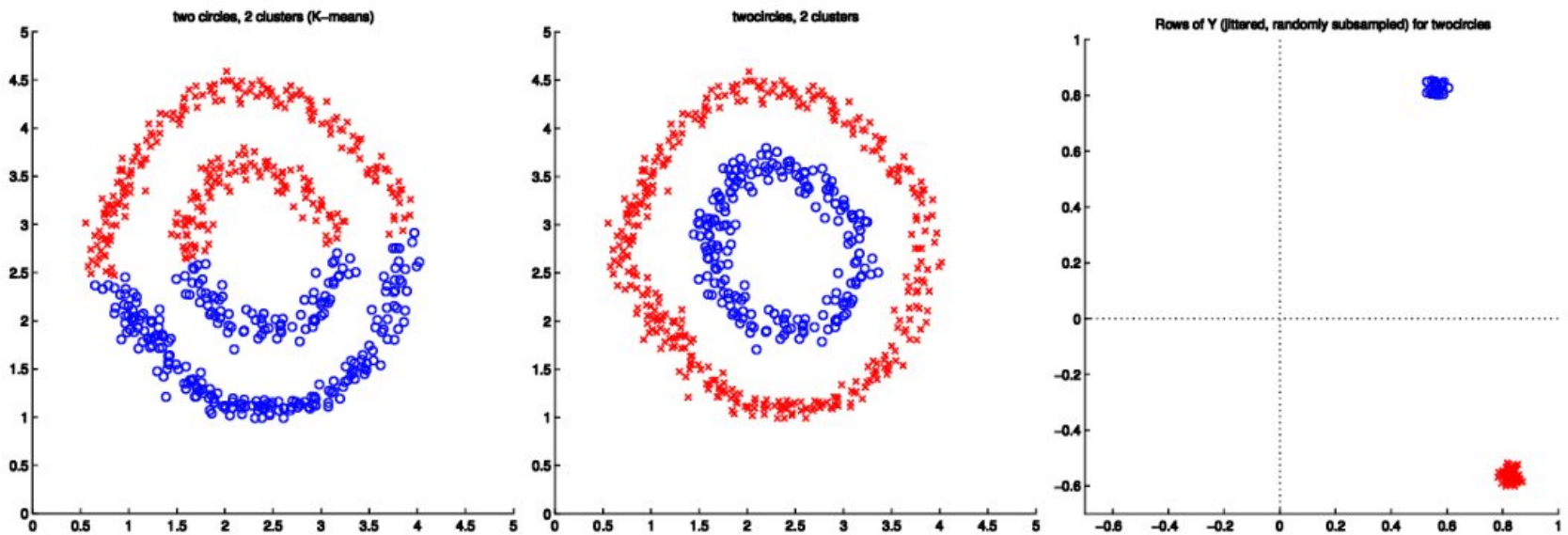
O resultado do método é um mapeamento

$$d_i \to b_i, \qquad b_i \in \mathbf{R}^n.$$

A partir de s_{ij} são produzidos os b_i, permitindo aplicar técnicas de agrupamento por densidade, como k-médias.

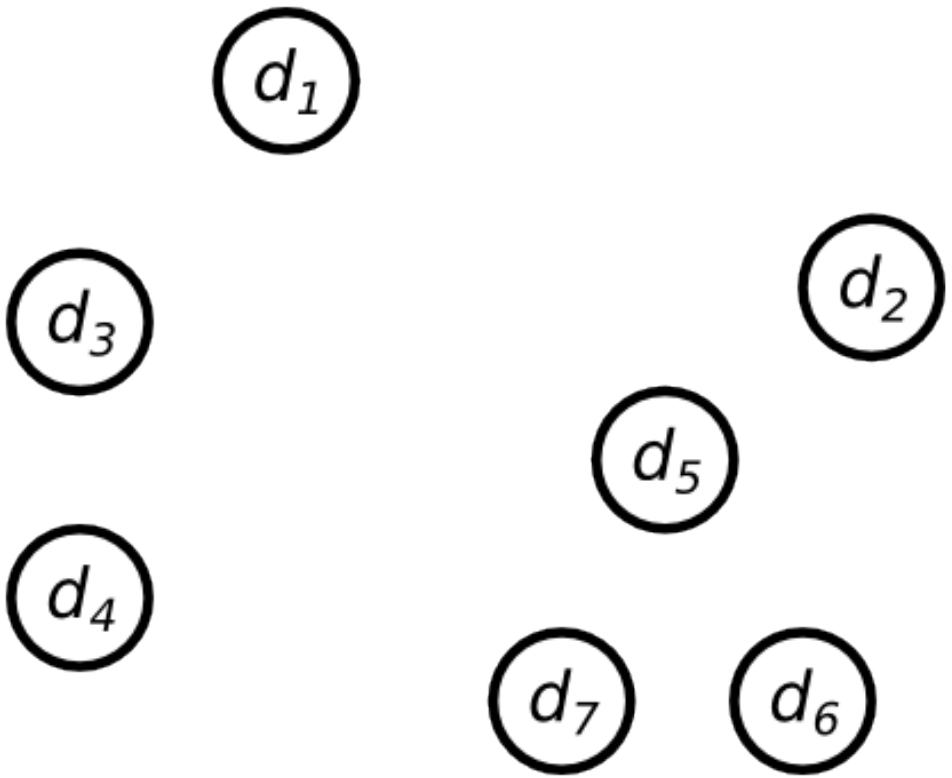
Exemplo

Algoritmos como k-médias não suportam regiões côncavas. A clusterização espectral lida com isto, simplificando a análise.

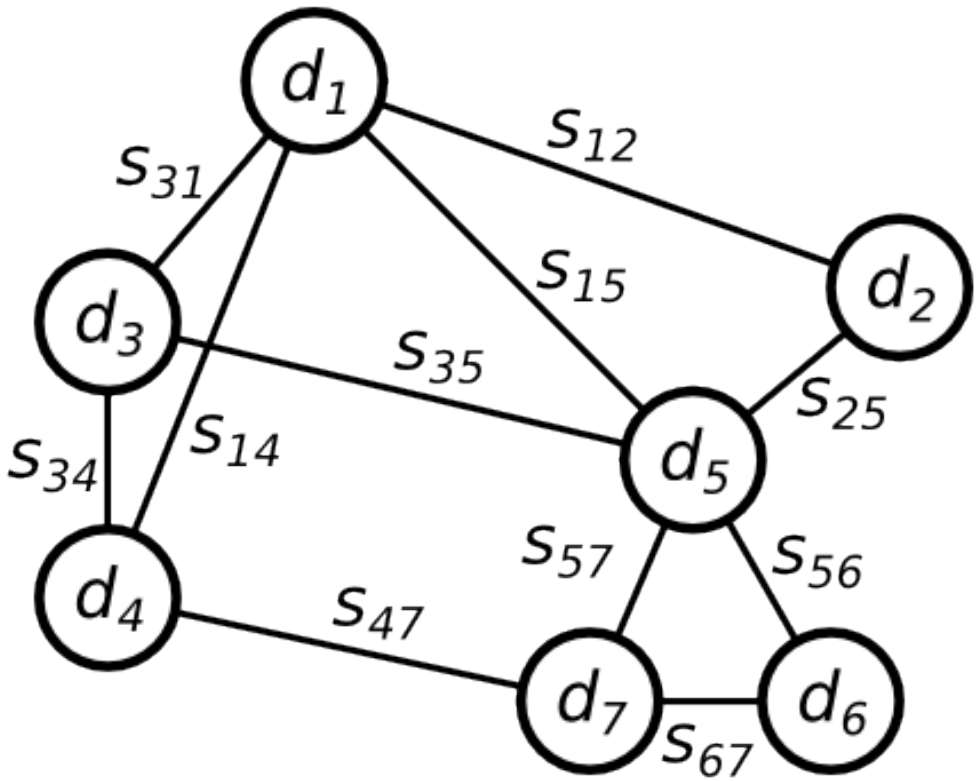


(Figura de Ng et al. [2002].)

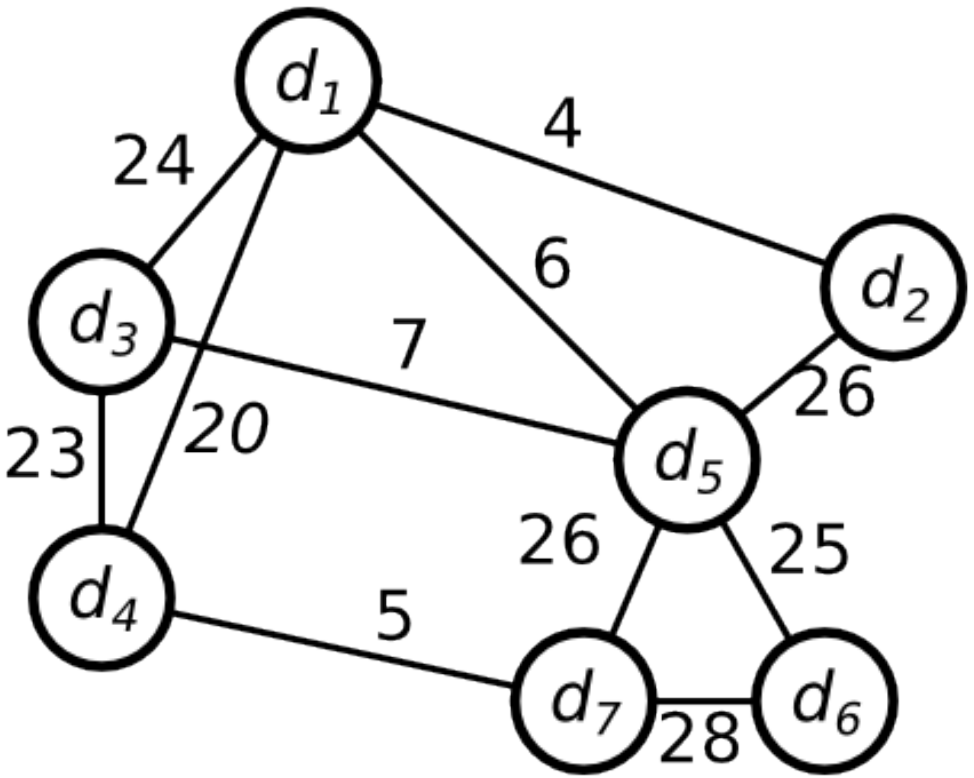
Modelo de grafo



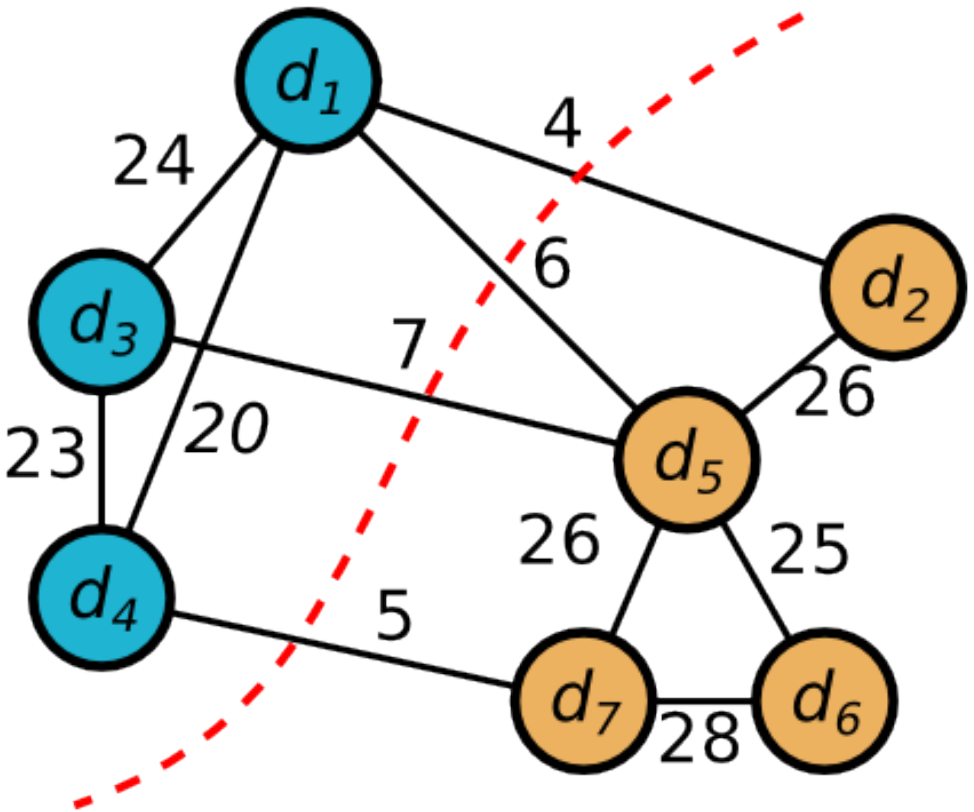
Modelo de grafo



Modelo de grafo



Modelo de grafo



Interpretações

- Corte de grafo — MinCut, RatioCut, etc.
- Cadeia de Markov. *(PageRank)*
- Teoria de perturbação.

Algoritmo

1. Determinar vértices vizinhos.
2. Montar matriz com valores de similaridade.
3. Calcular matriz Laplaciana.
4. Encontrar menores autovalores e autovetores.
5. Utilizar linhas dos autovetores como coordenadas de um espaço transformado.
6. Limiarizar, ou aplicar k-médias ou outros algoritmos de agrupamento mais simples.

Experimento

Foi simulada uma classificação de alunos a partir de suas respostas em um teste.

A classificação indicaria grupos de alunos com dificuldades nas mesmas disciplinas ou tópicos.

Modelo probabilístico

O teste possui questões de múltipla escolha com 4 opções. Há diferentes tipos de PDF para cada questão:

Item	A	B	C	D
Questão fácil	90%	3,3%	3,3%	3,3%
Questão difícil	70%	10%	10%	10%
Erro sistemático	10%	70%	10%	10%
Chute	25%	25%	25%	25%

A resposta de um teste é uma amostra da PDF conjunta. Simulamos 30 questões respondidas por 100 alunos.

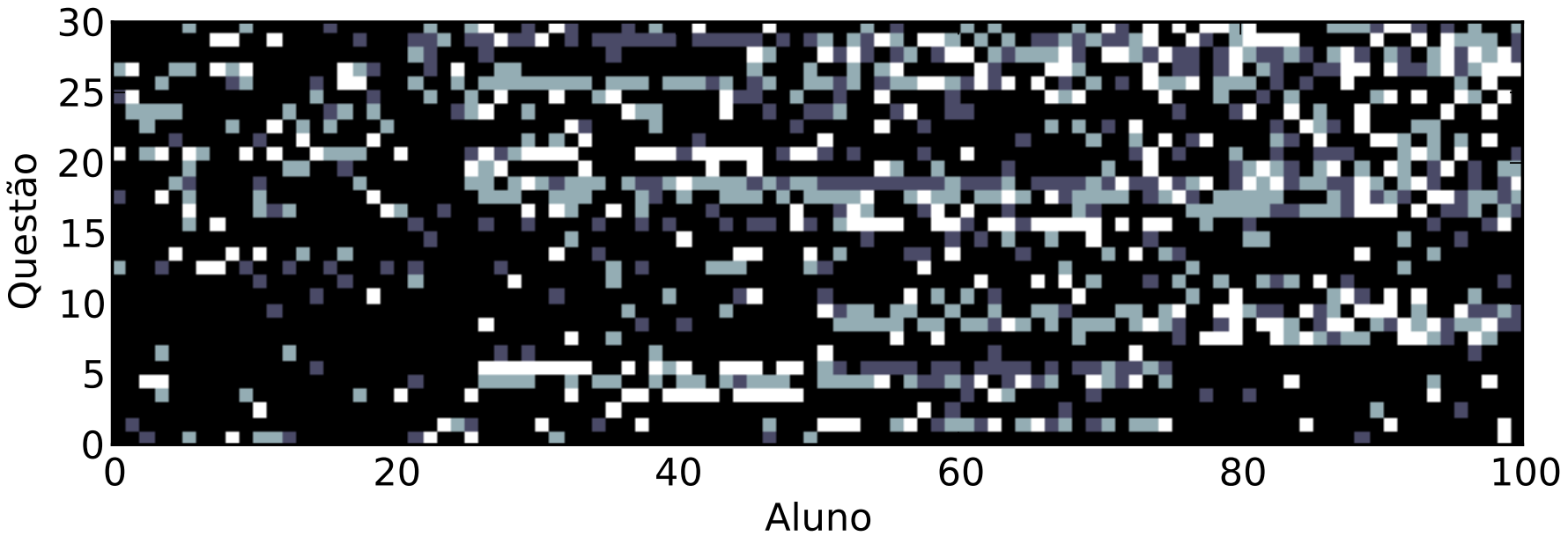
Definição das classes

Há quatro grupos de 25 alunos. Cada classe possui PDFs diferentes para cada questão.

A primeira classe é o *caso base*.

1. 10 questões fáceis, 10 médias e 10 difíceis.
2. Erros sistemáticos em 9 questões.
3. 6 erros sistemáticos, 5 chutes puros.
4. Chute puro em 11 questões.

Dados produzidos



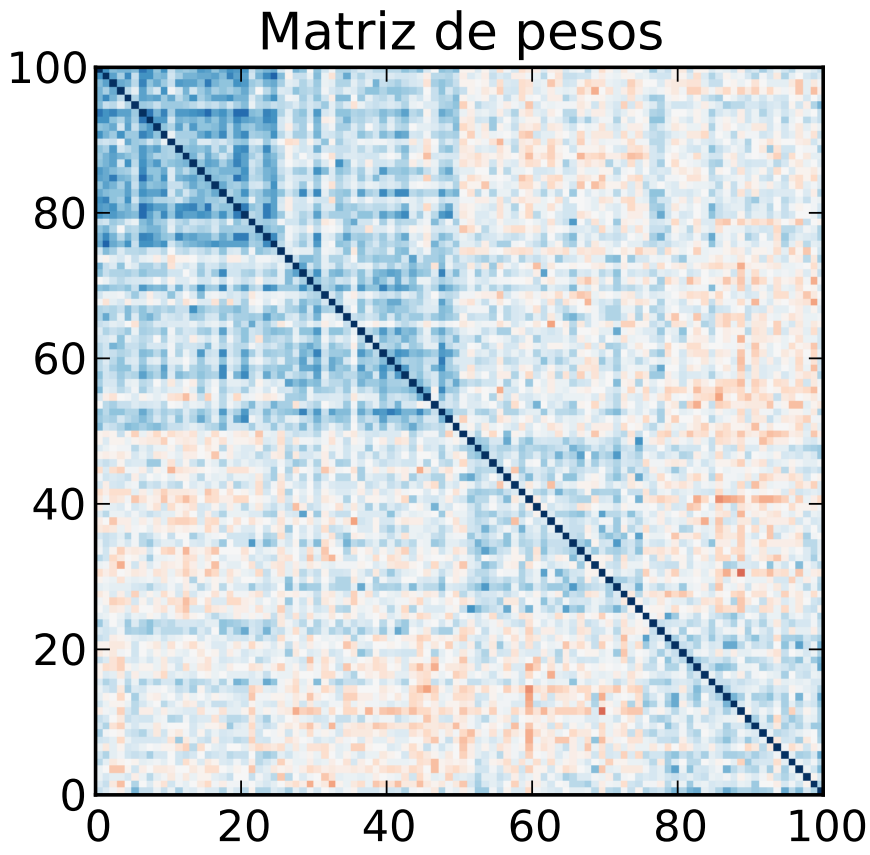
Função de Similaridade

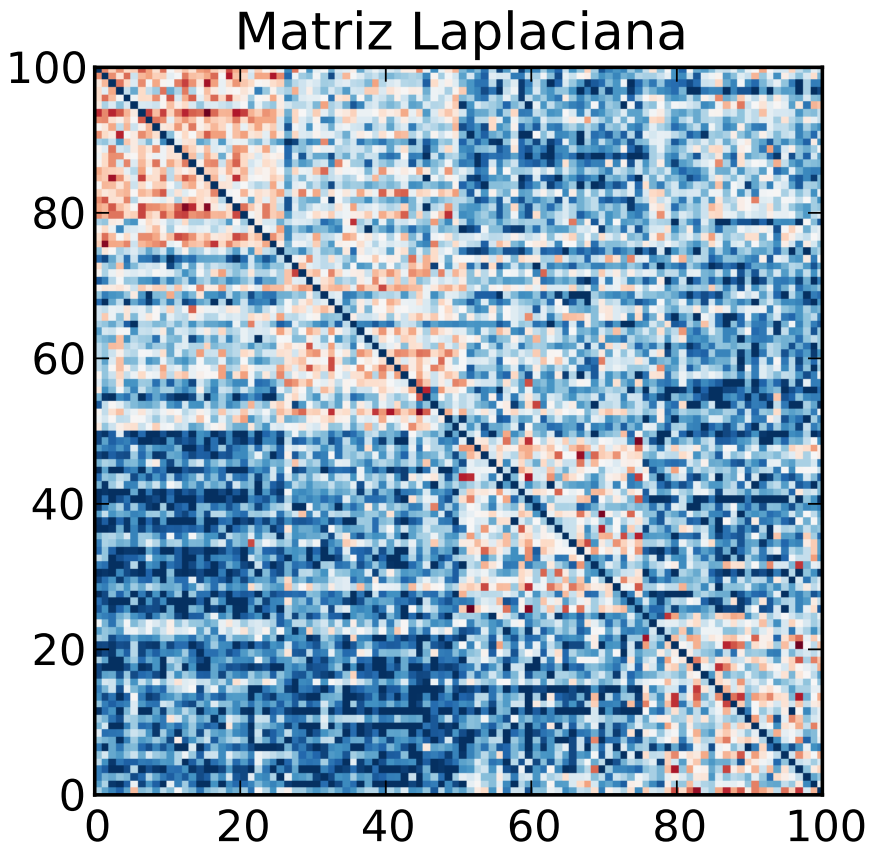
A similaridade entre as respostas de dois alunos é uma soma das similaridades de cada questão, pela tabela:

	A	B	C	D
A	1.0	0.0	0.0	0.0
B	0.0	1.0	0.5	0.5
C	0.0	0.5	1.0	0.5
D	0.0	0.5	0.5	1.0

- Certas ou erradas idênticas $\rightarrow$ 1.0,
- Resposta certa + errada $\rightarrow$ 0.0,
- Erradas diferentes $\rightarrow$ 0.5.

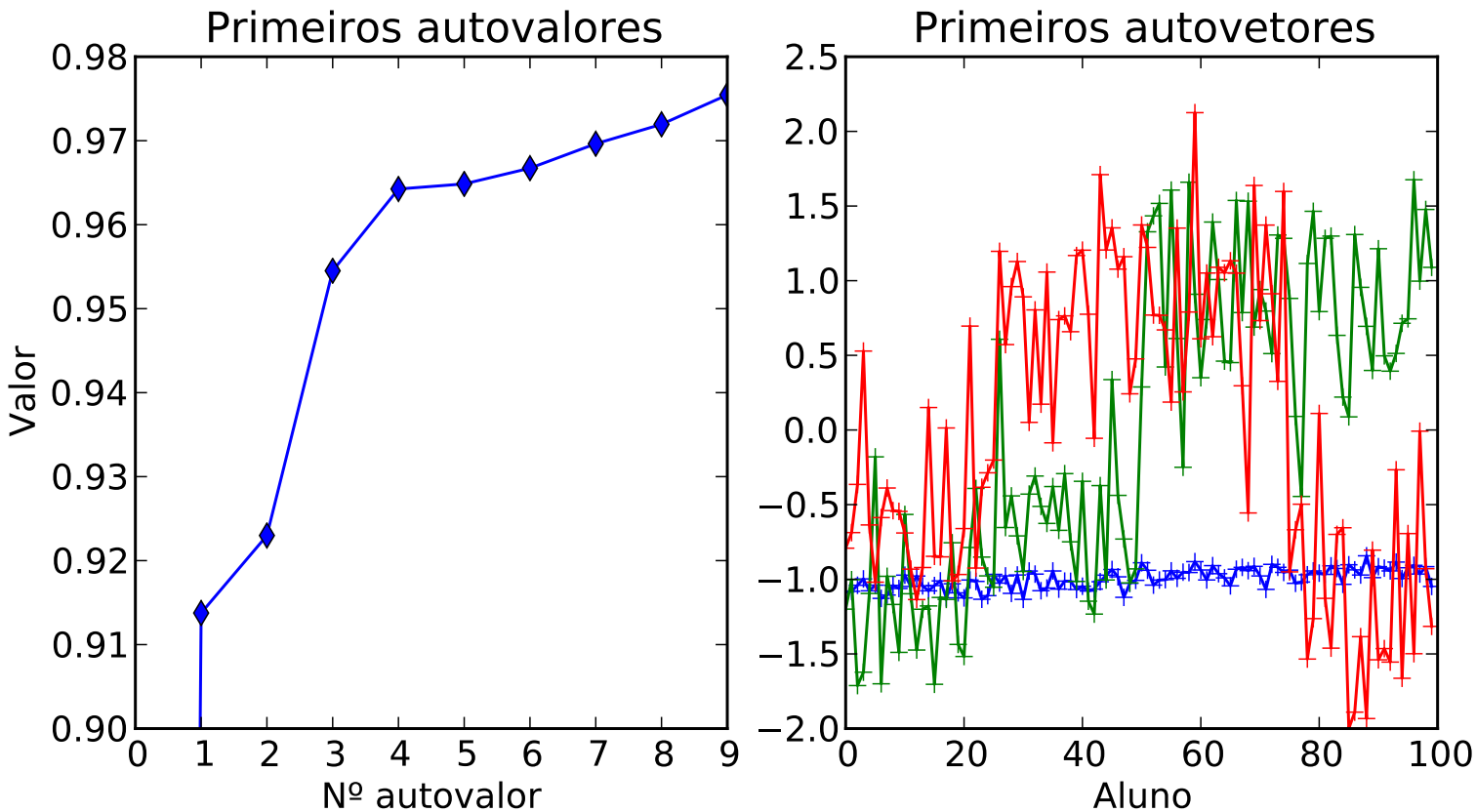
Matrizes do problema





Edgels e retas

Edgels são pontos amostrados sobre curvas ou retas.



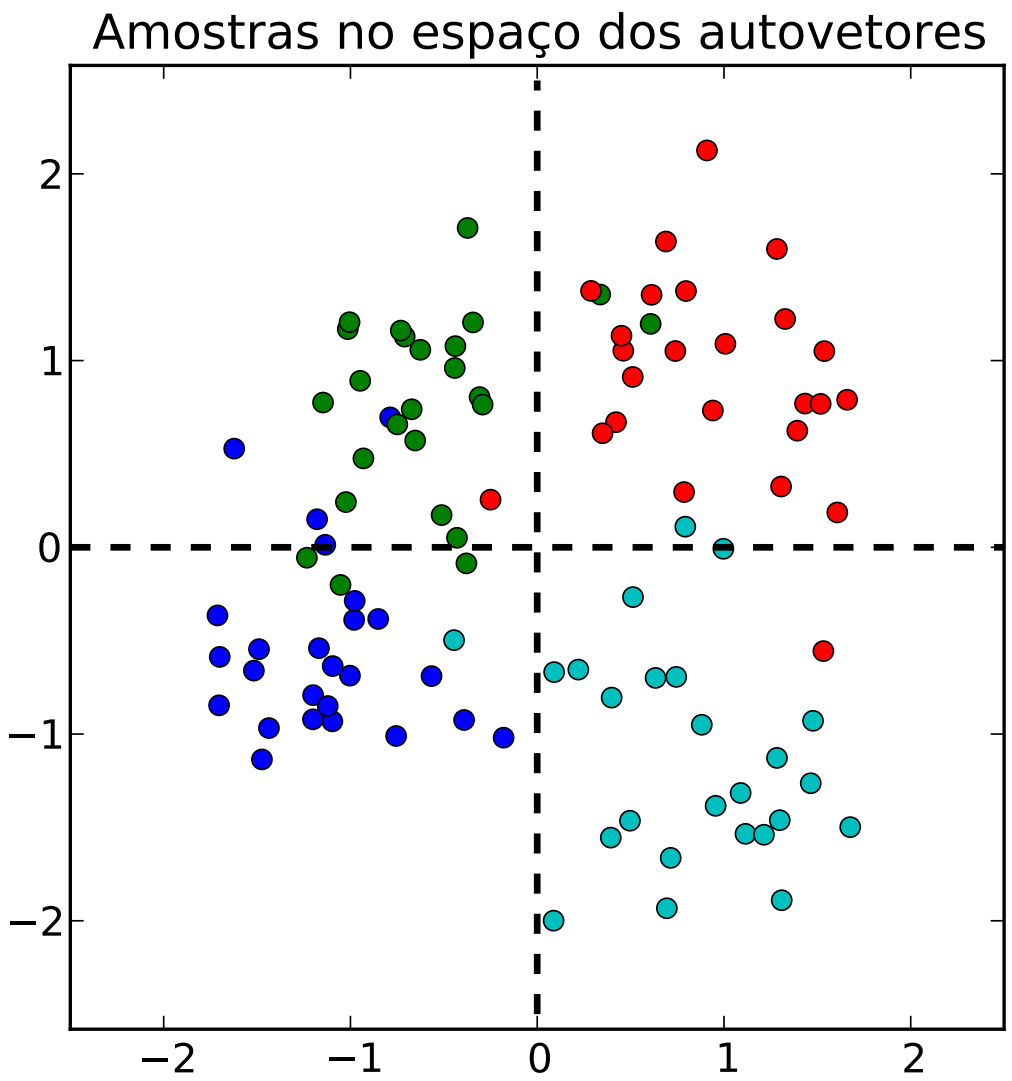
Classificação no espaço transformado

As amostras no espaço dos autovetores podem agora ser classificadas utilizando métodos convencionais.

SVM, *k-médias*, *ANN*...

Fizemos uma simples classificação de acordo com o quadrante de cada amostra.

Classificação no espaço transformado



Resultado da classificação

Matriz de confusão

	C_1	C_2	C_3	C_4
C_1	21	4	0	0
C_2	3	20	2	0
C_3	0	1	23	1
C_4	1	0	1	23

Desempenho do classificador

	C_1	C_2	C_3	C_4	
Precisão	84%	80%	88%	96%	μ: 87%
Revocação	84%	80%	92%	92%	μ: 87%
F-score	84%	80%	90%	94%	μ: 87%

Conclusão

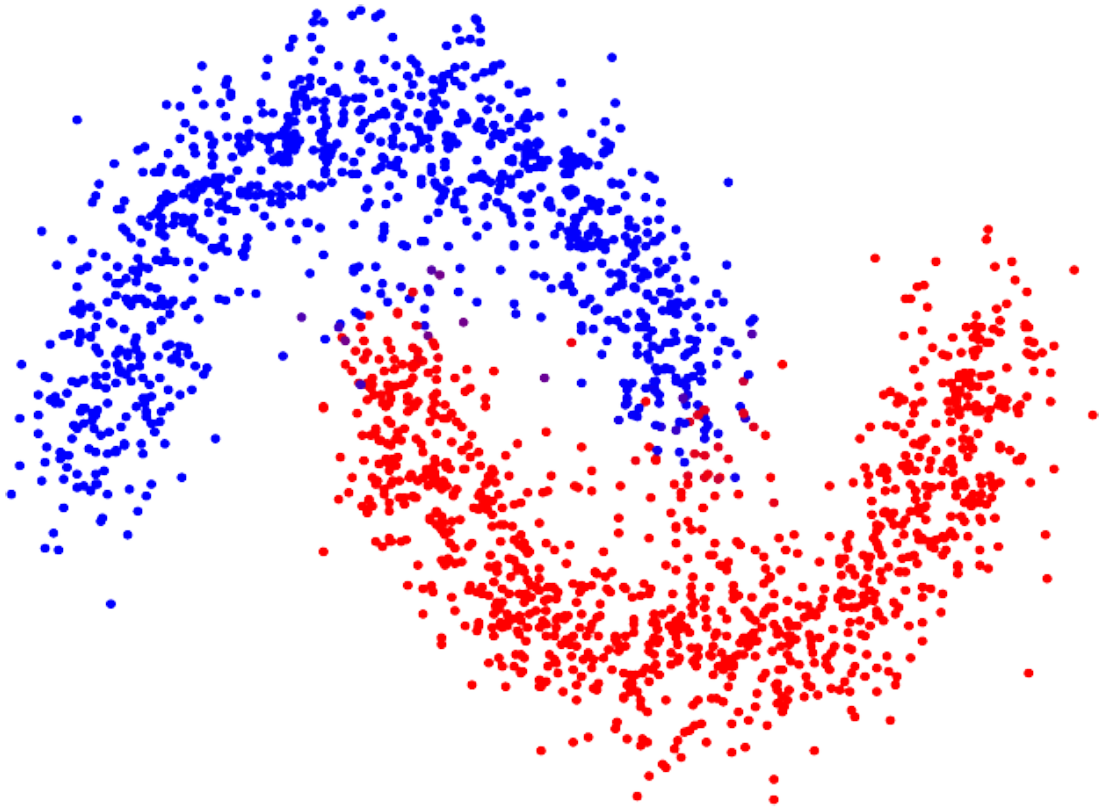
Apresentamos a técnica da clusterização espectral, e demonstramos como ela poderia ser útil para ensino.

Nosso experimento ilustra bem a técnica, mas:

1. O modelo probabilístico é bastante rudimentar.
2. É preciso analisar dados reais.
3. Uma aplicação com muitos dados precisa utilizar técnicas numeŕicas sofisticadas.

Fim

Obrigado!



Nicolau Werneck <nwerneck@gmail.com>

Referências Bibliográficas

Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Neural Information Processing Systems*, 2002. URL `http://books.google.com/books?hl=en&lr=&id=GbC8cqxGR7YC&oi=fnd&pg=PA849&dq=On+Spectral+CLustering:+Analysis+and+an+algorithm&ots=ZvN1H01DB5&sig=NsxAYwu8QzKmCeEo-FUfwMwkI4k`.

Ulrike von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, August 2007. ISSN 0960-3174. doi: 10.1007/s11222-007-9033-z. URL `http://www.springerlink.com/index/10.1007/s11222-007-9033-z`.