Mineração de Dados

Trabalho 4

Arthur do Prado Labaki - 11821BCC017

13-04, 2023

GBC212

Estudo do Conceito

As medidas que calculam a distancia entre objetos costumam ser denominadas medidas de dissimilaridade, que são as distancias Euclidiana, Manhattan e Minkowski. Também as medidas que calculam a proximidade entre objetos costumam ser denominadas medidas de similaridade, como SMC, cosseno e Jaccard. Explicando cada uma, temos:

A distância Euclidiana é uma medida de distância entre dois pontos em um espaço de n dimensões. É calculada como a raiz quadrada da soma dos quadrados das diferenças entre as coordenadas dos pontos. A distância Euclidiana é sempre um valor positivo e pode ser interpretada como a magnitude do vetor que liga os dois pontos.

A distância Manhattan é uma medida de distância entre dois pontos em um espaço de n dimensões. É calculada como a soma das diferenças absolutas entre as coordenadas dos pontos. A distância Manhattan é sempre um valor positivo e pode ser interpretada como a distância que um objeto teria que percorrer para se mover entre os dois pontos em uma cidade que segue um padrão de ruas quadradas.

A distância Minkowski é uma medida de distância entre dois pontos em um espaço de n dimensões. É uma generalização da distância Euclidiana e da distância Manhattan, que são casos especiais da distância Minkowski para p = 2 e p = 1, respectivamente. A distância Minkowski é calculada como a p-ésima raiz da soma das diferenças elevadas a p entre as coordenadas dos pontos. A distância Minkowski pode ser usada para definir diferentes tipos de distâncias, dependendo do valor de p escolhido.

O coeficiente de igualdade simples (também denominado Simple Matching Coefficient, ou SMC) é uma medida de similaridade entre duas variáveis nominais ou categóricas. É calculado como o número de categorias que as duas variáveis têm em comum dividido pelo número total de categorias. O coeficiente de igualdade simples varia de 0 a 1, sendo que 0 indica nenhuma igualdade e 1 indica igualdade perfeita entre as variáveis.

O coeficiente de Jaccard é uma medida de similaridade entre dois conjuntos. É calculado como o número de elementos que estão presentes em ambos os conjuntos dividido pelo número total de elementos em pelo menos um dos conjuntos. O coeficiente de Jaccard varia de 0 a 1, sendo que 0 indica nenhum elemento em comum e 1 indica que os conjuntos são iguais.

A similaridade do cosseno é uma medida de similaridade entre dois vetores que representam pontos em um espaço de n dimensões. É calculada como o produto interno dos vetores dividido pelo produto da norma dos vetores. A similaridade do cosseno varia de -1 a 1, sendo que -1 indica vetores opostos e 1 indica vetores idênticos em direção.

Resolução do Exercício 1)

Para os seguintes objetos x e y, calcule a medida de proximidade indicada.

a) x=(1,1,1,1), y=(2,2,2,2): Euclidiana, cosseno

Distância Euclidiana:

Utilizando a formula, temos:
$d(x, y) = \sqrt{(x1 - y1)^2 + (x2 - y2)^2 + ... + (xn - yn)^2}$

Substituindo os valores, temos:
$d(x, y) = \sqrt{(1 - 2)^2 + (1 - 2)^2 + (1 - 2)^2 + (1 - 2)^2}$
$d(x, y) = \sqrt{4} = 2$

Portanto, a distância Euclidiana entre x e y é 2.

Similaridade do cosseno:

$sim(x, y) = (xy)/(||x|| * ||y||)$
Onde x · y é o produto interno entre os vetores x e y e $||x||$ e $||y||$ são as normas dos vetores x e y, respectivamente.

Substituindo os valores, temos:
$sim(x, y) = (1 * 2 + 1 * 2 + 1 * 2 + 1 * 2)/(\sqrt{1^2 + 1^2 + 1^2 + 1^2} * \sqrt{2^2 + 2^2 + 2^2 + 2^2})$
$sim(x, y) = 8/(2 * 4)$
$sim(x, y) = 1/2$

Portanto, a similaridade do cosseno entre x e y é 1/2 ou 0.5.

b) x=(0,1,0,1), y=(1,0,1,0): Euclidiana, SMC, Jaccard

Distância Euclidiana:

Utilizando a formula, temos:
$d(x,y) = \sqrt{(x1-y1)^2 + (x2-y2)^2 + ... + (xn-yn)^2}$

Substituindo os valores, temos:
$d(x,y) = \sqrt{(0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2}$
$d(x,y) = \sqrt{4}$
$d(x,y) = 2$

Portanto, a distância Euclidiana entre x e y é 2.

Coeficiente de igualdade simples:

Numero de atributos iguais a 0 = 0
Numero de atributos iguais a 1 = 0
Número de atributos onde em X é 0 e em Y é 1: 2
Número de atributos onde em X é 1 e em Y é 0: 2

Logo: SMC = (números de posições iguais / número de posições)
= 0 / 4 = 0

Portanto, o coeficiente de igualdade simples entre x e y é 0.

Coeficiente de Jaccard:

Considerando os mesmos números do SMC, temos que:
posições iguais a 1 / (posições diferentes + posições iguais a 1)
= 0 / (2 + 2 + 0) = 0

Portanto, o coeficiente de Jaccard entre x e y é 0.

c) x=(0,-1,0,1), y=(1,0,-1,0): Euclidiana, cosseno

Distância Euclidiana:

Utilizando a formula, temos:
$d(x,y) = \sqrt{(x1-y1)^2 + (x2-y2)^2 + ... + (xn-yn)^2}$

Substituindo os valores, temos:
$d(x,y) = \sqrt{(0-1)^2 + (-1-0)^2 + (0+1)^2 + (1-0)^2}$
$d(x,y) = \sqrt{4}$
$d(x,y) = 2$

Portanto, a distância Euclidiana entre x e y é 2.

Similaridade do cosseno:

$sim(x,y) = (xy)/(||x|| * ||y||)$
Onde x · y é o produto interno entre os vetores x e y e $||x||$ e $||y||$ são as normas dos vetores x e y, respectivamente.

Substituindo os valores, temos:
$sim(x,y) = (0*1 + -1*0 + 0*-1 + 1*0)/(\sqrt{0^2 + -1^2 + 0^2 + 1^2} * \sqrt{1^2 + 0^2 + -1^2 + 0^2})$
$sim(x,y) = 0/\sqrt{2} * \sqrt{2}$
$sim(x,y) = 0$

Portanto, a similaridade do cosseno entre x e y é 0.

d) x=(1,1,0,1,0,1), y=(1,1,1,0,0,1): SMC, Jaccard, cosseno

Coeficiente de igualdade simples:

Numero de atributos iguais a 0 = 1
Numero de atributos iguais a 1 = 3
Número de atributos onde em X é 0 e em Y é 1: 1
Número de atributos onde em X é 1 e em Y é 0: 1

Logo: SMC = (números de posições iguais / número de posições)

= 3 + 1 / 1 + 1 + 3 + 1 = 0.666

Portanto, o coeficiente de igualdade simples entre x e y é 0.666.

Coeficiente de Jaccard:

Considerando os mesmos números do SMC, temos que:
posições iguais a 1 / (posições diferentes + posições iguais a 1)
= 3 / 1 + 1 + 3 = 0.6

Portanto, o coeficiente de Jaccard entre x e y é 0.6.

Similaridade do cosseno:

$sim(x, y) = (xy)/(||x|| * ||y||)$
Onde x · y é o produto interno entre os vetores x e y e $||x||$ e $||y||$ são as normas dos vetores x e y, respectivamente.

Substituindo os valores, temos:
$sim(x, y) = (1 * 1 + 1 * 1 + 0 * 1 + 1 * 0 + 0 * 0 + 1 * 1)/(\sqrt{1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} * \sqrt{1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 1^2})$
$sim(x, y) = 3/(\sqrt{4} * \sqrt{4})$
$sim(x, y) = 3/4$

Portanto, a similaridade do cosseno entre x e y é 0.75.

e) x=(2,-1,0,2,0,-3), y=(-1,1,-1,0,0,-1): Euclidiana, cossen

Distância Euclidiana:

Utilizando a formula, temos:
$d(x, y) = \sqrt{(x1 - y1)^2 + (x2 - y2)^2 + ... + (xn - yn)^2}$

Substituindo os valores, temos:
$d(x, y) = \sqrt{(2 - (-1))^2 + (-1 - 1)^2 + (0 - (-1))^2 + (2 - 0)^2 + (0 - 0)^2 + (-3 - (-1))^2}$
$d(x, y) = \sqrt{22}$

$d(x, y) = 4.6904$

Portanto, a distância Euclidiana entre x e y é 4.6904.

Similaridade do cosseno:

$sim(x, y) = (xy)/(||x|| * ||y||)$
Onde x · y é o produto interno entre os vetores x e y e $||x||$ e $||y||$ são as normas dos vetores x e y, respectivamente.

Substituindo os valores, temos:
$sim(x, y) = (2 * (-1)) + (-1 * 1) + (0 * (-1)) + (2 * 0) + (0 * 0) + (-3 * (-1))/$
$(\sqrt{2^2 + (-1)^2 + 0^2 + 2^2 + 0^2 + (-3)^2} * \sqrt{(-1)^2 + 1^2 + (-1)^2 + 0^2 + 0^2 + (-1)^2})$
$sim(x, y) = 0/\sqrt{18} * \sqrt{4}$
$sim(x, y) = 0$

Portanto, a similaridade do cosseno entre x e y é 0.