Mineração de Dados

Aula 12: Clustering: análise de agrupamento/segmentação

Rafael Izbicki

Aprendizado não supervisionado

O Problema

Como dividir sua amostra em grupos de indivíduos que são parecidos entre si? Isto é, grupos diferentes uns dos outros, mas homogêneos entre si.

Formalmente: queremos criar uma partição da nossa amostra $C_1, \ldots, C_K$. Isto é, devemos ter:

$$C_1 \bigcup C_2 \bigcup \ldots \bigcup C_K = \{1, 2 \ldots, n\}$$

e

$$C_i \bigcap C_j = \ \forall i \neq j$$

Ex: $K = 2$, $C_1 = \{3, 4, 6\}$, $C_2 = \{1, 2, 5, 7\}$

O Problema

Como dividir sua amostra em grupos de indivíduos que são parecidos entre si? Isto é, grupos diferentes uns dos outros, mas homogêneos entre si.

Formalmente: queremos criar uma partição da nossa amostra $C_1, \ldots, C_K$. Isto é, devemos ter:

$$C_1 \bigcup C_2 \bigcup \ldots \bigcup C_K = \{1, 2 \ldots, n\}$$

e

$$C_i \bigcap C_j = \ \forall i \neq j$$

Ex: $K = 2$, $C_1 = \{3, 4, 6\}$, $C_2 = \{1, 2, 5, 7\}$

O Problema

Como dividir sua amostra em grupos de indivíduos que são parecidos entre si? Isto é, grupos diferentes uns dos outros, mas homogêneos entre si.

Formalmente: queremos criar uma partição da nossa amostra $C_1, \ldots, C_K$. Isto é, devemos ter:

$$C_1 \bigcup C_2 \bigcup \ldots \bigcup C_K = \{1, 2 \ldots, n\}$$

e

$$C_i \bigcap C_j = \ \forall i \neq j$$

Ex: $K = 2$, $C_1 = \{3, 4, 6\}$, $C_2 = \{1, 2, 5, 7\}$

Conceito essencial: como medir dissimilaridade (ou similaridade) entre dois indivíduos.

Várias possíveis medidas.

Ex: Distância Euclidiana

$$d^2(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} (x_{i,k} - x_{j,k})^2$$

Para variáveis discretas: criar variáveis "dummies"

Conceito essencial: como medir dissimilaridade (ou similaridade) entre dois indivíduos.

Várias possíveis medidas.

Ex: Distância Euclidiana

$$d^2(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{p} (x_{i,k} - x_{j,k})^2$$

Para variáveis discretas: criar variáveis "dummies"

K-Médias

O K-Médias supõe que a medida de dissimilaridade usada é a distância Euclidiana. Para usá-lo, é necessário especificar de antemão K, quantos clusters se deseja.

Para o K-médias, buscar o melhor clustering significa buscar a uma partição $C_1, \ldots, C_K$ da nossa amostra tal que

$$\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} d^2(\mathbf{x}_i, \mathbf{x}_j)$$

seja baixo. Este quantidade é a soma de quadrados dentro de cada cluster.

O método para encontrar o resposta para esse problema é um algoritmo iterativo. (vocês devem estudar isso mais a fundo em multivariada).

K-Médias

O K-Médias supõe que a medida de dissimilaridade usada é a distância Euclidiana. Para usá-lo, é necessário especificar de antemão K, quantos clusters se deseja.

Para o K-médias, buscar o melhor clustering significa buscar a uma partição $C_1, \ldots, C_K$ da nossa amostra tal que

$$\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} d^2(\mathbf{x}_i, \mathbf{x}_j)$$

seja baixo. Este quantidade é a soma de quadrados dentro de cada cluster.

O método para encontrar o resposta para esse problema é um algoritmo iterativo. (vocês devem estudar isso mais a fundo em multivariada).

K-Médias

O K-Médias supõe que a medida de dissimilaridade usada é a distância Euclidiana. Para usá-lo, é necessário especificar de antemão K, quantos clusters se deseja.

Para o K-médias, buscar o melhor clustering significa buscar a uma partição $C_1, \ldots, C_K$ da nossa amostra tal que

$$\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} d^2(\mathbf{x}_i, \mathbf{x}_j)$$

seja baixo. Este quantidade é a soma de quadrados dentro de cada cluster.

O método para encontrar o resposta para esse problema é um algoritmo iterativo. (vocês devem estudar isso mais a fundo em multivariada).

K-Médias

O K-Médias supõe que a medida de dissimilaridade usada é a distância Euclidiana. Para usá-lo, é necessário especificar de antemão K, quantos clusters se deseja.

Para o K-médias, buscar o melhor clustering significa buscar a uma partição $C_1, \ldots, C_K$ da nossa amostra tal que

$$\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,j \in C_k} d^2(\mathbf{x}_i, \mathbf{x}_j)$$

seja baixo. Este quantidade é a soma de quadrados dentro de cada cluster.

O método para encontrar o resposta para esse problema é um algoritmo iterativo. (vocês devem estudar isso mais a fundo em multivariada).

Algoritmo de Lloyd

1. Escolha aleatoriamente k centróides $c_1, \ldots, c_k$.
Itere:

2. **Atribuição**: Defina o cluster C_j ($j = 1, \ldots, k$) como sendo

$$C_j = \{x_i : \arg\min_r d(x_i, c_r) = r\}$$

3. **Atualização**: Calcule os novos centróides usando os grupos que foram criados:

$$c_j \longleftarrow \frac{1}{|C_j|} \sum_{j : x_j \in C_j} x_j$$

`https://www.youtube.com/watch?v=ZMfwPUrOFsE`

Algoritmo de Lloyd

1. Escolha aleatoriamente k centróides $c_1, \dots, c_k$.

Itere:

2. **Atribuição**: Defina o cluster C_j $(j = 1, \dots, k)$ como sendo

$$C_j = \{x_i : \arg\min_r d(x_i, c_r) = r\}$$

3. **Atualização**: Calcule os novos centróides usando os grupos que foram criados:

$$c_j \longleftarrow \frac{1}{|C_j|} \sum_{j : x_j \in C_j} x_j$$

https://www.youtube.com/watch?v=ZMfwPUrOFsE

Algoritmo de Lloyd

1. Escolha aleatoriamente k centróides $c_1, \ldots, c_k$.

Itere:

2. **Atribuição**: Defina o cluster C_j $(j = 1, \ldots, k)$ como sendo

$$C_j = \{x_i : \arg\min_r d(x_i, c_r) = r\}$$

3. **Atualização**: Calcule os novos centróides usando os grupos que foram criados:

$$c_j \longleftarrow \frac{1}{|C_j|} \sum_{j : x_j \in C_j} x_j$$

https://www.youtube.com/watch?v=ZMfwPUrOFsE

Algoritmo de Lloyd

1. Escolha aleatoriamente k centróides $c_1, \ldots, c_k$.
Itere:

2. **Atribuição**: Defina o cluster C_j $(j = 1, \ldots, k)$ como sendo

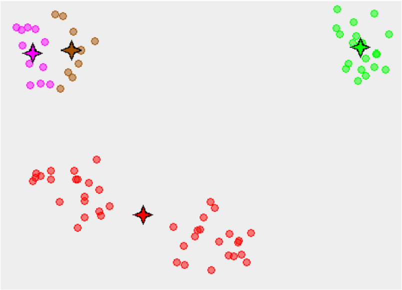
$$C_j = \{x_i : \arg\min_r d(x_i, c_r) = r\}$$

3. **Atualização**: Calcule os novos centróides usando os grupos que foram criados:

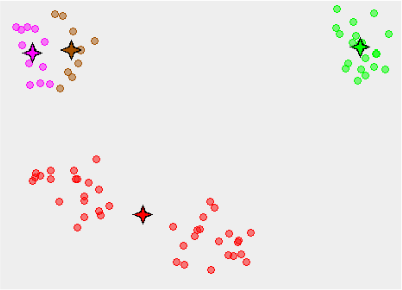
$$c_j \longleftarrow \frac{1}{|C_j|} \sum_{j : x_j \in C_j} x_j$$

`https://www.youtube.com/watch?v=ZMfwPUrOFsE`

O algoritmo depende de escolhas iniciais, e portanto o resultado pode ser um mínimo local dependendo da inicialização.



O algoritmo depende de escolhas iniciais, e portanto o resultado pode ser um mínimo local dependendo da inicialização.



Uma melhoria: k-médias++

1. Escolha c_1 aleatoriamente entre $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ e defina $C = \{c_1\}$.
2. Para $j = 2, \ldots, k$:
   2.1 Calcule $D(\mathbf{x}_i) = \min_{c \in C} ||\mathbf{x}_i - c||$ para cada $\mathbf{x}_i$
   2.2 Escolha uma amostra $\mathbf{x}_i$ aleatoriamente entre todas as amostras observadas com probabilidade
   $$p_i = \frac{D^2(\mathbf{x}_i)}{\sum_{j=1}^{n} D^2(\mathbf{x}_j)}$$
   2.3 Defina c_j como sendo o ponto escolhido. Atualize
   $$C \leftarrow C \cup \{c_j\}$$

Métodos Hierárquicos

Um problema do K-médias é que K deve ser especificado de antemão. Métodos Hierárquicos são uma forma de evitar isso.

(1) Atribua cada observação a um cluster diferente. Calcule cada uma das $\binom{n}{2}$ distâncias entre esses clusters.

(2) Para $i = n, n-1, \ldots, 2$:

(a) Procure entre todos os pares formados por dois dos i clusters aqueles mais parecidos. Junte esses dois clusters em um só. A dissimilaridade entre esses dois clusters indica a altura do dendrograma em que a junção será feita.

(b) Calcule cada uma das distâncias entre os novos i-1 clusters.

Métodos Hierárquicos

Um problema do K-médias é que K deve ser especificado de antemão. Métodos Hierárquicos são uma forma de evitar isso.

(1) Atribua cada observação a um cluster diferente. Calcule cada uma das $\binom{n}{2}$ distâncias entre esses clusters.

(2) Para $i = n, n-1, \ldots, 2$:

(a) Procure entre todos os pares formados por dois dos i clusters aqueles mais parecidos. Junte esses dois clusters em um só. A dissimilaridade entre esses dois clusters indica a altura do dendrograma em que a junção será feita.

(b) Calcule cada uma das distâncias entre os novos i-1 clusters.

Métodos Hierárquicos

Um problema do K-médias é que K deve ser especificado de antemão. Métodos Hierárquicos são uma forma de evitar isso.

(1) Atribua cada observação a um cluster diferente. Calcule cada uma das $\binom{n}{2}$ distâncias entre esses clusters.

(2) Para $i = n, n-1, \ldots, 2$:

(a) Procure entre todos os pares formados por dois dos i clusters aqueles mais parecidos. Junte esses dois clusters em um só. A dissimilaridade entre esses dois clusters indica a altura do dendrograma em que a junção será feita.

(b) Calcule cada uma das distâncias entre os novos i-1 clusters.

Métodos Hierárquicos

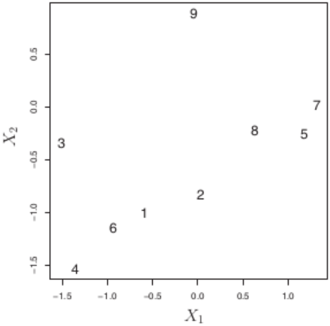
Um problema do K-médias é que K deve ser especificado de antemão. Métodos Hierárquicos são uma forma de evitar isso.

(1) Atribua cada observação a um cluster diferente. Calcule cada uma das $\binom{n}{2}$ distâncias entre esses clusters.

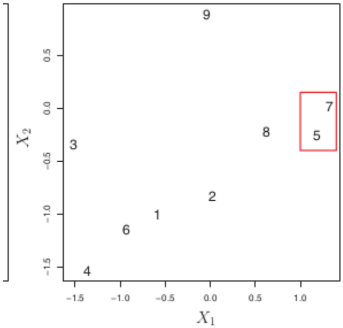
(2) Para $i = n, n-1, \ldots, 2$:

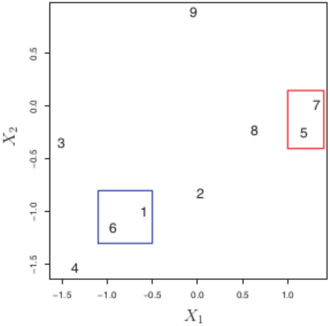
(a) Procure entre todos os pares formados por dois dos i clusters aqueles mais parecidos. Junte esses dois clusters em um só. A dissimilaridade entre esses dois clusters indica a altura do dendrograma em que a junção será feita.

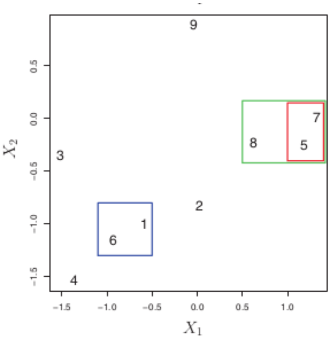
(b) Calcule cada uma das distâncias entre os novos i-1 clusters.



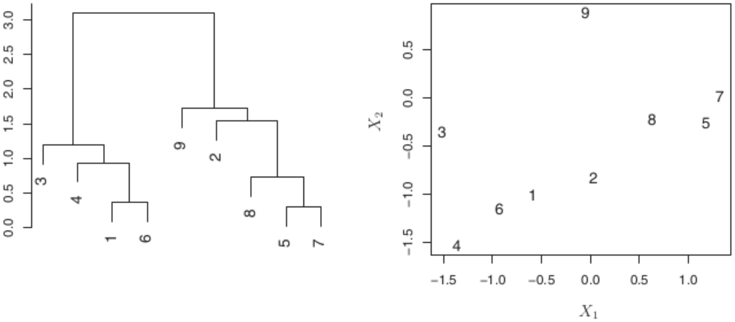
9
7
8
5
3
2
1
6
4
X_1
X_2
-1.5
-1.0
-0.5
0.0
0.5
1.0


9
7
8
5
3
2
1
6
4
0.5
0.0
−0.5
−1.0
−1.5
X_2
−1.5
−1.0
−0.5
0.0
0.5
1.0
X_1


9
7
8
5
3
2
1
6
4
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
X_1
X_2



9
7
8
5
3
2
1
6
4
0.5
0.0
−0.5
−1.0
−1.5
X_2
−1.5
−1.0
−0.5
0.0
0.5
1.0
X_1

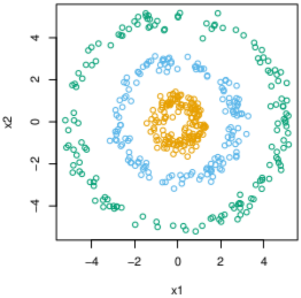

3.0
2.5
2.0
1.5
1.0
0.5
0.0
3
4
1
6
9
2
8
5
7
X_2
X_1
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0

Há várias formas de se definir a distância entre dois clusters:

Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

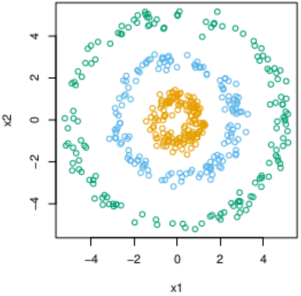
Na prática, é comum tentarmos vários métodos de clustering (diferentes distâncias, linkages etc), e buscarmos a solução mais interpretável.

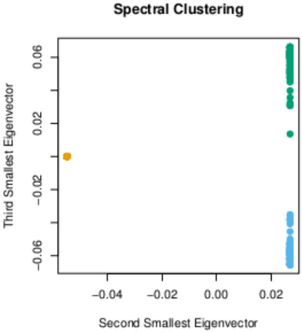
Clustering Espectral



Redução de dimensionalidade (Kernel PCA ou variações) $+$
Técnicas tradicionais de clustering

Clustering Espectral





Redução de dimensionalidade (Kernel PCA ou variações) + Técnicas tradicionais de clustering nas variáveis reduzidas