Conceitos gerais - representação vetorial, medidas de similaridade, normalização, noções de entropia e de informação mútua

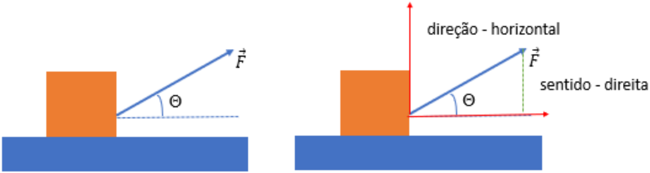
Prof. Dra. Sarajane Marques Peres

Fevereiro de 2020

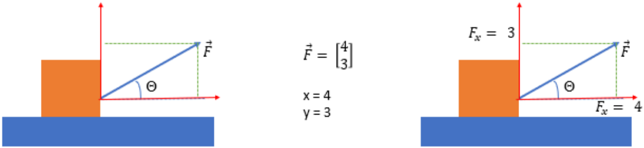
Disciplina: Inteligência Artificial
Bacharelado em Sistemas de Informação
http://www.each.usp.br/si

Espaço vetorial - vetor

- na Física: representa grandezas com valor (módulo), direção e sentido.



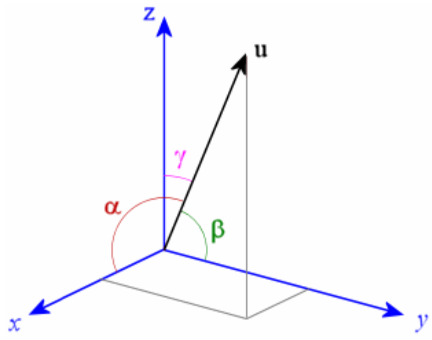
- na Matemática: o ente que representa o conjunto de segmentos orientados de uma reta que têm mesmo módulo, mesma direção e mesmo sentido.
- Matemática: qualquer matriz coluna.



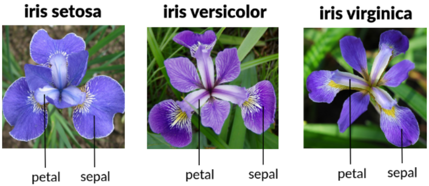
Espaço vetorial

Espaço vetorial

Conjunto não vazio, cujos elementos são chamados de vetores, com os quais podemos efetuar combinações lineares (representação de um vetor por meio de operações sobre outros vetores).



Representação vetorial - exemplo



THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

BY R. A. FISHER, Sc.D., F.R.S.

I. DISCRIMINANT FUNCTIONS

WHEN two or more populations have been measured in several characters, $x_1, \ldots, x_s$, special interest attaches to certain linear functions of the measurements by which the populations are best discriminated. At the author's suggestion use has already been made of this fact in craniometry (*a*) by Mr E. S. Martin, who has applied the principle to the sex differences in measurements of the mandible, and (*b*) by Miss Mildred Barnard, who showed how to obtain from a series of dated series the particular compound of cranial measurements showing most distinctly a progressive or secular trend. In the present paper the application of the same principle will be illustrated on a taxonomic problem; some questions connected with the precision of the processes employed will also be discussed.

II. ARITHMETICAL PROCEDURE

Table I shows measurements of the flowers of fifty plants each of the two species *Iris setosa* and *I. versicolor*, found growing together in the same colony and measured by Dr E. Anderson, to whom I am indebted for the use of the data. Four flower measurements are given. We shall first consider the question: What linear function of the four measurements

$$X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$$

will maximize the ratio of the difference between the specific means to the standard deviations within species? The observed means and their differences are shown in Table II. We may represent the differences by d_p, where $p = 1, 2, 3$ or 4 for the four measurements.

The sums of squares and products of deviations from the specific means are shown in Table III. Since fifty plants of each species were used these sums contain 98 degrees of freedom. We may represent these sums of squares or products by S_{pq}, where p and q take independently the values 1, 2, 3 and 4.

Then for any linear function, X, of the measurements, as defined above, the difference between the means of X in the two species is

$$D = \lambda_1 d_1 + \lambda_2 d_2 + \lambda_3 d_3 + \lambda_4 d_4,$$

while the variance of X within species is proportional to

$$S = \sum_{p=1}^{4} \sum_{q=1}^{4} \lambda_p \lambda_q S_{pq}.$$

The particular linear function which best discriminates the two species will be one for

Iris-dataset (criado por R. A. Fisher): Conjunto de dados contendo três classes de 50 instâncias (exemplares) cada, sendo que cada classe diz respeito a um tipo de planta Iris. Uma classe é linearmente separável das outras duas; essas duas não são linearmente separáveis entre si.

- atributos (medidas): sepal length em cm; sepal width em cm; petal length em cm; petal width em cm;
- classes: Iris Setosa; Iris Versicolour; Iris Virginica;

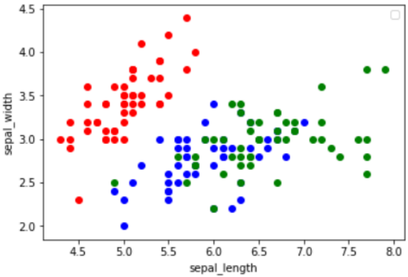
Representação vetorial - exemplo

```
5.0,3.4,1.5,0.2,Iris-setosa  7.0,3.2,4.7,1.4,Iris-versicolor  5.6,2.8,4.9,2.0,Iris-virginica
4.4,2.9,1.4,0.2,Iris-setosa  6.4,3.2,4.5,1.5,Iris-versicolor  7.7,2.8,6.7,2.0,Iris-virginica
4.9,3.1,1.5,0.1,Iris-setosa  6.9,3.1,4.9,1.5,Iris-versicolor  6.3,2.7,4.9,1.8,Iris-virginica
5.4,3.7,1.5,0.2,Iris-setosa  5.5,2.3,4.0,1.3,Iris-versicolor  6.7,3.3,5.7,2.1,Iris-virginica
4.8,3.4,1.6,0.2,Iris-setosa  6.5,2.8,4.6,1.5,Iris-versicolor  7.2,3.2,6.0,1.8,Iris-virginica
4.8,3.0,1.4,0.1,Iris-setosa  5.7,2.8,4.5,1.3,Iris-versicolor  6.2,2.8,4.8,1.8,Iris-virginica
4.3,3.0,1.1,0.1,Iris-setosa  6.3,3.3,4.7,1.6,Iris-versicolor  6.1,3.0,4.9,1.8,Iris-virginica
5.8,4.0,1.2,0.2,Iris-setosa  4.9,2.4,3.3,1.0,Iris-versicolor  6.4,2.8,5.6,2.1,Iris-virginica
5.7,4.4,1.5,0.4,Iris-setosa  6.6,2.9,4.6,1.3,Iris-versicolor  7.2,3.0,5.8,1.6,Iris-virginica
5.4,3.9,1.3,0.4,Iris-setosa  5.2,2.7,3.9,1.4,Iris-versicolor  7.4,2.8,6.1,1.9,Iris-virginica
```

Representação vetorial - exemplo

```
5.0,3.4,1.5,0.2,Iris-setosa  7.0,3.2,4.7,1.4,Iris-versicolor  5.6,2.8,4.9,2.0,Iris-virginica
4.4,2.9,1.4,0.2,Iris-setosa  6.4,3.2,4.5,1.5,Iris-versicolor  7.7,2.8,6.7,2.0,Iris-virginica
4.9,3.1,1.5,0.1,Iris-setosa  6.9,3.1,4.9,1.5,Iris-versicolor  6.3,2.7,4.9,1.8,Iris-virginica
5.4,3.7,1.5,0.2,Iris-setosa  5.5,2.3,4.0,1.3,Iris-versicolor  6.7,3.3,5.7,2.1,Iris-virginica
4.8,3.4,1.6,0.2,Iris-setosa  6.5,2.8,4.6,1.5,Iris-versicolor  7.2,3.2,6.0,1.8,Iris-virginica
4.8,3.0,1.4,0.1,Iris-setosa  5.7,2.8,4.5,1.3,Iris-versicolor  6.2,2.8,4.8,1.8,Iris-virginica
4.3,3.0,1.1,0.1,Iris-setosa  6.3,3.3,4.7,1.6,Iris-versicolor  6.1,3.0,4.9,1.8,Iris-virginica
5.8,4.0,1.2,0.2,Iris-setosa  4.9,2.4,3.3,1.0,Iris-versicolor  6.4,2.8,5.6,2.1,Iris-virginica
5.7,4.4,1.5,0.4,Iris-setosa  6.6,2.9,4.6,1.3,Iris-versicolor  7.2,3.0,5.8,1.6,Iris-virginica
5.4,3.9,1.3,0.4,Iris-setosa  5.2,2.7,3.9,1.4,Iris-versicolor  7.4,2.8,6.1,1.9,Iris-virginica
```

Plotando o conjunto de dados, considerando um espaço vetorial de duas dimensões:



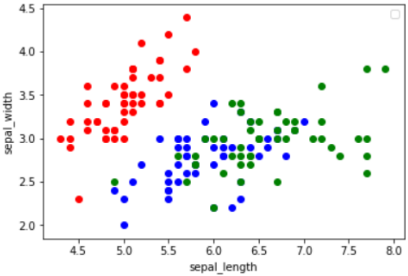
Representação vetorial - exemplo

sepal length sepal width

```
5.0,3.4,1.5,0.2,Iris-setosa
4.4,2.9,1.4,0.2,Iris-setosa
4.9,3.1,1.5,0.1,Iris-setosa
5.4,3.7,1.5,0.2,Iris-setosa
4.8,3.4,1.6,0.2,Iris-setosa
4.8,3.0,1.4,0.1,Iris-setosa
4.3,3.0,1.1,0.1,Iris-setosa
5.8,4.0,1.2,0.2,Iris-setosa
5.7,4.4,1.5,0.4,Iris-setosa
5.4,3.9,1.3,0.4,Iris-setosa
```

```
7.0,3.2,4.7,1.4,Iris-versicolor
6.4,3.2,4.5,1.5,Iris-versicolor
6.9,3.1,4.9,1.5,Iris-versicolor
5.5,2.3,4.0,1.3,Iris-versicolor
6.5,2.8,4.6,1.5,Iris-versicolor
5.7,2.8,4.5,1.3,Iris-versicolor
6.3,3.3,4.7,1.6,Iris-versicolor
4.9,2.4,3.3,1.0,Iris-versicolor
6.6,2.9,4.6,1.3,Iris-versicolor
5.2,2.7,3.9,1.4,Iris-versicolor
```

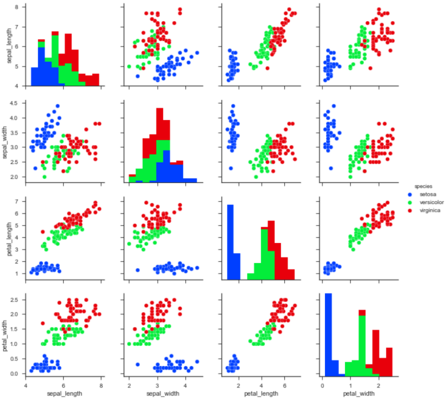
```
5.6,2.8,4.9,2.0,Iris-virginica
7.7,2.8,6.7,2.0,Iris-virginica
6.3,2.7,4.9,1.8,Iris-virginica
6.7,3.3,5.7,2.1,Iris-virginica
7.2,3.2,6.0,1.8,Iris-virginica
6.2,2.8,4.8,1.8,Iris-virginica
6.1,3.0,4.9,1.8,Iris-virginica
6.4,2.8,5.6,2.1,Iris-virginica
7.2,3.0,5.8,1.6,Iris-virginica
7.4,2.8,6.1,1.9,Iris-virginica
```



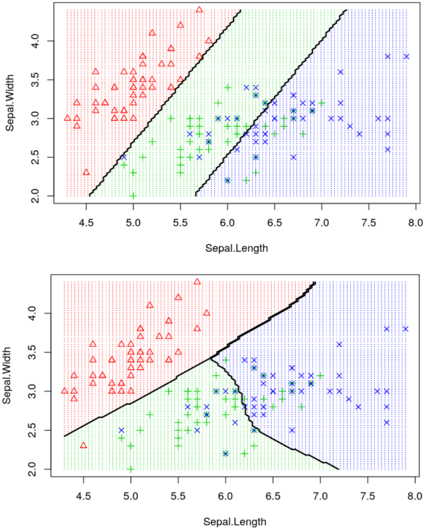
setosa versicolor virgínica

Representação vetorial - exemplo

Analisando a distribuição dos *datapoints* (exemplares - vetores) compondo o plano cartesiano usando como eixos diferentes combinações de atributos.



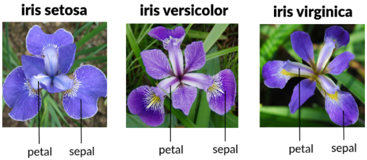
Separabilidade linear (?)

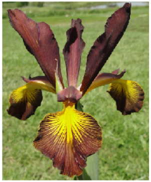


Similaridade

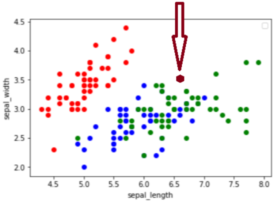
Por que estudar/medir similaridade?

Quando trabalhamos com análise de dados, em geral, nós estamos interessados em buscar por modelos que expliquem os dados com algum nível de generalização. De forma aplicada, queremos descobrir comportamentos ou perfis existentes dentro dos dados, ou queremos achar uma lei que os explique e permita tomar decisões que estão baseadas no que sabemos sobre eventos provenientes de um fenômeno.





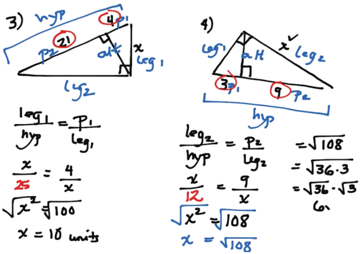
?



Medindo similaridade

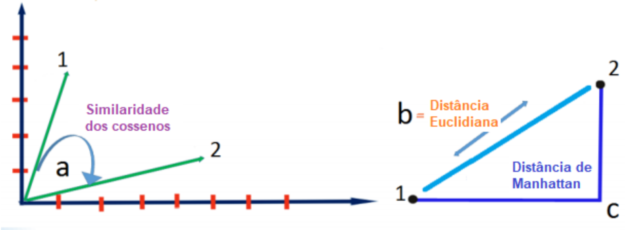
Como medir similaridade?

- Usando medidas de distância (p.ex.: distância Euclidiana ou distância de Manhattan ou medidas de ângulos - similaridade cosseno), nós podemos medir a similaridade sob uma ótica geométrica.
- Usando cálculo de entropia ou informação mútua, nós podemos medir a similaridade sob uma ótica informacional.



Medindo similaridade

- distância euclidiana: $(\sum |x_i - y_i|^2)^{\frac{1}{2}}$
- distância de Manhattan: $\sum |x_i - y_i|$
- similaridade dos cossenos: $\frac{\langle x,y \rangle}{\|x\|_2 \cdot \|y\|_2} = cos(\theta)$, sendo $\langle \ \rangle$ o produto interno $\| \ \|_2$ a norma euclidiana e $\cdot$ uma multiplicação.

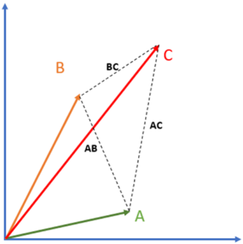


Medindo similaridade

Intuição da distância euclidiana e da similaridade dos cossenos.

Distância Euclidiana

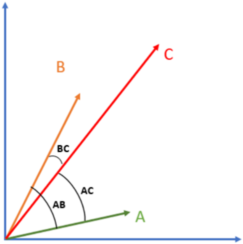
Retas: AC > AB > BC



Vetores A e C são os **mais distantes** entre si.

Similaridade de cossenos

Ângulos: AB > AC > BC

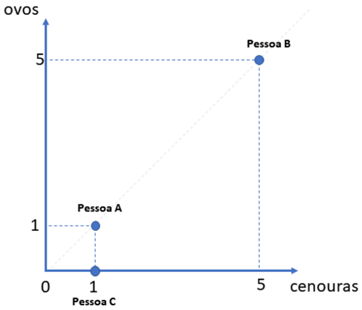


Vetores A e B são os mais distantes entre si.

- Maior ângulo
- Menor cosseno do ângulo (**menos** similares)
- Maior (1 - cosseno) do ângulo (**mais** distantes)

Medindo similaridade

Intuição da distância euclidiana e da similaridade dos cossenos.



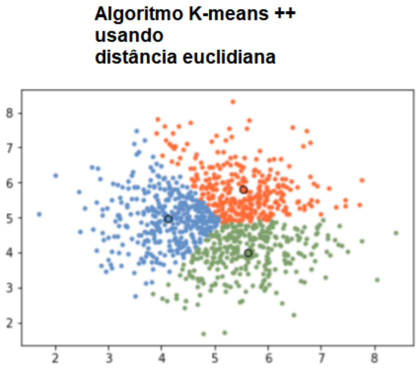
COMER	ovos	cenouras
Pessoa A	1	1
Pessoa B	5	5
Pessoa C	0	1

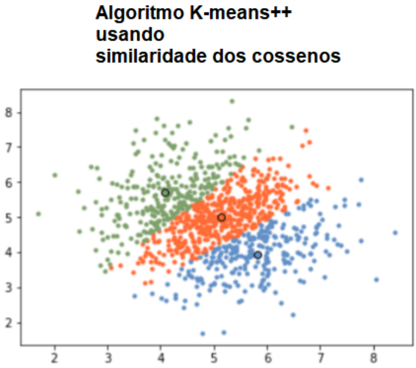
Sob análise da distância euclidiana, as pessoas A e C são mais parecidas entre si, do que elas são parecidas com a pessoa B.
A e C são pessoas comedidas. B é uma pessoa gulosa.

Sob a análise da similaridade dos cossenos, as pessoas A e B são mais parecidas entre si, do que são pareceidas com a pessoa C.
C é vegana, enquanto A e B não são.

Medindo similaridade

Intuição da distância euclidiana e da similaridade dos cossenos (resultado de um algoritmo de agrupamento).





Normalização

Normalização

Normalização é um procedimento de pré-processamento de dados cujo objetivo é escalar os valores dos atributos de forma que todos fiquem ou dentro de um intervalo específico, por exemplo [0, 1] ou [-1, 1], ou distribuídos em torno de sua média de acordo com seu desvio padrão.

Esse procedimento é especialmente útil quando os algoritmos de análise de dados são baseados em distância. Também é útil para acelerar o processo de "convergência" de um algoritmo de Machine Learning (como por exemplo, redes neurais artificiais).

Reescala e Standardizing

São procedimentos semelhantes. A reescala é aplicada para mudanças na unidade de medida, por exemplo, conversão de temperatura de graus Celsius para Graus Fahrenheit.

A standardizing subtrai uma medida de localização (como média) e divide por uma medida de escala (como o desvio). Ela é usada quando os dados seguem uma distribuição normal (Gaussiana) e queremos que eles sigam uma distribuição normal padrão (média $= 0$ e desvio $= 1$).

Normalização - motivação

```
dataset =

    5.0000  257.0000
    8.0000   73.6000
    2.0000  772.0000
    4.5000  901.5600
    9.3000   45.2000
    5.0000  764.7000
    2.1000  556.8000
    3.7000  878.0000
    6.0000  924.5000
    7.0000  114.5600
    2.0000  336.4000
    5.1000  722.0000
    1.7000  556.2000
    4.8000  777.0000
    5.5000  332.8000
```

(a) Conjunto de dados

```
x =

    5.0000
    8.0000
    2.0000
    4.5000
    9.3000
    5.0000
    2.1000
    3.7000
    6.0000
    7.0000
    2.0000
    5.1000
    1.7000
    4.8000
    5.5000
```

(b) Atributo X

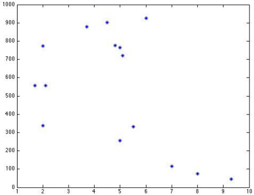
```
y =

  257.0000
   73.6000
  772.0000
  901.5600
   45.2000
  764.7000
  556.8000
  878.0000
  924.5000
  114.5600
  336.4000
  722.0000
  556.2000
  777.0000
  332.8000
```

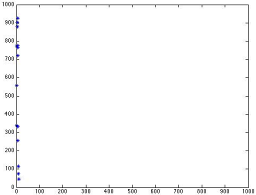
(c) Atributo Y

Normalização - motivação

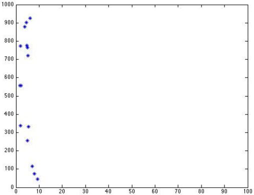
Plotando o conjunto de dados com os valores originais, usando diferentes escalas **para visualização** no eixo x (atributo X).



(d)



(e)



(f)

Normalização - motivação

Distância euclidiana entre exemplares (entre os vetores), considerando apenas o atributo X, apenas o atributo Y (vetores unidimensionais) e considerando ambos os atributos (vetores bidimensionais) - ilustrando apenas a parte inicial da matriz de distâncias:

0	3.0000	3.0000
3.0000	0	6.0000
3.0000	6.0000	0
0.5000	3.5000	2.5000
4.3000	1.3000	7.3000
0	3.0000	3.0000
2.9000	5.9000	0.1000
1.3000	4.3000	1.7000
1.0000	2.0000	4.0000

(g) Dist. X

0	183.4000	515.0000
183.4000	0	698.4000
515.0000	698.4000	0
644.5600	827.9600	129.5600
211.8000	28.4000	726.8000
507.7000	691.1000	7.3000
299.8000	483.2000	215.2000
621.0000	804.4000	106.0000
667.5000	850.9000	152.5000

(h) Dist. Y

0	183.4245	515.0087
183.4245	0	698.4258
515.0087	698.4258	0
644.5602	827.9674	129.5841
211.8436	28.4297	726.8367
507.7000	691.1065	7.8924
299.8140	483.2360	215.2000
621.0014	804.4115	106.0136
667.5007	850.9024	152.5524

(i) Dist. XY

Normalização - motivação

Distância euclidiana entre exemplares (entre os vetores), considerando apenas o atributo X, apenas o atributo Y (vetores unidimensionais) e considerando ambos os atributos (vetores bidimensionais) - ilustrando apenas a parte inicial da matriz de distâncias:

0	3.0000	3.0000
3.0000	0	6.0000
3.0000	6.0000	0
0.5000	3.5000	2.5000
4.3000	1.3000	7.3000
0	3.0000	3.0000
2.9000	5.9000	0.1000
1.3000	4.3000	1.7000
1.0000	2.0000	4.0000

0	183.4000	515.0000
183.4000	0	698.4000
515.0000	698.4000	0
644.5600	827.9600	129.5600
211.8000	28.4000	726.8000
507.7000	691.1000	7.3000
299.8000	483.2000	215.2000
621.0000	804.4000	106.0000
667.5000	850.9000	152.5000

0	183.4245	515.0087
183.4245	0	698.4258
515.0087	698.4258	0
644.5602	827.9674	129.5841
211.8436	28.4297	726.8367
507.7000	691.1065	7.8924
299.8140	483.2360	215.2000
621.0014	804.4115	106.0136
667.5007	850.9024	152.5524

O atributo Y tem muito mais influência nos valores de distância do que o atributo X. Algoritmos que se basearem na distância serão influenciados pelos atributo Y - gerando um viés.

Normalização

Normalização Min-max

Trata-se de uma transformação linear sobre os valores originais de um atributo A. Sendo min_A e max_A os valores mínimos e máximos de um atributo, o procedimento mapeia um valor v de A para v' no intervalo $[new\_min_A, new\_max_A]$, estabelecidos pelo analista de dados, computando:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_maxA - new\_minA) + new\_min_A$$

Essa transformação preserva o relacionamento entre os valores originais.

Observe que os valores min_A e max_A precisam ser defindos com cuidado, ou uma entrada futura pode cair fora desses intervalos e causar um problema na preservação dos relacionamentos originais.

Os min_A e max_A precisam ser armazenados para que possam ser usados na normalização de novos exemplares.

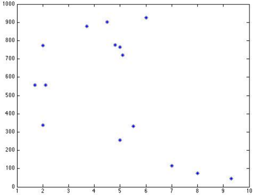
Normalização

Plotando o conjunto de dados normalizado - Min-max - intervalo [0,1]

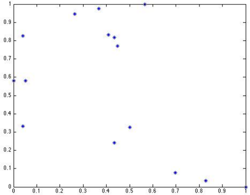
```
dataset_minmax =

    0.4342    0.2409
    0.8289    0.0323
    0.0395    0.8266
    0.3684    0.9739
    1.0000         0
    0.4342    0.8183
    0.0526    0.5818
    0.2632    0.9471
    0.5658    1.0000
    0.6974    0.0789
    0.0395    0.3312
    0.4474    0.7697
         0    0.5811
    0.4079    0.8323
    0.5000    0.3271
```

(j) Min-max



(k) Original



(l) Normalizado

Os valores de mínimo e máximo do atributo A foram determinados dentre os valores existentes no atributo.

$min_x = 1.7$

$min_y = 45.2$

$max_x = 9.3$

$max_y = 924.5$

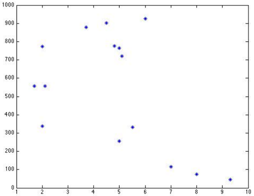
Normalização

Plotando o conjunto de dados normalizado - Min-max - intervalo [0,1]

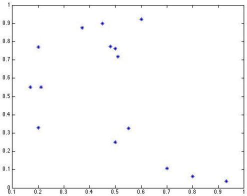
```
dataset_minmax_b =

    0.5000    0.2495
    0.8000    0.0642
    0.2000    0.7697
    0.4500    0.9006
    0.9300    0.0356
    0.5000    0.7623
    0.2100    0.5523
    0.3700    0.8768
    0.6000    0.9237
    0.7000    0.1056
    0.2000    0.3297
    0.5100    0.7192
    0.1700    0.5517
    0.4800    0.7747
    0.5500    0.3261
```

(m) Min-max



(n) Original



(o) Normalizado

Os valores de mínimo e máximo do atributo A foram determinados nos limites do domínio dos atributos.

$min_x = 0$

$min_y = 10$

$max_x = 10$

$max_y = 1000$

Normalização

Considerando cada um dos casos de escolha dos valores de mínimo e máximo do atributo A, e tomando como entrada para a normalização, um novo exemplar:

novo exemplar = (10, 999)

novo exemplar x = 10 e novo exemplar y = 999

Seguindo a primeira normalização (mínimo e máximo dentro dos valores dos exemplares existentes):

x_norm = 1.09

y_norm = 1.08

Seguindo a segunda normalização (mínimo e máximo dentro dos limites do domínio):

x_norm = 1

y_norm = 0.999

Normalização

Distância euclidiana entre exemplares (entre os vetores), considerando o conjunto de dados original e o conjunto de dados normalizado (Minmax)

0	183.4245	515.0087
183.4245	0	698.4258
515.0087	698.4258	0
644.5602	827.9674	129.5841
211.8436	28.4297	726.8367
507.7000	691.1065	7.8924
299.8140	483.2360	215.2000
621.0014	804.4115	106.0136
667.5007	850.9024	152.5524

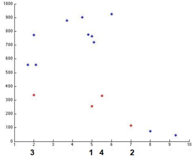
(p) Dist. XY

0	0.4465	0.7063
0.4465	0	1.1199
0.7063	1.1199	0
0.7360	1.0482	0.3604
0.6149	0.1741	1.2672
0.5774	0.8795	0.3948
0.5117	0.9511	0.2451
0.7267	1.0756	0.2541
0.7704	1.0028	0.5542

(q) Dist. XY norm

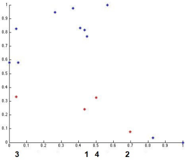
Normalização

Considere os exemplares (originais): (5 257) (7 114,56) (2 336,4) (5,5 332,8) - em vermelho no gráfico. Observe as distâncias (euclidiana).



0	142.4540	79.4567	75.8016
142.4540	0	221.8963	218.2452
79.4567	221.8963	0	5.0210
75.8016	218.2452	5.0210	0

Considere os exemplares (agora normalizados): (0,4342 0,2409) (0,6974 0,0789) (0,0395 0,3312) (0,5000 0,3271) - em vermelho no gráfico. E observe as distâncias (euclidiana).



0	0.3091	0.4049	0.1084
0.3091	0	0.7046	0.3171
0.4049	0.7046	0	0.4605
0.1084	0.3171	0.4605	0

Normalização

Outras opções, considerando a normalização de um atributo A:

- escalonamento decimal: $v' = \frac{v}{10^j}$, em que j é igual a 1 se o maior valor absoluto no conjunto de valores do atributo A é < 10, é igual a 2 se o maior valor absoluto no conjunto de valores do atributo A é ≥ 10 e < 100, e assim por diante.
- z-score (*standardizing*): $v' = \frac{v - \bar{A}}{\sigma_A}$, em que $\bar{A}$ é a média dos valores existentes no atributo A e σ_A é o desvio padrão do mesmo conjunto de valores.
- desvio absoluto da mediana (indicado quanto temos outliers): $v' = \frac{v - medianaA}{MAD}$, em que $MAD = mediana\{|v_i - mediana\{A\}|\}$.

Entropia

A entropia é uma grandeza termodinâmica que mede o grau de irreversibilidade de um sistema. É comumente associada ao que se entende por "desordem" (não em senso comum) de um sistema termodinâmico.

Exemplo

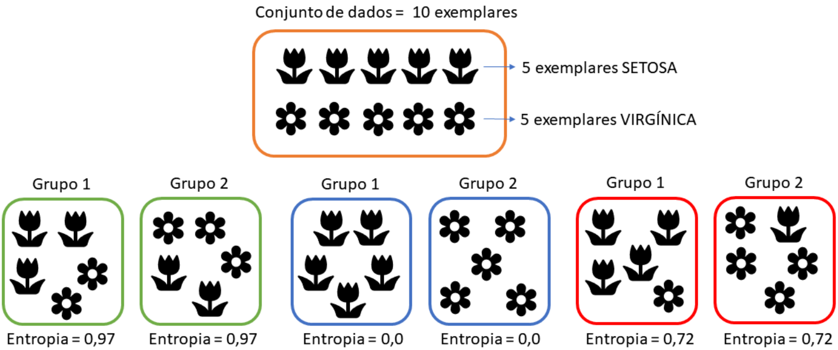
- antes de se misturarem tem-se
  - café
  - leite
  - açucar/adoçante
- depois de se misturarem teremos apenas uma informação
  - café com leite adoçado

(r) Baixa entropia - muita informação

(s) Alta entropia - pouca informação

Entropia

Medindo a entropia de grupos em um agrupamento, contra um *ground truth*:



	|s|	|v|	|Cx|	p(s)	log(p(s))	p(s)*log(p(s))	p(v)	log(p(v))	p(v)*log(p(v))	Entropia - soma()	
Cluster 1	3	2	5	0,60	-0,74	-0,44	0,40	-1,32	-0,53	0,97	
Cluster 2	2	3	5	0,40	-1,32	-0,53	0,60	-0,74	-0,44	0,97	
Cluster 1	5	0	5	1,00	0,00	0,00	1,00	0,00	0,00	0,00	Com Laplace
Cluster 2	0	5	5	1,00	0,00	0,00	1,00	0,00	0,00	0,00	
Cluster 1	4	1	5	0,80	-0,32	-0,26	0,20	-2,32	-0,46	0,72	
Cluster 2	1	4	5	0,20	-2,32	-0,46	0,80	-0,32	-0,26	0,72	

Entropia

Estudando entropia por meio da resolução de uma tarefa: discretização

Considere D um conjunto de dados (tuplas/linhas) definido por um conjunto de atributos e um atributo de rótulo de classe. O atributo de rótulo de classe fornece a informação sobre a classe, por dado (tupla/linha).

```
5.0, 3.4, 1.5, 0.2, Iris-setosa      7.0, 3.2, 4.7, 1.4, Iris-versicolor
4.4, 2.9, 1.4, 0.2, Iris-setosa      6.4, 3.2, 4.5, 1.5, Iris-versicolor
4.9, 3.1, 1.5, 0.1, Iris-setosa      6.9, 3.1, 4.9, 1.5, Iris-versicolor
5.4, 3.7, 1.5, 0.2, Iris-setosa      5.5, 2.3, 4.0, 1.3, Iris-versicolor
```

O método básico para discretização baseada em entropia de um atributo A dentro do conjunto consiste de:

1. Cada valor de A pode ser considerado como um potencial limite de intervalo ou ponto de divisão (*split-point*) para particionar os valores de A. Ou seja, o ponto de divisão para A divide os dados de D em dois subconjuntos que satisfazem as condições $A \leq split\_point$ e $A > split\_point$, criando uma discretização binária.

Entropia

A

```
5.0, 3.4, 1.5, 0.2, Iris-setosa
4.4, 2.9, 1.4, 0.2, Iris-setosa
4.9, 3.1, 1.5, 0.1, Iris-setosa
5.4, 3.7, 1.5, 0.2, Iris-setosa
...
```

```
...
7.0, 3.2, 4.7, 1.4, Iris-versicolor
6.4, 3.2, 4.5, 1.5, Iris-versicolor
6.9, 3.1, 4.9, 1.5, Iris-versicolor
5.5, 2.3, 4.0, 1.3, Iris-versicolor
```

A > 5.0

Parte 1

```
5.0, 3.4, 1.5, 0.2, Iris-setosa
4.4, 2.9, 1.4, 0.2, Iris-setosa
4.9, 3.1, 1.5, 0.1, Iris-setosa
```

Parte 2

```
5.4, 3.7, 1.5, 0.2, Iris-setosa
7.0, 3.2, 4.7, 1.4, Iris-versicolor
6.4, 3.2, 4.5, 1.5, Iris-versicolor
6.9, 3.1, 4.9, 1.5, Iris-versicolor
5.5, 2.3, 4.0, 1.3, Iris-versicolor
```

A > 5.4

Parte 1

```
5.0, 3.4, 1.5, 0.2, Iris-setosa
4.4, 2.9, 1.4, 0.2, Iris-setosa
4.9, 3.1, 1.5, 0.1, Iris-setosa
5.4, 3.7, 1.5, 0.2, Iris-setosa
```

Parte 2

```
7.0, 3.2, 4.7, 1.4, Iris-versicolor
6.4, 3.2, 4.5, 1.5, Iris-versicolor
6.9, 3.1, 4.9, 1.5, Iris-versicolor
5.5, 2.3, 4.0, 1.3, Iris-versicolor
```

Entropia

2 Suponha que nós queremos classificar as tuplas de D particionando-as no atributo A usando algum *split-point*. Idealmente, nós gostaríamos que esta partição resultasse na classificação exata dos dados (tuplas/linhas). Ou seja, se nós tivéssemos duas classes, nós esperaríamos que todas as tuplas da classe C_1 caíssem em uma parte, e todas as tuplas de C_2 caíssem em outra parte.

Sendo isso improvável, **quanta informação é ainda necessária** para obter uma classificação perfeita após esse particionamento? Esse montante é chamado de **expected information requirement** para classificar um dado de D baseado no particionamento de A.

E é dado por ....

Entropia

$$InfoNecessaria_A(D) = \frac{|D_1|}{|D|}Entropy(D_1) + \frac{|D_2|}{|D|}Entropy(D_2),$$

em que

- D_1 e D_2 correspondem aos dados (tuplas/linhas) em D que satisfazem as condições $A \leq split\_point$ e $A > split\_point$, respectivamente;
- $|D|$ é o número de dados no conjunto D ($|.|$ é o número de dados);
- a função $Entropy(.)$ para um dado conjunto é calculada com base na distribuição das classes dentro deste conjunto. Assim, dado m classes, C_1, C_2, ..., C_m, a entropia de D_1 é ...

Entropia

$$Entropy(D_1) = -\sum_{i=1}^{m} p_i log_2(p_i),$$

em que

- p_i é a probabilidade da classe C_i em D_1, determinada dividindo-se o número de dados da classe C_i em D_1 por $|D_1|$.

Portanto, quando selecionamos um ponto de divisão para o atributo A, nós queremos escolher o valor de atributo que nos dá o mínimo *expected information requirement* (i.e., $\min(Info_A(D))$).
Isto resultaria no montante mínimo de informação esperada AINDA necessária para classificar perfeitamente os dados depois de usar esta partição. **Isto é equivalente a escolher o par "valor-atributo" com o máximo ganho de informação.**

Entropia

3. o processo de determinar o ponto de divisão é recursivamente aplicado para cada parte obtida, até algum critério de parada ser alcançado (por exemplo, quando a mínima *expected information requirement* de todos os pontos de divisão é menor do que um limiar; ou quando o número de intervalos é maior do que um limiar).

Discretização baseada em entropia simplifica os dados e cria um conceito de hierarquia, usando a informação de classe. Ela ajuda a produzir resultados de classificação mais precisos, já que embute informação na representação dos dados RESUMIDOS.

Informação Mútua

Este também é um conceito vindo da teoria da informação, assim como o conceito de entropia. A informação mútua mede a independência mútua entre duas variáveis randômicas.

	B_1	B_2	$\dots$	B_S	Sums
A_1	n_{11}	n_{12}	$\dots$	n_{1S}	a_1
A_2	n_{21}	n_{22}	$\dots$	n_{2S}	a_2
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
A_R	n_{R1}	n_{R2}	$\dots$	n_{RS}	a_R
Sums	b_1	b_2	$\dots$	b_S	n

$$H(A) = -\sum_{i=1}^{R} \frac{a_i}{n} log \frac{a_i}{n}$$

é a entropia;

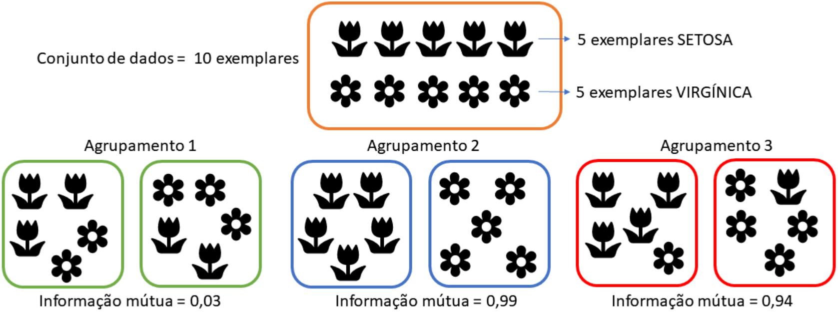
$$H(A, B) = -\sum_{i=1}^{R} \sum_{j=1}^{S} \frac{n_{i,j}}{n} log \frac{n_{i,j}}{n}$$

é a entropia conjunta e

$$IM(A, B) = H(A) + H(B) - H(A, B)$$

é a informação mútua.

Medindo a informação mútua de um agrupamento em relação a um *ground truth*:



n	H(A)	H(B)	nij/n	log*(nij/n)	prod	soma	H(A,B)	Informação Mútua IM = (H(A)+H(B)-H(A,B))
10	1,00	1,00	0,30	-1,74	-0,52	-1,97	1,97	0,03
			0,20	-2,32	-0,46			
			0,20	-2,32	-0,46			
			0,30	-1,74	-0,52			
10	1,00	1,00	0,50	-1,00	-0,50	-1,01	1,01	0,99
			0,99	-0,01	-0,01			
			1,01	0,01	0,01			
			0,50	-1,00	-0,50			
10	1,00	1,00	0,40	-1,32	-0,53	-1,06	1,06	0,94
			0,99	-0,01	-0,01			
			1,00	0,01	0,01			
			0,40	-1,32	-0,53			

Bibliografia

- Russel e Norvig: entropia
- Han et al.: entropia, normalização, distâncias
- Fausset: separabilidade linear
- Silva et al: entropia, normalização, distâncias
- Amelio e Pizzuti: informação mútua

Profa. Dra. Sarajane Marques Peres
Universidade de São Paulo
Escola de Artes, Ciências e Humanidades
Sala 320-A - Bloco I1
sarajane@usp.br www.each.usp.br/sarajane