Algoritmo k-means

Aprendizagem Estatística

Paulo C. Marques F.

Primeiro semestre de 2019

Insper

Aprendizagem não supervisionada

Insper

Até o momento estudamos problemas de aprendizagem supervisionada, fazendo o uso de diversos métodos de classificação e regressão.

Os problemas de *aprendizagem não supervisionada* são aquelas situações inferenciais em que as observações $x_1, \dots, x_n$ não estão associadas a respostas $y_1 \dots, y_n$ fornecidas por um "supervisor".

Nesta aula, os dados não possuem mais rótulos que os classifiquem, ou respostas quantitativas correspondentes.

Em termos da teoria disponível, a aprendizagem não supervisionada é uma área muito menos desenvolvida da Aprendizagem Estatística.

Análise de clusters (1)

O problema exemplar em aprendizagem não supervisionada é a *análise de clusters* (conglomerados).

O objetivo é definir classes de equivalência tais que os dados dentro de uma mesma classe sejam “similares” segundo alguma perspectiva.

Intuitivamente, queremos que dentro de cada *cluster* os objetos sejam muito similares; e que objetos de dois *clusters* distintos sejam muito diferentes.

Queremos encontrar estruturas nos dados.

Para cada unidade amostral $i = 1, \ldots, n$, conhecemos apenas o vetor $x_i = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$.

Formalmente, um cluster C_i é o conjunto dos índices dos dados que pertencem a ele.

Queremos construir $k \geq 1$ clusters $C_1, \ldots, C_k$ tais que

$$\cup_{i=1}^k C_i = \{1, \ldots, n\},$$

e $C_i \cap C_j = \emptyset$ (*hard clustering*), quando $i \neq j$, de modo a minimizar a dispersão intra-clusters

$$W = \frac{1}{2} \sum_{r=1}^{k} \frac{1}{n_r} \sum_{i \in C_r} \sum_{j \in C_r} \|x_i - x_j\|^2,$$

em que n_r é o número de observações no cluster C_r e $\|x_i - x_j\|^2 = \sum_{\ell=1}^{p} (x_{i\ell} - x_{j\ell})^2$ é o quadrado da distância Euclidiana entre x_i e x_j.

Análise de clusters (3)

De quantas maneiras $A(n, k)$ podemos formar os k clusters a partir das n observações?

Suponha que você é a n-ésima unidade amostral.

Você pode fazer duas escolhas, mutuamente exclusivas.

Você pode decidir criar um cluster apenas para você e as demais $n - 1$ pessoas se agruparão em $k - 1$ clusters de $A(n - 1, k - 1)$ maneiras.

Ou você pode escolher entrar em um de k clusters e as demais $n - 1$ pessoas se agruparão nestes k clusters de $A(n - 1, k)$ maneiras.

Deste modo, obtemos a relação de recorrência

$$A(n, k) = A(n - 1, k - 1) + k \cdot A(n - 1, k),$$

com as condições $A(n, 1) = A(n, n) = 1$.

Análise de clusters (4)

Insper

```r
A <- function(n, k) {
    if (k == 1 || k == n) return(1)

    return(A(n - 1, k - 1) + k * A(n - 1, k))
}
```

Há casos factíveis: $A(10, 4) = 34\,105$, por exemplo.

No entanto, $A(30, 4) \approx 10^{16}$ e o problema se torna computacionalmente intratável.

Os $A(n, k)$ são conhecidos na literatura de combinatória como números de Stirling de segunda espécie.

De fato, é possível (Knuth) resolver a relação de recorrência e encontrar

$$A(n, k) = \frac{1}{k!} \sum_{r=1}^{k} (-1)^{k-r} \binom{k}{r} r^n.$$

Em busca de uma aproximação (1)

Insper

Lema 1. Vale a identidade

$$\sum_{i \in C_r} \sum_{j \in C_r} \|x_i - x_j\|^2 = 2\, n_r \sum_{i \in C_r} \|x_i - \bar{x}_r\|^2,$$

em que $\bar{x}_r = \sum_{i \in C_r} x_i / n_r$ é a média das observações pertencentes ao cluster C_r.

A idéia da demonstração é usar que

$$\|u - v\|^2 = \langle u - v, u - v \rangle = \|u\|^2 - 2\, \langle u, v \rangle + \|v\|^2$$

e "somar zero" no lugar adequado.

Em busca de uma aproximação (2)

Insper

Demonstração

$$
\begin{aligned}
\sum_{i \in C_r} \sum_{j \in C_r} \|x_i - x_j\|^2 &= \sum_{i \in C_r} \sum_{j \in C_r} \|(x_i - \bar{x}_r) - (x_j - \bar{x}_r)\|^2 \\
&= \sum_{i \in C_r} \sum_{j \in C_r} \left(\|x_i - \bar{x}_r\|^2 - 2\,\langle x_i - \bar{x}_r, x_j - \bar{x}_r\rangle + \|x_j - \bar{x}_r\|^2\right) \\
&= \sum_{i \in C_r} \left(n_r \|x_i - \bar{x}_r\|^2 - 2 \left\langle x_i - \bar{x}_r, \sum_{j \in C_r} (x_j - \bar{x}_r) \right\rangle + \sum_{j \in C_r} \|x_j - \bar{x}_r\|^2 \right) \\
&= 2\, n_r \sum_{i \in C_r} \|x_i - \bar{x}_r\|^2 .
\end{aligned}
$$

Em busca de uma aproximação (3)

Portanto, pelo Lema 1, o problema original

$$\arg\min_{C_1,\ldots,C_k} \frac{1}{2}\sum_{r=1}^{k}\frac{1}{n_r}\sum_{i\in C_r}\sum_{j\in C_r}\|x_i - x_j\|^2$$

é equivalente a

$$\arg\min_{C_1,\ldots,C_k} \sum_{r=1}^{k}\sum_{i\in C_r}\|x_i - \bar{x}_r\|^2.$$

Em busca de uma aproximação (4)

Insper

Lema 2. Para os vetores $u_1, \ldots, u_m \in \mathbb{R}^p$, a quantidade

$$\sum_{i=1}^{m} \|u_i - c\|^2$$

é minimizada escolhendo-se o vetor $c = \bar{u} = \sum_{i=1}^{m} u_i/m$.

Demonstração

$$\begin{aligned}
\sum_{i=1}^{m} \|u_i - c\|^2 &= \sum_{i=1}^{m} \|(u_i - \bar{u}) - (c - \bar{u})\|^2 \\
&= \sum_{i=1}^{m} \left(\|u_i - \bar{u}\|^2 - 2\langle u_i - \bar{u}, c - \bar{u}\rangle + \|c - \bar{u}\|^2\right) \\
&= \sum_{i=1}^{m} \|u_i - \bar{u}\|^2 + m\|c - \bar{u}\|^2.
\end{aligned}$$

Em busca de uma aproximação (5)

Usando o Lema 2, o problema original equivale a minimizar o "custo estendido"

$$\arg \min_{\substack{C_1,\ldots,C_k \\ m_1,\ldots,m_k}} \sum_{r=1}^{k} \sum_{i \in C_r} \|x_i - m_r\|^2.$$

Esta representação do problema sugere uma solução iterativa em que primeiramente fixamos os m_r's e minimizamos o custo estendido escolhendo os C_r's adequadamente, e posteriormente fixamos os C_r's e minimizamos o custo estendido escolhendo os m_r's como sendo as médias das observações nos respectivos clusters.

Algoritmo k-means (Lloyd (1957))

Insper

1. Inicializamos arbitrariamente $m_1, \ldots m_k$.
2. Alocamos a observação x_i no cluster C_r tal que

$$r = \arg \min_{1 \leq r \leq k} \|x_i - m_r\|,$$

   para $i = 1, \ldots, n$. Deste modo, determinamos $C_1, \ldots, C_k$.
3. Fazemos $m_r = \bar{x}_r$, para $r = 1, \ldots, k$.
4. Iteramos os dois passos anteriores até que o valor de W fique inalterado.

O algoritmo k-means converge

Insper

Note que os dois passos iterativos do algoritmo k-means reduzem o valor do custo estendido

$$\sum_{r=1}^{k} \sum_{i \in C_r} \|x_i - m_r\|^2.$$

Uma vez que o conjunto envolvido na iteração é finito, o algoritmo k-means eventualmente converge.

No entanto, não há nenhuma garantia de que encontraremos um mínimo global de W.

De fato, o algoritmo k-means fornece uma configuração de clusters que produz um mínimo local para W.

Por este motivo, os praticantes executam o algoritmo diversas vezes com inicializações distintas para os m_r's.

Levantamento

Insper

Vamos analisar um conjunto de dados da bilioteca `datasets` que traz os resultados de um levantamento feito entre funcionários dos escritórios de uma grande organização financeira. Os dados agregam as respostas de questionários aplicados a aproximadamente 35 empregados de 30 departamentos escolhidos ao acaso. Para sete questões, temos o percentual de respostas favoráveis em cada departamento.

```
library(datasets)

str(attitude)
```

```
## 'data.frame':    30 obs. of  7 variables:
##  $ rating    : num  43 63 71 61 81 43 58 71 72 67 ...
##  $ complaints: num  51 64 70 63 78 55 67 75 82 61 ...
##  $ privileges: num  30 51 68 45 56 49 42 50 72 45 ...
##  $ learning  : num  39 54 69 47 66 44 56 55 67 47 ...
##  $ raises    : num  61 63 76 54 71 54 66 70 71 62 ...
##  $ critical  : num  92 73 86 84 83 49 68 66 83 80 ...
##  $ advance   : num  45 47 48 35 47 34 35 41 31 41 ...
```

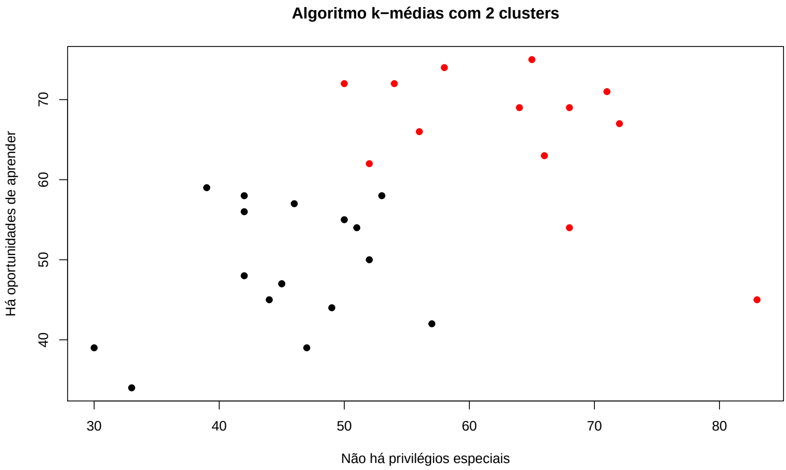
```
survey <- attitude[, 3:4]
```

Selecionando as respostas referentes à “existência de privilégios especiais” e à “oportunidade de aprender”, o algoritmo k-médias nos fornece a seguinte configuração com dois clusters.

Dois clusters

Insper

```
set.seed(1234)
k_means <- kmeans(survey, centers = 2, nstart = 100)
plot(survey, col = k_means$cluster, main = "Algoritmo k-médias com 2 clusters",
     xlab = "Não há privilégios especiais", ylab = "Há oportunidades de aprender", pch = 20, cex = 1.5)
```



Como escolher k? (1)

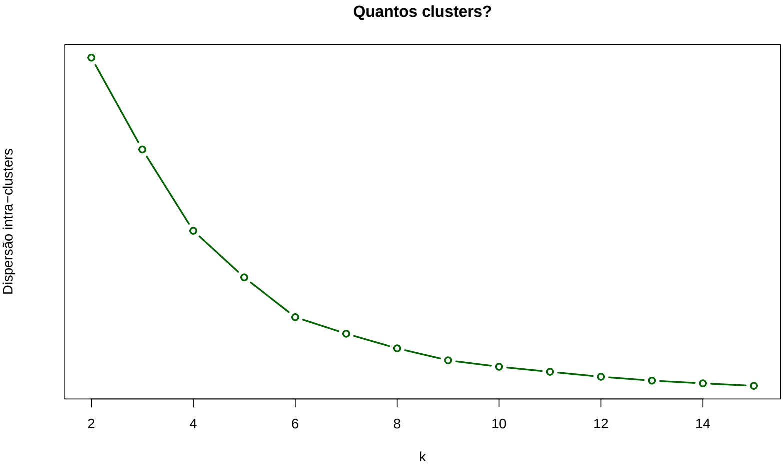
Insper

Escolher o número de clusters é sempre uma questão delicada. Uma técnica que ampara a escolha de k é calcular as dispersões totais intra-clusters para diversos valores de k e examinar o comportamento da curva procurando um “cotovelo”.

```r
set.seed(1234)

k_range <- 2:15

W <- numeric(length(k_range))

for (i in 1:length(k_range)) {
    k_means <- kmeans(survey, centers = k_range[i], nstart = 100)
    W[i] <- k_means$tot.withinss
}
```

Como escolher k? (2)

```r
plot(k_range, W, type = "b", lwd = 2, col = "dark green", yaxt = "n",
     xlab = "k", ylab = "Dispersão intra-clusters", main = "Quantos clusters?")
```



Seis clusters

Insper

```
set.seed(1234)
k_means <- kmeans(survey, centers = 6, nstart = 100)
plot(survey, col = k_means$cluster, main = "Algoritmo k-médias com 6 clusters",
     xlab = "Não há privilégios especiais", ylab = "Há oportunidades de aprender", pch = 20, cex = 1.5)
```