Métricas ponto a ponto (continuação)

Dados do tipo monóide

Considere-se o espaço de atributos $\mathcal{A} = A \times A \times A \times ... \times A = A^l$. onde $A = \{\epsilon, a, b, c, ..., z\}$.

Para trabalhar este tipo de dados, vamos considerar o caso em que $l = 5$. Ou seja consideramos palavras até 5 letras.

Uma função dissemelhaça poderá ser,
$d(x, y) =$ número de letras de x que são diferentes de x'.

Exemplo 16: Mostrar que é uma dissemelhança não simétrica, não está normalizada, e não verifica difinitness.

Usar definição de função de dissemelhança e arranjar contra exemplo que mostre que $d(x, y) \neq d(y, x)$. Mostrar que pode ter valores fora de $[0, 1]$ e que existem casos que $d(x, y) = 0$ e $x \neq y$.

Uma outra função de semelhança poderá ser,

$$s(x,y)=\frac{1}{5}(s1(x,y)+s2(x,y)+s3(x,y)+s4(x,y)+s5(x,y))$$

onde,

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

Exemplo 17: Verificar se é uma semelhança simétrica, se está normalizada, e se verifica difinitness.

Exemplo 18: Calcule $s('aula','aulas')$, $s('aulas','aula')$.

Métricas com dados Reais

$\mathcal{A} = \mathbb{R}^I$, $x = (x_1, x_2, ..., x_I) \in \mathcal{A}$, é elemento genérico que pode ser evento e^n duma base de dados D.

Métrica de Minkowski :
Seja $r > 0$. A Métrica de Minkowski é:

$$d_r(x, y) = (\sum_{i=1}^{I} |x_i - y_i|^r)^{1/r}.$$

De acordo com os valores de r obtêm-se diferentes métricas. Se $r \in]0, 1[$ não se verifica a desigualdade triangular.

Exemplo 19: Considere $r = 1/4$; $x = (0, 0)$, $y = (0, 1)$ e $z = (1, 1)$. Verifique se a métrica em causa verifica a desigualdade triangular.

$d_{\frac{1}{4}}(x, y) = 1$; $d_{\frac{1}{4}}(y, z) = 1$; $d_{\frac{1}{4}}(x, z) = 16$;

Logo, $d_{\frac{1}{4}}(x, z) \leq d_{\frac{1}{4}}(x, y) + d_{\frac{1}{4}}(y, z)$ é falso.

Métricas com dados Reais

Distância de Manhattan: Fazer $r = 1$. Obtém–se,

$$d_1(x, y) = \sum_{i=1}^{I} |x_i - y_i|$$

Exemplo 20: $x, y \in \mathbb{R}^2$. Seja $x = (1, 1)$ e $y = (2, 2)$. Calcule a distância de Manhattan entre x e y.

$$d_1(x, y) = |1 - 2| + |1 - 2| = 2.$$

Distância de Euclidiana:

Fazer $r = 2$. Obtém–se,

$$d_2(x, y) = \sqrt{\sum_{i=1}^{I} |x_i - y_i|^2}.$$

Exemplo 21: Mostrar que a métrica Euclidiana é uma distância.

Métricas em Machine Learning

É necessário mostrar que,

a) $d_2(x, y) \geq 0$. (positividade)
b) $d_2(x, y) = d_2(y, x)$. (simetria)
c) $d_2(x, y) = 0 \equiv x = y$. (definitness)
d) $d_2(x, z) \leq d_2(x, y) + d_2(y, z)$ (desigualdade triangular).

Também se consegue provar que,

a) Compatibilidade com a adição - $d_2(0, x) = \sqrt{x_1^2 + x_2^2}$ e $d_2(y, x + y) = \sqrt{x_1^2 + x_2^2}$.

b) Compatibilidade com a multiplicação por cosntante λ,

$d_2(\lambda x, \lambda y) = |\lambda| d_2(x, y)$.

Distância de Chebyshev: Fazer $r \to \infty$. Obtém-se,

$d_\infty(x, y) = max|x_i - y_i|$ com $i = 1, ...I$.

Exemplo 22:

Sejam $x, y \in \mathbb{R}^2$, onde $x = (1, 1)$ e $y = (2, 3)$. Calcule $d_\infty(x, y)$.

Solução : 2.

Dissemelhaça de Rook : Considerando $\mathcal{A} = \mathbb{R}^2$, define-se como:

$$d_{Rook}\,(x, y) = \begin{cases} |x_2 - y_2| & se\ x_1 = y_1 \\ |x_1 - y_1| & se\ x_2 = y_2 \\ +\infty & se\ noutros\ casos \end{cases}$$

Exemplo 23: Mostre que d_{Rook} não verifica a desigualdade triangular.

3. Métricas para sub conjuntos (subset metrics)

- A idéia é usar métricas ponto a ponto, para definir métricas para sub conjuntos.
- O cálculo do representante dum conjunto C pode ser importante. Este vai depender da métrica usada + operações algébricas possíveis no espaço dos atributos.

Dados reais - representante dum conjunto

- $\mathcal{A} = \mathbb{R}$, $D = \{x^n : n = 1, ...N\}$ onde $x^n \in \mathbb{R}$.
- Vamos assumir que os dados são ordenados.
- Para determinar o representante $m \in \mathcal{A}$ dum conjunto C vamos definir a funcional,

$$E(m; C) = \sum_{n=1}^{N} (d_2(m, x^n))^2 = \sum_{n=1}^{N} (x^n - m)^2.$$

se usarmos a métrica Euclidiana.

ou seja,

$$E(m; C) = (d_2(m, x^1))^2 + (d_2(m, x^2))^2 + ... + (d_2(m, x^N))^2$$

O objetivo é determinar m que minimiza E,

$$\overline{m} = \arg\min_m E(m; C).$$

Usando uma métrica Euclidiana vem,

$$E(m; C) = (x^1 - m)^2 + (x^2 - m)^2 + ... + (x^N - m)^2$$

Derivando e igualando a 0 vem,

$$-2(x^1 - m) - 2(x^2 - m) - ... - 2(x^N - m) = 0.$$

Resolvendo em ordem a m,

$$m = \frac{1}{N} \sum_{i=1}^{N} x^i,$$

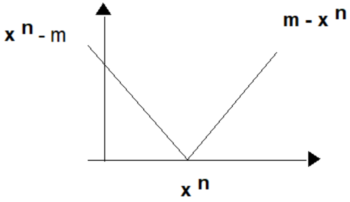
ou seja, o representante de C é o valor médio dos valores de C.

Métricas em Machine Learning

Se a métrica usada fôr a de Manhattan, temos,

$$E(m, C) = \sum_{n=1}^{N} |m - x^n|.$$

Notar que a funcional não é derivável.

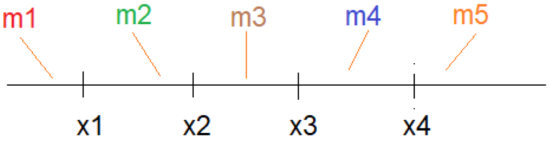


Notar que à esquerda de x^n é estritamente decrescente com declive -1, e à direita de x^n é estritamente decrescente com declive 1. Podemos definir,

$$E'(m; C) = \sum_{n=1}^{N} sign(m - x^n), \text{ com } m \neq x^1, x^2, ...x^N, \text{ onde } sign \text{ é -1 ou 1.}$$

Exemplo 24: Seja $C = \{x^1, x^2, x^3, x^4\}$ onde $x^n \in \mathbb{R}$. Calcule representante de C usando a métrica de Manhattan.

Pretende-se então ver qual o valor de m para o qual a funcional é mínima.



$E'(m1; C) = -4$, $E'(m2; C) = -2$, $E'(m3; C) = 0$, $E'(m4; C) = 2$, $E'(m5; C) = 4$.

Assim, se temos $C = \{x^1, x^2, ..., x^N\}$, com $x^n \in \mathbb{R}$ podemos definir como **representantes** de C, o valor médio com métrica Euclidana. Poderá ser interpretado como mediana se usamos a métrica de Manhattan. (média mto. diferente de mediana $\rightarrow$ outlyers)

Métricas entre subconjuntos (Inter cluster metrics)

Existem na literatura várias métricas entre dois sub conjuntos (clusters). Apresentamos algumas, e denominamos de dd.

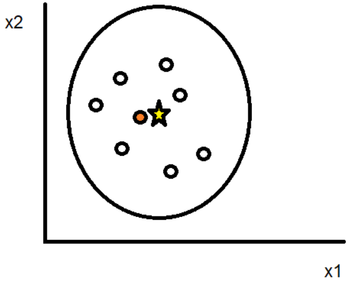
1) Sejam m e m' os representantes de C e C', então $dd(C, C') = d(m, m')$. Neste caso estamos a usar uma métrica ponto a ponto entre os representantes dos clusters.

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'.

Contróide versus Medóide



Centróide é o 'centro de massa' ; Medóide pertence ao conjunto.

Métricas em Machine Learning

Definição : $\overline{m} \in \mathcal{A}$ é um **centróide** de C se $\forall x \in \mathcal{A}\ E(x; C) \geq E(\overline{m}; C)$, escreve-se,

$$\overline{m} = \arg\min_{x \in \mathcal{A}} E(x; C)$$

Definição : $\overline{m} \in C$ é um **medóide** de C se $\forall x \in C\ E(x; C) \geq E(\overline{m}; C)$, escreve-se,

$$\overline{m} = \arg\min_{x \in C} E(x; C)$$

Notar que no segundo caso $x \in C$.

Dados em $\mathbb{R}^I$ - representante do tipo centróide

$\mathcal{A} = \mathbb{R}^I$, $C = \{x^n : \ n = 1, ...N\}$ onde $x_i^n \in \mathbb{R}$ representa o atributo $i = 1, ...I$ do evento $n = 1, ..., N$.

Neste caso, definimos a função custo,

$$f(m) = \sum_{n=1}^{N} d(m, x^n) \quad \text{, onde ,}$$

$$d(m, x^n) = \sum_{i=1}^{I} (m_i - x_i^n)^2 \text{ (se } d \text{ fôr distância Euclidiana) .}$$

Assim f representa a soma para todos os eventos de C da distância do representante de C que é m ao elemento de $x_i^n \in C$.

Como os elementos de C têm I componentes, usando a **distância Euclidiana**, fazemos a soma para todas as componentes de $(m_i - x_i^n)^2$.

Assim, podemos escrever a função custo,

$$f(m) = \sum_{n=1}^{N} \sum_{i=1}^{I} (m_i - x_i^n)^2 = \sum_{i=1}^{I} \sum_{n=1}^{N} (m_i - x_i^n)^2.$$

Se definirmos, $f_i(m_i) = \sum_{n=1}^{N} (m_i - x_i^n)^2$, para encontrar o representante de C, temos que encontrar $m \in \mathbb{R}^I$ que minimiza f, que é o mesmo que **encontrar os** $m_i \in \mathbb{R}$ **que minimizam** f_i.

Assim, para cada componente temos o problema de otimização semelhante ao caso 1D, que vimos ter como solução, o valor médio. Podemos então escrever,

$$\overline{m} = \frac{1}{N} \begin{bmatrix} \sum_{n=1}^{N} x_1^N \\ \sum_{n=1}^{N} x_2^N \\ .... \\ \sum_{n=1}^{N} x_I^N \end{bmatrix}$$

m que é o representante de C é um centróide(normalmente não vai dar um elemento de C) .

Se optarmos por usar a **Métrica de Manhattan**, temos que a função custo é agora,

$$f(m) = \sum_{n=1}^{N} \sum_{i=1}^{I} |m_i - x_i^n| = \sum_{i=1}^{I} \sum_{n=1}^{N} |m_i - x_i^n|.$$

Se definirmos, $f_i(m_i) = \sum_{n=1}^{N} |m_i - x_i^n|$, para encontrar o representante de C, temos que encontrar $m \in \mathbb{R}^I$ que minimiza f, que é o mesmo que **encontrar os** $m_i \in \mathbb{R}$ **que minimizam** f_i. Vimos para o caso 1D que o representante é a mediana. Assim, neste caso, escrevemos,

$$\overline{m} = \begin{bmatrix} \overline{m}_1 \\ \overline{m}_2 \\ \dots. \\ \overline{m}_N \end{bmatrix}$$

onde $\overline{m}_i$ é a mediana dos x_i^n.

Exemplo 24 : Considere $C = \{x^1, x^2, x^3, x^4, x^5\}$ onde $x^1 = (0,0)^T$, $x^2 = (2,1)^T$ e $x^3 = (2,4)^T$, $x^4 = (1,-2)^T$ e $x^5 = (-2,4)^T$.

a) Calcule o representante de C usando a métrica Euclidiana.
b) Calcule o representante de C usando a métrica de Manhattan.

Dados em $\mathbb{R}^l$ - representante do tipo medóide

O raciocínio é o mesmo. temos que minimizar uma função custo $f(m)$, mas acrescentamos a restrição de que esse representante tem que ser membro do conjunto C. É computacionalmente mais pesado.

Assim, pretende-se minimizar a função custo,

$$f(m) = \sum_{n=1}^{N} d(m, x^n)$$

com a restrição de que $m \in C$. Ou seja,

$$\overline{m} = \arg\min_{x \in C} f(m)$$

Exemplo 25 : Considere $C = \{x^1, x^2, x^3, x^4, x^5\}$ onde $x^1 = (0,0)^T$, $x^2 = (2,1)^T$ e $x^3 = (2,4)^T$, $x^4 = (1,-2)^T$ e $x^5 = (-3,1)^T$.
Calcule o representante de C do tipo medóide, usando a métrica de Manhattan.