Métricas em Machine Learning

Avaliação duma partição

Comecemos por ver possíveis avaliações dum cluster.

- Se o temos o representante do cluster, podemos quantificar as dissemelhanças entre o representante e os seus membros.
- Se não temos o representante, fazemos avaliação direta entre pares de membros do cluster.

  Entre várias possíveis métricas, algumas das mais populares são:

  Dissemelhança média de cluster :

$$d_1(m; C) = \frac{1}{|C|} \sum_{x^n \in C} d(m, x^n)$$

  (minimiza impacto de outlyers)

Métricas em Machine Learning

Avaliação duma partição

Dissemelhança quadrática do cluster :

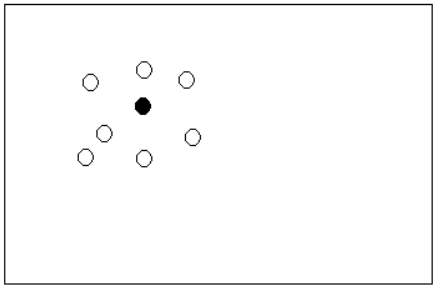
$$d_2(m; C) = \sqrt{\frac{1}{|C|} \sum_{x^n \in C} d(m, x^n)^2}$$

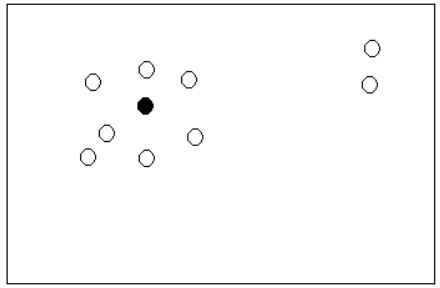
Dissemelhança máxima do cluster :

$$d_\infty(m; C) = max\{d(m, x^n), \ \ x^n \in C\}$$

(maximiza impacto dos outliers)

Métricas em Machine Learning





$d1 \approx d_2 \approx d_\infty$ $\quad$ $d1 <\approx d_2 < d_\infty$

Pode-se introduzir uma métrica mais sofisticada.

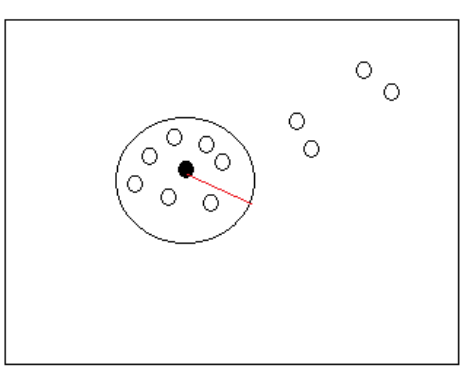
$$N(r) = N(r; m, C) = |\{x \in C,\ d(m, C) \leq r\}|$$

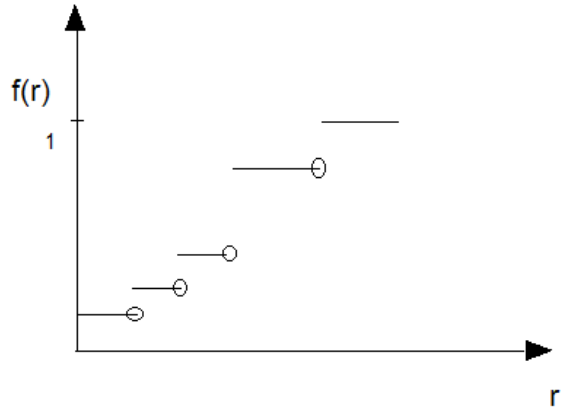
que dá o número de membros de C que estão num raio r do representante m do conjunto C. Defina-se,

$$f(r) = f(r; m, C) = \frac{N(r)}{N}$$

que dá a frequência relativa de elementos do cluster C função do raio r medido a partir do representante m.

Métricas em Machine Learning





Podemos usar esta métrica para detectar outliers.

Outra métrica bastante usada na literatura para avaliar um cluster é **diâmetro do cluster**, a saber,

$$\Delta(C) = \max_{x \in C, x' \in C} (d(x, x'))$$

pode ser interpretada como a "distância" máxima entre quaisquer pontos do cluster.

Métricas em Machine Learning

Recordem-se algumas métricas inter-cluster estudadas nas aulas anteriores.

1) Sejam m e m' os representantes de C e C', então $dd(C, C') = d(m, m')$. Neste caso estamos a usar uma métrica ponto a ponto entre os representantes dos clusters.

2) $dd(C, C') = min\ d(x, x'),\ \ x \in C,\ x' \in C'$. Chamada de **Single Linkage**, calcula a menor distância entre pontos de C e pontos de C'.

3) $dd(C, C') = max\ d(x, x'),\ \ x \in C,\ x' \in C'$. Chamada de **Complete Linkage**, calcula a maior distância entre pontos de C e pontos de C'.

4) $dd(C, C') = \frac{1}{|C|.|C'|} \sum_{x \in C, x' \in C'} d(x, x'),\ \ x \in C,\ x' \in C'$. Chamada de **Average**, calcula a média das distâncias entre pontos de C e pontos de C'.

Métricas em Machine Learning

Para avaliar o resultado dum algoritmo de clustering, vamos considerar que obtivémos a partição, $\mathcal{P} = \{C^1, C^2, ..., C^k\}$.

Vamos ainda considerar que temos o conjunto de representantes $\mathcal{M} = \{m^1, m^2, ...m^k\}$ (podíamos não ter).

Podemos por exemplo definir,

$$E(m^k, C^k) = \frac{1}{|C^k|} \sum_{x^n \in C^k} (d(m^k, C^k))^2$$

que avalia o cluster C^k usando a dissemelhança média do cluster C^k.

Podemos definir uma métrica que avalia toda a partição,

$$EE(\mathcal{M}, \mathcal{P}) = \sum_{i=1}^{k} E(m^i, C^i)$$

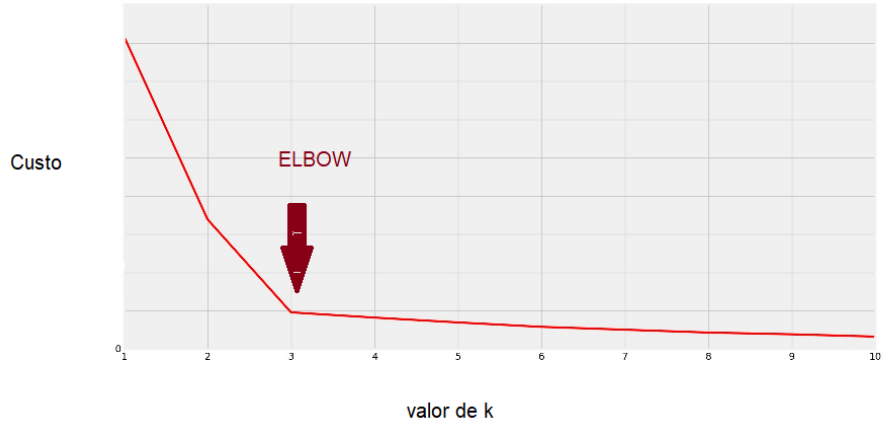
Notar que neste caso estamos apenas a usar métricas intra-cluster. Outras métricas mais elaboradas podem ser definidas, envolvendo métricas intra-cluster e inter-cluster como **Silhouette Score**. TPC - Investigar.

Algoritmo Lloyd - como estimar k?

Na literatura existem vários métodos como o Elbow method.

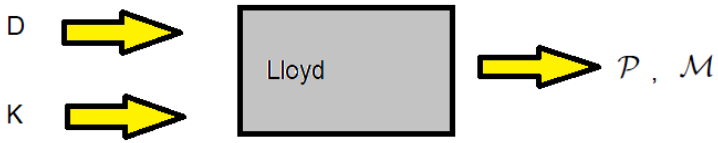
Vejamos como funciona o **Elbow method**:

- Definimos atrás uma função que avalia a qualidade duma partição $EE(\mathcal{M}, \mathcal{P})$.
- Corremos o algoritmo de LLoyd para vários valores de k e calculamos $EE(\mathcal{M}, \mathcal{P})$.
- Representamos o gráfico de $EE(\mathcal{M}, \mathcal{P})$ função de k.
- A escolha de k é o valor de k onde aparece o **elbow** - valor a partir do qual $EE(\mathcal{M}, \mathcal{P})$ diminui menos rapidamente.



Um classificador

Vamos agora apresentar como construir um possível classificador após um processo de clusterização, como o algoritmo de Lloyd. O objetivo é caso surja um novo evento, ter uma ferramenta que permita classificar esse evento.



1) Labelizar os elementos dos vários clusters.

2) Avaliar a dissemelhança entre o novo evento e os representanates dos vários clusters. classificar o novo evento como sendo da classe do representante mais próximo.

3) Alternativamente ao passo 2, determinar os K elementos mais práximos do novo evento. Atribuir ao novo evento o label que ocorrer mais vezes nesse K elementos mais próximos.

Métricas em Machine Learning

4) Avaliar a performance do classificador usando matrizes de confusão.

Bases de dados comuns na literatura para testar cluestering:

- IRIS dataset (dados reais de $\mathbb{R}^4$)
  `https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html`

- MNIST dataset (imagens com números escritos à mão com 28x28 com 16 níveis de cinza)
  `https://paperswithcode.com/dataset/mnist`