SCC0276 – Aprendizado de Máquina

Vetores Binários

Profa. Dra. Roseli Ap. Francelin Romero

ICMC USP SÃO CARLOS

Similaridade entre vetores binários

- Algumas vezes, objetos p e q têm apenas valores binários
  - Ex.: 0110 e 1100
- Similaridades podem ser computadas usando:
  - M_{01} = número de atributos em que p = 0 e q = 1
  - M_{10} = número de atributos em que p = 1 e q = 0
  - M_{00} = número de atributos em que p = 0 e q = 0
  - M_{11} = número de atributos em que p = 1 e q = 1

Similaridade entre vetores binários

- Coeficiente de Casamento Simples

$$CCS = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$$

- Coeficiente Jaccard

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11})$$

- Agrupamento de dados

Exercício

- Que medida de similaridade binária gera o maior valor de similaridade entre vetores *p* e *q?*

p = 1 0 0 1 1 0 1 0 1 1 1 0
q = 0 1 0 0 1 1 0 0 1 0 1 1

Similaridade cosseno

- Muito usado quando dados são textos
  - *Bag of words*
    - Grande número de atributos
    - Vetores esparsos
- Sejam p e q vetores representando documentos
  - $\cos(p, q) = ||p||\ ||q||\ \cos\theta = (p \bullet q) / (||p||\ ||q||)$
    - $\bullet$: vector produto interno entre vetores
    - $||\ p\ ||$: é o tamanho (norma) do vetor p

Distância cosseno

- Distância angular entre dois vetores
  - Invariante a escala dos atributos
  - 1 – similaridade cosseno

$$dist_{\cos seno} = 1 - \frac{\sum_{k=1}^{m} p_k . q_k}{\sum_{k=1}^{m} p_k^2 . \sum_{k=1}^{m} q_k^2}$$

Distância de Pearson

- Muito usada em bioinformática e séries temporais
  - 1 – correlação entre dois vetores

$$dist_{Pearson} = 1 - \frac{\sum_{k=1}^{m} (p_k - \bar{p}).(q_k - \bar{q})}{\sqrt{\sum_{k=1}^{m} (p_k - \bar{p})^2 . \sum_{k=1}^{m} (q_k - \bar{q})^2}}$$

Propriedade de Distâncias

- Medidas de distância, em geral, têm as seguintes propriedades
  - Seja $d(p, q)$ a distância (dissimilaridade) entre dois objetos p e q
    - $d(p, q) \geq 0 \; \forall \; p$ e q e $d(p, q) = 0$ se somente se $p = q$ (definida positiva)
    - $d(p, q) = d(q, p) \; \forall \; p$ e q (simetria)
    - $d(p, r) \leq d(p, q) + d(q, r) \; \forall \; p, q$ e r (desigualdade triangular)
- Medidas que satisfazem essas propriedades são denominadas métricas

Propriedade de Distâncias

- Medidas de similaridade também têm propriedades bem definidas:
  - Seja *s(p, q)* a similaridade entre dois objetos *p* e *q*
    - *s(p, q) = 1* (similaridade máxima) apenas se *p = q*
    - *s(p, q) = s(q, p)* $\forall$ *p* e *q* (simetria)

Input space