T1 - Dissemelhanças com Dados Binários

Micro-Projeto

Lucas Mello, Diogo Lopes, Fábio Gonçalves
Joel Carvalho, Tiago Gonçalves, Pedro Ribeiro

Universidade do Minho

Objetivos

Objetivos

- Definição de uma Estrutura Binária;
- Métricas de Similaridade associadas a dados binários;
- Métricas de Distância associadas a dados binários;
- Análise Crítica, Estudos e Implementação Prática;
- Aplicações das Métricas abordadas.

Estrutura Binária

Estrutura Binária

Representamos uma base de dados binária com o seguinte exemplo: Um ambiente ecológico (x, y) é caracterizado por várias `espécies` de gramíneas onde $a_1, a_2, ..., a_n$, representa os n numero de `espécies` de gramíneas, é reportado o resultado na tabela seguinte, onde:

- o espaço dos Atributos é $\mathcal{A} = \{0, 1\}^n$ e
  $x = (x_1, x_2, ..., x_n)$,$y = (y_1, y_2, ..., y_n)$ onde $x_i, y_i \in \mathcal{A}$
- 1 e 0 Representam Presença e Ausência de um Atributo

n	$espcie_1$ a_1	$espcie_2$ a_2	$espcie_3$ a_3	$espcie_4$ a_4	....	$espcie_n$ a_n
x	1	1	0	0	...	1
y	1	0	1	0	...	1

Tabela 1: Tabela de dados de sítio ecológico

Matriz de Confusão

Para a aplicação das métricas de Distância e Similaridade, é necessário antes definir a Matriz de confusão definida por:

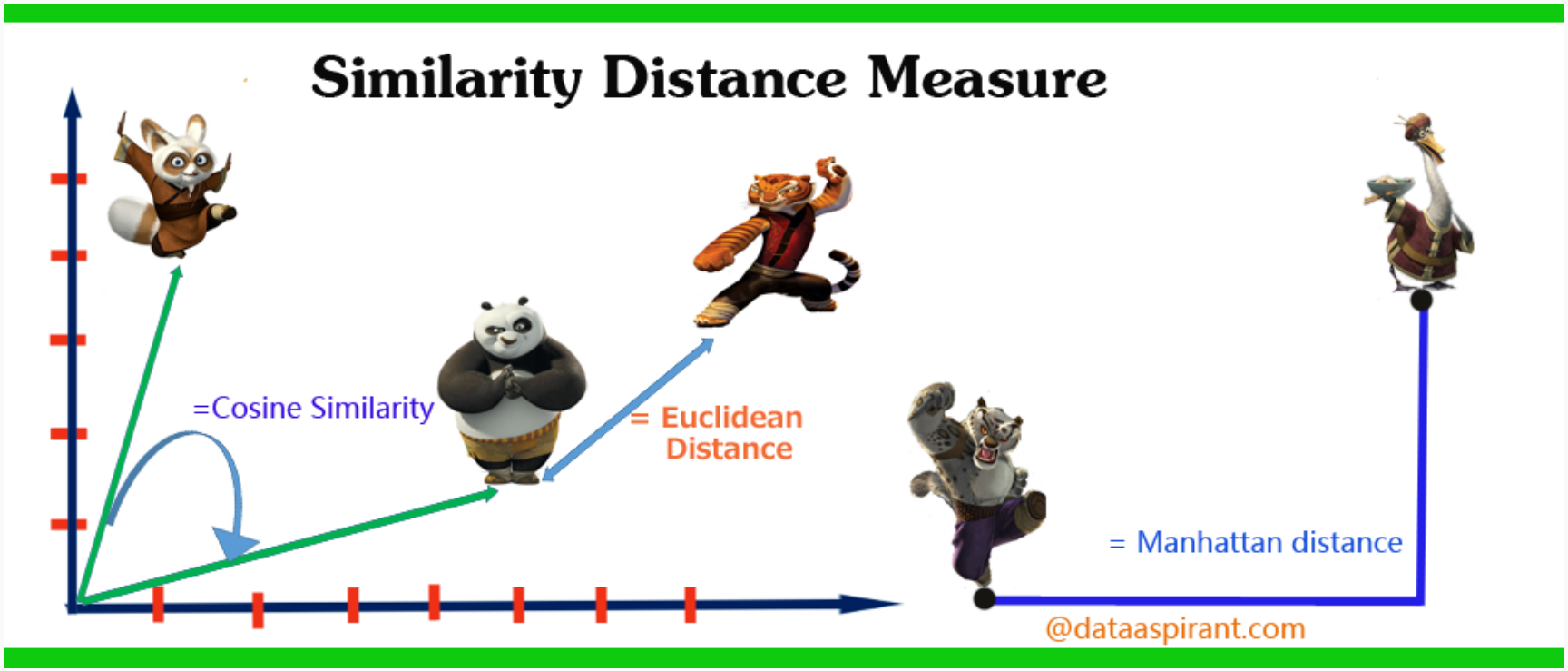
- $a = |\{i \in 1, ..., N; x_i = 1 \wedge y_i = 1\}| = \sum_{i=1}^{N} (x_i)(y_i);$
- $b = |\{i \in 1, ..., N; x_i = 1 \wedge y_i = 0\}| = \sum_{i=1}^{N} (x_i)(1 - y_i);$
- $c = |\{i \in 1, ..., N; x_i = 0 \wedge y_i = 1\}| = \sum_{i=1}^{N} (1 - x_i)(y_i);$
- $d = |\{i \in 1, ..., N; x_i = 0 \wedge y_i = 0\}| = \sum_{i=1}^{N} (1 - x_i)(1 - y_i);$

x/y	1	0
1	a	c
0	b	d

Tabela 2: Matriz de Confusão

Métricas associadas a Dados Binários

Métricas associadas a Dados Binários



Similaridade Binária

Similaridade Binária

- Definimos uma forma genérica de Similaridade através de:
  $x, y \in \mathcal{A} = \{0, 1\}$
  $x, y \Rightarrow s(x, y) \in \mathbb{R}$
- Deste modo, temos as seguintes propriedades satisfeitas:
  - $s(x, y) \in [0, 1]$
  - Simetria $s(x, y) = s(y, x)$
  - Normalização $s(x, x) = 1$
  - *Definiteness* $s(x, y) = 1 \Rightarrow x = y$
- Exemplo de Semelhanças Aditivas:
  - $s_i(x, y) = 1$ se $x_i = y_i, i = 1, 2, 3, 4$
  - $s(x, y) = \frac{1}{4} \sum_{i=1}^{4} s_i(x, y)$;

3 cenários possíveis de Semelhanças Aditivas:

- $x = (1, 0, 0, 1)$, $y = (0, 1, 0, 0)$, $s(x, y) = \frac{1}{4} * 0 = 0$
- $x = (1, 0, 0, 1)$, $y = (1, 1, 0, 1)$, $s(x, y) = \frac{1}{4} * 2 = 0.50$
- $x = (1, 1, 1, 1)$, $y = (1, 1, 1, 1)$, $s(x, y) = \frac{1}{4} * 4 = 1$

Métricas de similaridade

De seguida apresentamos 4 Métricas de Similaridade

- Similaridade de Sokal Michener
- Similaridade de Jaccard
- Similaridade de Dice
- Similaridade de Russel and Rao

A tabela 3 será utilizada como exemplo genérico para calcularmos e compararmos as diferentes métricas. Como isto obtemos: `a = 1, b = 1, c = 2 e d = 1`.

Objetos	Esfera	Doce	> 8cm	Crocante	Pesado
x = Maçã	1	1	0	1	0
y = Banana	0	1	1	0	0

Tabela 3: Exemplo geral para o cálculo da similaridade

Similaridade Sokal Michener

Similaridade de Sokal Michener

- $S_{SM} = \frac{a+d}{a+b+c+d} = \frac{atributos(correspondentes)}{atributos(total)}$
- Consiste na proporção de correspondências com o número total de valores.
- Peso atribuido de igual forma a correspondências e não correspondências.
- Bastante útil quando os valores positivos e negativos carregam informações simétricas/iguais.
- Semelhança simétrica: $S_{SM}(x_i, y_i) = S_{SM}(y_i, x_i)$.

Similaridade de Jaccard

- $S_{Jaccard} = \frac{a}{a+b+c}$
- $S_{SM} = 1$ quando valores de x_i e y_i são iguais a 1.
- $S_{SM} = 0$ quando $a = 0$.
- Relacionando-a com a S_{SM}, a S_{SM} caso $d = 0$, este valor não é contabilizado para variar a distância entre 2 objetos, apenas é valorizado quando os objetos são presentes.
- Semelhança simétrica: $S_{Jaccard}(x_i, y_i) = S_{Jaccard}(y_i, x_i)$.

Similaridade de Dice

Similaridade de Dice

- $S_{Dice} = \frac{2a}{2a+b+c}$
- Muito semelhante a Jaccard, porém estamos a duplicar a importância de a (TP).
- Se $a > 1$, significa que havendo TP damos-lhe muita importância.
- Consequentemente sendo $a < 1$, como 0.1, o b, c vão continuar a ter preponderância.
- Comparativamente com Jaccard, quando $a \neq 0$, Dice dá mais peso aos casos positivos.

Similaridade de Russel and Rao

- $S_{Russel_Rao} = \frac{a}{a+b+c+d}$
- Misto de Jaccard (numerador) e Sokal Michener (denominador).
- Peso atribuído de igual forma a correspondências e não correspondências.
- Ao contrário da Jaccard, caso $d = 1$, ou seja todos valores de x_i e y_i são iguais a 0, o valor é indeterminado, nesta métrica não existe valores indeterminados pela presença do atributo d na fórmula.

Comparação de Resultados

- **Sokal Michener** Através da aplicação de Sokal Michener, concluímos que a similaridade é igual a 0.4
- **Dice** Através da aplicação da Similaridade de Dice, concluímos que a similaridade é igual a 0.4
- **Jaccard** Através da aplicação da Similaridade de Jaccard, concluímos que a similaridade é igual a 0.25
- **Russel and Rao** Através da aplicação de Russel and Rao, concluímos que a similaridade é igual a 0.2
-

Sokal Michener	Jaccard	Dice	Russel and Rao
0.4	0.25	0.4	0.2

Tabela 4: Comparação entre os valores finais das diferentes similaridades

Distância Binária

Distância Binária

- Definimos uma forma genérica a Distância através de:
  $\forall x, y \in \mathcal{A} = \{0, 1\}$
  $d(x, y) \Rightarrow [0, +\infty]$
- Deste modo, temos as seguintes propriedades satisfeitas:
  - $d(x, y) \in [0, +\infty]$
  - Simetria $d(x, y) = d(y, x)$
  - Definiteness $d(x, y) = 0 \Rightarrow x = y$
- Exemplo de Distâncias Aditivas (Tomamos distância como sendo igual a $[0, 1]$:
  - $d_i(x, y) = 0$ se $x_i = y_i, i = 1, 2, 3, 4$
  - $d(x, y) = 1 - \sum_{i=1}^{4} \frac{d_i(x,y)}{n}$;

3 cenários possíveis de distâncias aditivas:

- $x = (1, 0, 0, 1)$, $y = (0, 1, 0, 0)$, $d(x, y) = 1 - \frac{0}{4} = 1$
- $x = (1, 0, 0, 1)$, $y = (1, 1, 0, 1)$, $d(x, y) = 1 - \frac{2}{4} = 0.5$
- $x = (1, 1, 1, 1)$, $y = (1, 1, 1, 1)$, $d(x, y) = 1 - \frac{4}{4} = 0$

Métricas de Distância

Apresentamos 4 métricas de distância

- Distância de Sokal Michener
- Distância de Hamming
- Distância Euclidiana
- Distância do Produto

A tabela 5 será utilizada como exemplo genérico para calcularmos e compararmos as diferentes métricas. Como isto obtemos: `a = 1, b = 1, c = 2 e d = 1`.

Objetos	Esfera	Doce	> 8cm	Crocante	Pesado
x = Maçã	1	1	0	1	0
y = Banana	0	1	1	0	0

Tabela 5: Exemplo geral para o cálculo da distância

Distância Sokal Michener

Distância de Sokal Michener

- $D_{SM} = 1 - S_{SM} = \frac{(b+c)}{n} = \left[0 - 1\right]$
- S_{SM}: se $D_{SM} = 0$ então a $S_{SM} = 1$.
- Distância simétrica, contudo pode não acontecer sempre.
- $D_{(X,Y)} = \frac{(b*1+c*10)}{n}$, então com a atribuição de pesos garantimos que não há simetria: $D(X, Y) <> D(Y, X)$.

Distância de Hamming

Distância Hamming

- $D_{Hamming} = b + c$
- O cálculo matemático é bastante próximo da D_{SM}.
- Pode tornar-se muito grande à medida que o número de atributos aumenta.
- Caso uma BD contenha 10 atributos e algum evento não possui nenhum destes atributos, então não haverá valor para `b` ou `c`, não sendo possível alcançar o valor de `n`.
- D_{SM} é normalizada e não é possível identificar atributos não definidos, enquanto que a $D_{Hamming}$ consegue fazer essa distinção.

Distância Euclidiana

- $D_{Euclid} = \sqrt{b + c} = \sqrt{D_{Hamming}}$
- Uma vez que a Distância Euclidiana é a raíz quadrada da Distância de Hamming, logo tudo o que vai ser detetado por Hamming vai ser detetado pela Euclidiana.
- Importância inferior dos valores FP e FN em comparação da distância de Hamming.

Distância do Produto

- $D_{Produto} = D(X, Y) = \sqrt{b.c}$
- X = Y = 0, uma vez que não existem FP nem FN, porém poderá haver a possibilidade da existência dos mesmos mas estes terão de ser considerados como fatores não importantes.
- Multiplicação de FP e FN, logo noções de igualdade e diferença são diferentes do comum. $(FP = 0 \text{ ou } FN = 0) \Rightarrow X = Y$

Comparação de Resultados

- **Distância Sokal Michener** Através da aplicação da Distância de Sokal-Michener, concluímos que a distância é igual a 0.6. Caso apliquemos a fórmula de atribuição de pesos então, D(X,Y) = 4.2 e D(Y,X) = 2.4.
- **Distância de Hamming** Através da aplicação da Distância de Hamming, concluímos que a distância é igual a 3
- **Distância Euclidiana** Através da aplicação da Distância Euclidiana, concluímos que a distância é aproximadamente 1.73.
- **Distância do Produto** Através da aplicação da Distância do Produto,concluímos que a distância é igual a 1.4

Sokal-Michener	Hamming	Euclidiana	Produto
0.6	3	1.73	1.4

Tabela 6: Comparação entre os valores finais das diferentes distâncias

Implementações e Benchmark

Implementações e Benchmark

Começámos por definir 4 vetores sintéticos, denotados por `x`, `y`, `z` e `w`.

x = (1 1 1 0 0 0 0 0 0 0)
y = (1 1 1 1 0 0 0 0 0 0)
z = (0 0 0 1 1 1 1 0 0)
w = (0 0 0 1 1 0 0 0 1)

Matrizes Confusão

x/y	1	0
1	3	1
0	0	5

x/z	1	0
1	0	4
0	3	2

x/w	1	0
1	0	3
0	3	3

Implementações e Benchmark

Similaridades

Matrizes	Sokal Michener	Jaccard	Dice	Russel and Rao
(x,y)	0.8889	0.75	0.8571	0.3333
(x,z)	0.2222	0	0	0
(x,w)	0.3333	0	0	0

Distâncias

Matrizes	Sokal Michener	Euclidiana	Hamming	Produto
(x,y)	0.1111	1	1	0
(x,z)	0.7778	2.6458	7	3.4641
(x,w)	0.6667	2.4495	6	3

Aplicações com as Bases de Dados

Base de Dados SCADI

Colunas	Atributo/Valor
Género	**Masculino - 1**
	Feminino - 0
Idade	5,6,7 ...
205 Atividades na ICF-CY	**Tem - 1**
	Não Tem - 0
Classes	Classe 1 até Classe 7

Aplicação das Métricas a SCADI

Similaridades

Matriz	Sokal-Michener	Jaccard
(x,y)	0.9508	0.7059

Distâncias

Matriz	Sokal-Michener	Euclidiana
(x,y)	0.0492	3.1623

Clusters de SCADI

Número de clusters: 7

Epsilon das Similaridades

Sokal-Michener	Jaccard
0.9	0.48

Epsilon das Distâncias

Sokal-Michener	Euclidiana
0.1	4.5

Base de Dados Emojis

Aplicação das Métricas a Emojis

Similaridades

Matriz	Sokal-Michener	Jaccard
(x,y)	0.8311	0.5422

Distâncias

Matriz	Sokal-Michener	Euclidiana
(x,y)	0.1689	6.1644

Clusters de Emojis

Número de clusters: 11

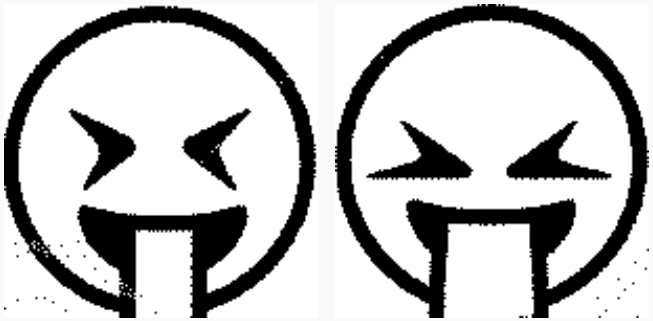
Epsilon das Similaridades

Sokal-Michener	Jaccard
0.9	0.645

Epsilon das Distâncias

Sokal-Michener	Euclidiana
0.1	4.7

Clusters de Emojis



Elementos 6 e 8

Elementos 20 e 23

Questões?

Obrigado pela atenção!

Lucas Mello, Diogo Lopes, Fábio Gonçalves
Joel Carvalho, Tiago Gonçalves, Pedro Ribeiro

Universidade do Minho