Dados nominativos - Representante

Quando lidamos com dados categóricos, não é óbvia a escolha dum representante. Uma das vias passa por usar as frequências que quantificam a probabilidade de **match** dos elementos mais frequentes.

Um exemplo:

Seja $\mathcal{A} = A_1 \times A_2 \times ... \times A_I$ o espaço admissível dos atributos, onde $A_i = \{a_{i,1}, a_{i,2}, ..., a_{i,J_i}\}$ é uma lista de Ji elementos que caracterizam A_i.

Podemos definir uma função de semelhança entre $x,\ y\ \in \mathcal{A}$,

$$s_i(x,y) = \begin{cases} 1 & se\ x_i = y_i \\ 0 & se\ x_i \neq y_i \end{cases}$$

que diz se o atributo i do elemento x é ou não igual ao atributo i do elemento y. E a função de semelhança global,

$$s(x,y) = \sum_{i=1}^{I} s_i(x,y).$$

Métricas em Machine Learning

Seja $D = \{x^1, x^2, ..., x^N\}$ um conjunto de dados, para o qual procuramos um representante. Defina-se uma funcional S que avalie o grau de semalhança entre o seu representante m e os seus dados x^n.

$$S(m; D) = \sum_{n=1}^{N} s(m, x^n).$$

O melhor candidato a representante de D tem que maximizar S.

$$m = \arg\max_{m \in \mathcal{A}} S(m; D).$$

Ou seja, o representante de D vai maximizar a funcional,

$$S(m; D) = \sum_{n=1}^{N} \sum_{i=1}^{I} s_i(m, x^n) = \sum_{i=1}^{I} \sum_{n=1}^{N} s_i(m, x^n) = \sum_{i=1}^{I} S_i(m; D)$$

que corresponde a calcular o máximo valor de semelhança para cada S_i (maximizar semelhança entre m_i e as componentes i de todos os elementos de D).

Métricas em Machine Learning

Seja $N_{i,j}$ o número de elementos de D tal que $x_i^n = a_{i,j}$, então,

$S_i(a_{i,j}; D) = N_{i,j}$.

O conjunto D tem número finito de elementos $\rightarrow$ existe um máximo (não necessariamente único) a que chamamos $\overline{m}_i$.

Notar que $\overline{m}_i$ é o elemento da categoria A_i que ocorre mais vezes (**moda**).

Fazendo o mesmo para todas as componentes, obtém-se o representante de D,

$$\overline{m} = (\overline{m}_1, \overline{m}_2, ..., \overline{m}_I)$$

.

Trata-se de um representante do tipo centróide já que $\overline{m}$ não pertence a D.

Exemplo 26 :

Considere $\mathcal{A} = A_1 \times A_2 \times A_3$ onde $A_1 = \{boreal, temperado, tropical\}$, $A_2 = \{F, T\}$ e $A_3 = \{branco, amarelo, laranja, veremelho\}$. Considere ainda o conjunto D definido pela tabela,

n	A1	A2	A3
1	boreal	F	amarelo
2	tropical	T	amarelo
3	tropical	F	laranja
4	temperado	F	amarelo
5	boreal	T	branco
6	boreal	T	branco

Calcule o representante, usando a métrica definida atrás.

Clustering

Notações :

Considere-se o espaço dos atributos $\mathcal{A} = A_1 \times A_2 \times ... \times A_I$.

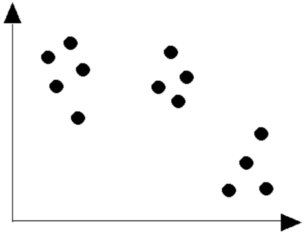
$x^n = (x_1^n, x_2^n ...., x_I^n)$ representa o evento n da base de dados
$D = \{x^n \in \mathcal{A} : \; n = 1, ..., N\}$,

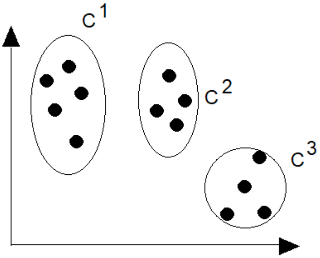
Definição : $\mathcal{P}$ é uma K-partição de D se,

$\mathcal{P} = \{C^1, C^2, ..., C^K\}$ onde,

a) $C^K \subseteq D$ com $C^k \neq \emptyset$;
b) $C^K \cap C^L = \emptyset$ para $K \neq L$;
c) $\displaystyle\bigcup_{k=1}^{K} C^k = D$.

Uma partição de um conjunto D é uma família de conjuntos composta por subconjuntos não vazios de D, disjuntos 2 a 2 e cuja união é D.





Uma a partição de D representada na figura acima poderá ser $\mathcal{P} = \{C^1, C^2, C^3\}$.

De seguida vamos apresentar um algoritmo do tipo **partitional clustering**:

1) Devemos analisar os dados e escolher métricas apropriadas;
2) Escolher técnica para determinar representante $\rightarrow$ escolher função custo e minimiza-la;
3) Escolher o tipo de algoritmo de Partitional Clustering (vamos falar de 2 : tipo Lloyd e Hierarchical Clustering).

Mappings

Seja $D = \{x^n : \; n = 1, ..., N\}$ uma base de dados com elementos $x^n \in \mathcal{A}$.

Seja $\mathcal{P} = \{C^1, C^2, ..., C^K\}$ uma partição de D.

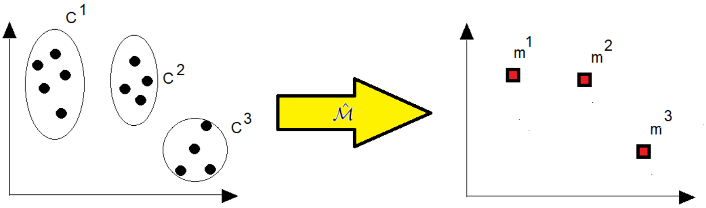
O mapa de representantes de $\mathcal{P} = \{C^1, C^2, ..., C^K\}$ faz a associação $m^k = Representante(C^k)$.

O conjunto dos representantes é dado por,

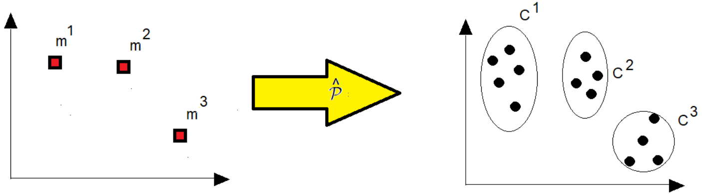
$\hat{\mathcal{M}}(\mathcal{P}) = \{Representante(C^1), Representante(C^2), ..., Representante(C^K)\}$

que é uma função de mapping, que tem por input uma partição $\mathcal{P}$ e tem por output um conjunto de representantes.

Métricas em Machine Learning



P mapping: dados um conjunto de representantes $\mathcal{M}$ e o conjunto D vai construir novos clusters obtendo uma nova partição de D.



Métricas em Machine Learning

Algoritmo do tipo LLoyd

- Para um valor de K (número de clusters), e a partir dos dados em D, gera K clusters convexos;

- Converge num número finito de iterações;

- O número de iterações pode ser grande;

- A partição final depende da condição inicial - $\mathcal{M}(0)$;

- Não há garantias que a partição obtida seja a melhor. (a função a minimizar pode admitir mínimos locais)

- para dados reais, usando a métrica Euclidiana é uma implementação do algoritmo **K-means**.

Algoritmo de LLoyd

K $\leftarrow$ ler

$\mathcal{M}(0) = \{m^1(0), m^2(0), ..., m^K(0)\}$

CP=0; t=1

While $\neg CP$

$\mathcal{P}(t) = \hat{\mathcal{P}}(\mathcal{M}(t-1))$

$\mathcal{M}(t) = \hat{\mathcal{M}}(\mathcal{P}(t))$

CP=condicao de paragem$(\mathcal{M}(t), \mathcal{M}(t-1))$

t=t+1

End While

Exemplo 27:

Aplicar o algoritmo de Lloyd com:

$(\mathcal{A}, d, m) = (\mathbb{R}^2, \textit{Euclidiana}, \textit{centroide})$;

$K = 2$;

$D = \{(1,1)^T, (1,2)^T, (1,3)^T, (2,1)^T, (2,4)^T, (3,4)^T, (4,4)^T\}$.