Fundamentos de Mineração de Dados

Exercício sobre conceitos essenciais

1- Para os seguintes vetores, x e y, calcule as medidas de semelhança ou distância indicadas.

a) distância euclidiana:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

b) Jaccard:

J = número de correspondências “11” / número de atributos não zero
= $(f_{11}) / (f_{01} + f_{10} + f_{11})$

c) Similaridade do cosseno

$$\cos(x, y) = \frac{x.y}{\| x \| \| y \|} \qquad x.y = \sum_{k=1}^{n} x_k y_k \qquad \| x \| = \sqrt{\sum_{k=1}^{n} x_k^{\,2}}$$

d) Correlação de Pearson

$$r_{xy} = \frac{Cov(X,Y)}{S_x S_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

ou

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{n s_x s_y} = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}$$

a) x = (1, 1, 1, 1), y = (2, 2, 2, 2) cosine, correlação, Euclidean
b) x = (0, 1, 0, 1), y = (1, 0, 1, 0) cosine, correlação, Euclidean, Jaccard
c) x = (0, −1, 0, 1), y = (1, 0, −1, 0) cosine, correlação, Euclidean
d) x = (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1) cosine, correlação, Jaccard
e) x = (2, −1, 0, 2, 0, −3), y = (−1, 1, −1, 0, 0, −1) cosine, correlação

Respostas no excel e no notebook

2- Considere uma matriz esparça (document-term matrix), em que tf_{ij} é a frequência da *i-ésima* palavra (termo) no j-ésimo documento e *m* é o número de documentos. Considere a transformação da variável que é definida por

$$tf'_{ij} = tf_{ij} * \log \frac{m}{df_i},$$

onde df_i é o número de documentos em que o *i-ésimo* termo aparece e é conhecido como a **frequência do documento** do termo. Essa transformação é conhecida como transformação **inversa da frequência do documento**.

a) Qual é o efeito dessa transformação se um termo ocorrer em um documento? E em todos os documentos?

Se o termo aparece apenas em um documento, o seu peso será maior do que se o mesmo termo aparecer em todos os documentos, uma vez que a razão dentro da função logarítmica terá valores menores dado que df_i será igual a "*m*" resultando em log 1 = 0.

b) Qual pode ser o propósito dessa transformação?

Essa transformação tem como propósito aumentar a importância de termos que aparecem em poucos documentos e diminuir a importância de termos que aparecem em muitos, justamente pelo fato dos termos de baixa freqüência serem, em geral, mais discriminantes

3 - Aqui, exploramos ainda mais as medidas de cosseno e correlação.

a) Qual é a faixa de valores possíveis para a medida do cosseno?

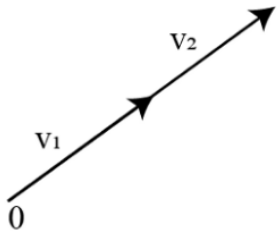
O valor desta métrica encontra-se entre -1 e 1, isto é, igual ao intervalo fechado [-1,1].

- ângulo de 0° - valor do cosseno = 1
- ângulo de 90° - ângulos ortogonais, valor do cosseno = 0
- ângulo de 180° - valor do cosseno = -1

b) Se dois objetos têm uma similaridade de cosseno de 1, eles são idênticos? Explique.

Não, indica apenas que o ângulo entre eles é zero, ou seja, os vetores se sobrepõem, porém, isso não significa que necessariamente eles são idênticos, mas sim que são similares.

Exemplo:



Os ângulo entre os vetores é zero, identificando similaridade entre eles, porém é possível observar que seu tamanhos são distintos, também conhecidos como vetores linearmente dependentes, sendo que $v_2 = a*v_1$, sendo a um número real.

4 - Discutir as diferenças entre a redução da dimensionalidade com base em agregação (ou seja, combinando uma ou mais *features*) e da redução da dimensionalidade com base em técnicas como PCA.

- A técnica de **Feature Selection** busca meios pelos quais seja possível avaliar a relevância ou a redundância de uma variável para a pergunta que se deseja responder e, a partir dessa avaliação, mantém ou elimina aquela variável do conjunto de dados.
- **PCA**: *Principal component Analysis* - Consiste em utilizar um algoritmo que aplica transformações lineares na matriz de features para combinar features reduzindo sua dimensionalidade. Com isso esta técnica invariavelmente gera perda de interpretabilidade, em favor de redução, e muitas vezes até mesmo melhora a performance e acurácia do modelo. É importante ressaltar que todas as colunas devem ser normalizadas antes de serem inseridas em um algoritmo de PCA.