Мере сличности

Ненад Митић

Математички факултет
nenad@matf.bg.ac.rs

Увод

- Како одредити сличност/различитост објеката, образаца, атрибута, догађаја, ...
- Подаци - различит тип, структура, расподела, димензионалност, ...
- Термин *близина* (енг. *proximity*) означава и сличност и различитост

Увод - наставак

- Функције сличности - већа вредност $\longrightarrow$ већа сличност
- Различитост - нумеричка мера колико су два објекта, атрибута, ... различити
- Сличност често [0,1], а различитост у $[0,+\infty)$
- *Растојање* - синоним за различитост
- Функција растојања - мања вредност $\longrightarrow$ већа сличност

Увод - наставак

Пример: функције сличности/различитости атрибута p и q

Тип атрибута	Сличност	Различитост
Номинални	$s = \begin{cases} 1 & \text{ako } p = q \\ 0 & \text{ako } p \neq q \end{cases}$	$d = \begin{cases} 1 & \text{ako } p \neq q \\ 0 & \text{ako } p = q \end{cases}$
Редни	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$	$d = \frac{\lvert p-q \rvert}{n-1}$
	Вредности се пресликавају у скуп $[0, n-1]$ где је n број вредности	
Интервални или размерни	$s = -d, s = \frac{1}{1+d},$ $s = 1 - \frac{d - min_d}{max_d - min_d}$	$d = \lvert p - q \rvert$

Мера и метрика

Функција растојања d је *метрика* ако важи

1. Позитивна одређеност
   - $d(p,q) \geq 0 \quad \forall \quad p,q$
   - $d(p,q) = 0$ акко $p = q$
2. Симетрија: $d(p,q) = d(q,p) \quad \forall \; p,q$
3. Неједнакост троугла:
   $d(p,r) \leq d(p,q) + d(q,r) \quad \forall \quad p,q,r$

Ултраметрика

Ако је функција растојања d *метрика* и ако важи

$$d(p,r) \leq max\{d(p,q), d(q,r)\} \quad \forall \quad p,q,r$$

тада је функција d *ултраметрика*

Примери мера

- које јесу метрика/ултраметрика?
- које нису метрика/ултраметрика?

Мере сличности за квантитативне податке

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

Растојање између две тачке у n димензионом простору $\overline{X} = (x_1, x_2, ..., x_n)$ і $\overline{Y} = (y_1, y_2, ..., y_n)$

- Хамингово растојање

$$Hamming(\overline{X}, \overline{Y}) = \sum_{i=1}^{n} q_i \quad \text{где је} \quad q_i = \begin{cases} 1, & \text{ako } x_i \neq y_i \\ 0, & \text{иначе} \end{cases}$$

- Најчешће коришћена мера је растојање Минковског или L_p мера

$$Dist(\overline{X}, \overline{Y}) = \left(\sum_{i=1}^{n} \mid x_i - y_i \mid^p \right)^{1/p}$$

Растојање Минковског

Специјални случајеви

- $p = 1$ - Градски блок (такси, Менхетн, ...)
- $p = 2$ - Еуклидско растојање
- $p \to \infty$ *супремум* растојање (L_{max}, L_∞ норма) = $max_{1 \leq i \leq n} |x_i - y_i|$
- Не мешати n (број димензија података) и p (величина параметра)

Растојање Минковског - недостаци

Није погодно за примену

- код ретких вишедимензионалних података са непознатом расподелом, шумовима, ...
- ако постоје локално ирелевантни атрибути (пример: анализа крви пацијената оболелих од различитих болести) због шума који се кумулира при израчунавању

Махаланобисово растојање

$$Maha(\overline{X}, \overline{Y}) = \sqrt{(\overline{X} - \overline{Y})\Sigma^{-1}(\overline{X} - \overline{Y})^T}$$

где је Σ^{-1} инверзна матрица матрице коваријанси података

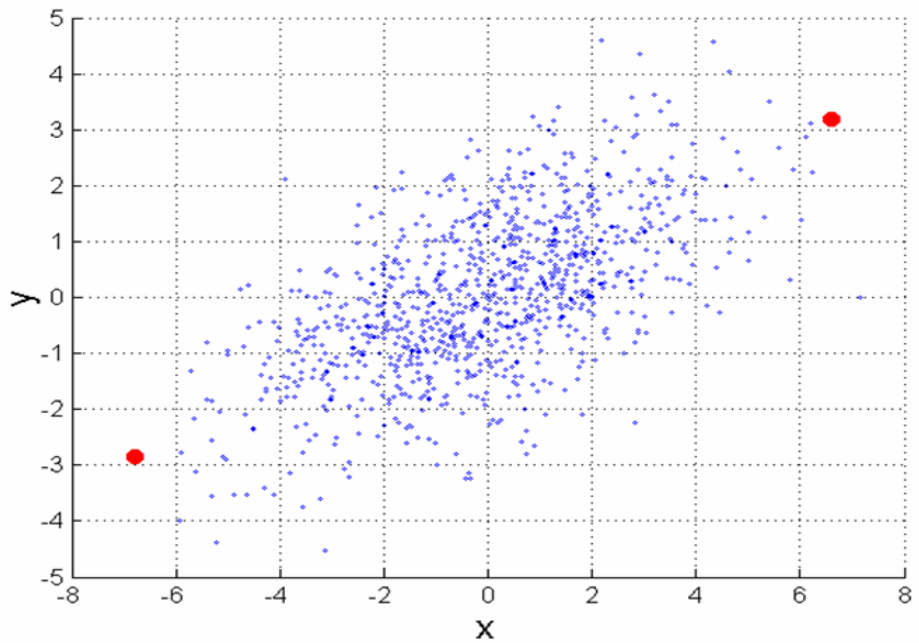
Махаланобисово растојање

Корисно је када важи

- атрибути су у корелацији
- атрибути имају различите опсеге вредности (различите варијансе)
- расподела података је приближно нормална (Гаусова)

Утицај расподеле на растојање

Међусобно растојање тачака $A(-6.8, -2.9)$ и $B(6.8, 3.1)$



Еуклидско растојање тачака је 14.7, а Махаланобисово 6

Растојање Минковског са тежинама

У случају да је потребно доделити тежинске факторе a_i хетерогеним атрибутим i

$$Dist(\overline{X}, \overline{Y}) = \left(\sum_{i=1}^{d} a_i \times \mid x_i - y_i \mid^p \right)^{1/p}$$

Мере сличности података са бинарним атрибутима

Сличност два слога $\overline{X} = (x_1, x_2, ..., x_d)$ и $\overline{Y} = (y_1, y_2, ..., y_d)$ са бинарним атрибутима се може дефинисати помоћу

- M_{01}=број атрибута који су једнаки 0 у $\overline{X}$ и 1 у $\overline{Y}$
- M_{10}=број атрибута који су једнаки 1 у $\overline{X}$ и 0 у $\overline{Y}$
- M_{00}=број атрибута који су једнаки 0 у $\overline{X}$ и 0 у $\overline{Y}$
- M_{11}=број атрибута који су једнаки 1 у $\overline{X}$ и 1 у $\overline{Y}$

Једноставно и Жакардово упаривање коефицијеана

- Једноставно упаривање коефицијената (енг. *Simple Matching Coefficient, SMC*)
  SMC=број упарених/број атрибута = $(M_{11} + M_{00})/(M_{01} + M_{10} + M_{11} + M_{00})$
- Жакардови коефицијенти - асиметрични атрибути
  J=број парова 11/број атрибута где нису обе вредности 0 = $M_{11}/(M_{01} + M_{10} + M_{11})$

Проширени Жакардови коефицијенти (коефицијенти Танимотоа)

- Варијанта Жакардових коефицијената применљива на атрибуте са непрекидним и пребројивим вредностима
- У случају атрибута са бинарним вредностима редукује се на Жакардове коефицијенте

$$T(\overline{X}, \overline{Y}) = \frac{\overline{X} \bullet \overline{Y}}{||\overline{X}||^2 + ||\overline{Y}||^2 - \overline{X} \bullet \overline{Y}}$$

Косинусна сличност

Нека су $\overline{X} = (x_1, x_2, ..., x_n)$ и $\overline{Y} = (y_1, y_2, ..., y_n)$ два вектора докумената. Њихова сличност може да се израчуна као

$$cos(\overline{X}, \overline{Y}) = \frac{\overline{X} \bullet \overline{Y}}{||\overline{X}|| \times ||\overline{Y}||}$$

odnosno

$$cos(\overline{X}, \overline{Y}) = \frac{\sum_{i=1}^{d} x_i \times y_i}{\sqrt{(\sum_{i=1}^{d} x_i^2)} \times \sqrt{(\sum_{i=1}^{d} y_i^2)}}$$

Косинусна сличност

Користи се код великог броја парова '00' при чему може да буде примењена и на не-бинарне векторе (касније-пример са документима)

Primer:

$d_1 = 3\ 2\ 0\ 5\ 0\ 0\ 0\ 2\ 0\ 0$

$d_2 = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 2$

$d_1 \cdot d_2 = 3*1+2*0+0*0+5*0+0*0+0*0+0*0+2*1+0*0+0*2 = 5$

$||d_1|| = (3*3+2*2+0*0+5*5+0*0+0*0+0*0+2*2+0*0+0*0)^{0.5} = (42)^{0.5} = 6.481$

$||d_2|| = (1*1+0*0+0*0+0*0+0*0+0*0+0*0+1*1+0*0+2*2)^{0.5} = (6)^{0.5} = 2.245$

$cos(d_1, d_2) = 0.34365$

Корелација

Корелација два објекта који имају бинарне или непрекидне атрибуте је мера линеарног односа између њихових атрибута

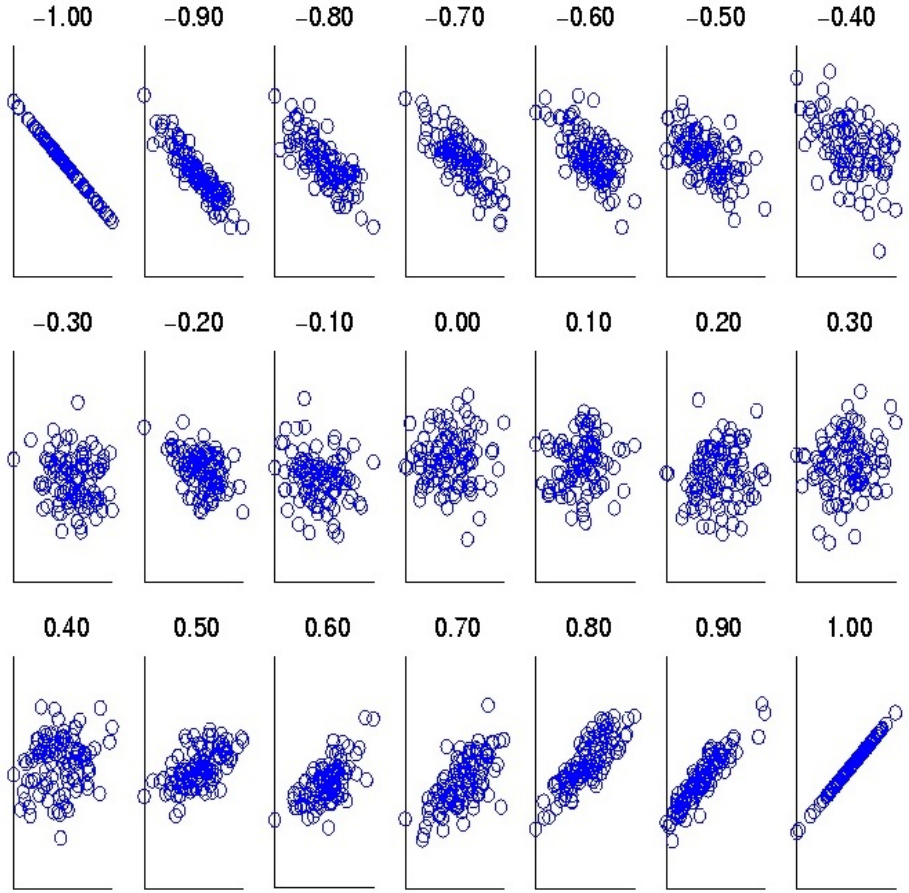
$$\text{коваријанса}(x, y) = cov_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})(y_k - \overline{y})$$

$$\text{стандардна девијација}(x) = \sigma_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \overline{x})^2}$$

$$\text{средња вредност}(x) = \overline{x} = \frac{1}{n} \sum_{k=1}^{n} x_k$$

$$\text{Пирсонов коефицијент корелације}(x, y) = \rho_{xy} = cov_{xy}/(\sigma_x * \sigma_y)$$

Корелација



Ако је корелација =1 (-1) $\rightarrow$ перфектно позитиван (негативан) линеарни однос $x_k = ay_k + b$

Мере сличности за категоричке податке

Сличност два податка $\overline{X} = (x_1, x_2, ..., x_n)$ i $\overline{Y} = (y_1, y_2, ..., y_n)$ са категоричким атрибутима се може дефинисати преко сличности појединачних атрибута

$$Sim(\overline{X}, \overline{Y}) = \sum_{i=1}^{n} S(x_i, y_i)$$

Мере сличности за категоричке податке

- Најједноставнији случај: $S(x_i, y_i) = \begin{cases} 1 & \text{ако } x_i = y_i \\ 0 & \text{иначе} \end{cases}$
- Не узима се у обзир релативна фреквенција атрибута
- Користи се *агрегирање статистичких особина*
- Мање фреквентне упарене вредности имају већу тежину

Сличност атрибута

Нека је $p_k(x)$ број слогова у којима k-ти атрибут узима вредност x

Мере које укључују учесталост (фреквенцију) су

- *Инверзна учесталост појављивања*

$$S(x_i, y_i) = \begin{cases} 1/p_k(x_i)^2, & ako\ x_i = y_i \\ 0, & \text{inače} \end{cases}$$

- *'Појављивање је добро'*. Већа сличност се постиже када је вредност мање учестала

$$S(x_i, y_i) = \begin{cases} 1 - p_k(x_i)^2, & ako\ x_i = y_i \\ 0, & \text{inače} \end{cases}$$

Сличност докумената

- Сличност два документа се боље оцењује ако се користе речи које су заједничке
- За нормализацију упаривања речи у случајевима када има речи које се ретко јављају и које се често јављају (везници, ...) користи се инверзна функција броја докумената n_i у коме се јавља реч i у укупном броју докумената n:

$$id_i = log(n/n_i)$$

- За смањење могућност да појава неке честе речи утиче на сличност докумената могу да се користе и функције

$$f(x_i) = sqrt(x_i)$$

$$f(x_i) = log(x_i)$$

Сличност докумената

- нормализована фреквенција за i-ту реч може да се дефинише као

$$h(x_i) = f(x_i) \cdot id_i$$

- Косинусно и Жакардово растојање докумената са нормализованом фреквенцијом речи су

$$cos(\overline{X}, \overline{Y}) = \frac{\sum_{i=1}^{d} h(x_i) \times h(y_i)}{\sqrt{\sum_{i=1}^{d} h(x_i)^2} \times \sqrt{\sum_{i=1}^{d} h(y_i)^2}}$$

$$J(\overline{X}, \overline{Y}) = \frac{\sum_{i=1}^{d} h(x_i) \times h(y_i)}{\sum_{i=1}^{d} h(x_i)^2 + \sum_{i=1}^{d} h(y_i)^2 - \sum_{i=1}^{d} h(x_i) \times h(y_i)}$$

Подаци са квантитативним и категоричким атрибутима

Сличност два слога $\overline{X} = (\overline{X}_n, \overline{X}_c)$ i $\overline{Y} = (\overline{Y}_n, \overline{Y}_c)$ са 'мешаним' (квантитативним и категоричким) атрибутима

$$Sim(\overline{X}, \overline{Y}) = \lambda \times NumSim(\overline{X}_n, \overline{Y}_n)) + (1 - \lambda) \times CatSim(\overline{X}_c, \overline{Y}_c))$$

где λ одређује релативну важност категоричких и нумеричких атрибута

Сличност дискретних података

Едит растојање, растојање за трансформације $\overline{X} = (x_1, x_2, ..., x_m)$ у $\overline{Y} = (y_1, y_2, ..., y_n)$.
За првих i симбола из $\overline{X}$ и првих j симбола $\overline{Y}$ цена трансформације је

$$Edit(i,j) = min \begin{cases} Edit(i-1,j) + \text{цена брисања} \\ Edit(i,j-1) + \text{цена уметања} \\ Edit(i-1,j-1) + I_{ij} \times \text{цена замене} \end{cases}$$

где је I_{ij} индикатор једнакости i-тог симбола $\overline{X}$ и j-тог симбола $\overline{Y}$

Пример: трансформација абабабабаб у бабабаба

Сличност дискретних података

Сличност на основу најдуже заједничке подниске

За првих i симбола из $\overline{X} = (x_1, x_2, ..., x_m)$ и првих j симбола из $\overline{Y} = (y_1, y_2, ..., y_n)$, у ознаци $\overline{X}_i$ и $\overline{Y}_i$ најдужа заједничка подниска (енг. *Longest Common SubSequence, LCSS*)

$$LCSS(i,j) = max \begin{cases} LCSS(i-1,j-1)+1 & ako \quad x_i = y_i \\ LCSS(i-1,j) & x_i \ nije \ upareno \\ LCSS(i,j-1) & y_j \ nije \ upareno \end{cases}$$

Већа вредност означава већу сличност; број подниски директно зависи од дужине ниски

Пример: одредити LCSS(агбфцгдђхе,аффбгцхдише)

Мере на основу информација

- Мере сличности засноване на теорији информација
- Ентропија
  - X - догађај са n могућих исхода $x_1, ..., x_n$
  - Вероватноћа исхода је $p_1, ..., p_n$
  - Ентропија догађаја X је

$$H(X) = -\sum_{i=1}^{n} p_i log_2 p_i$$

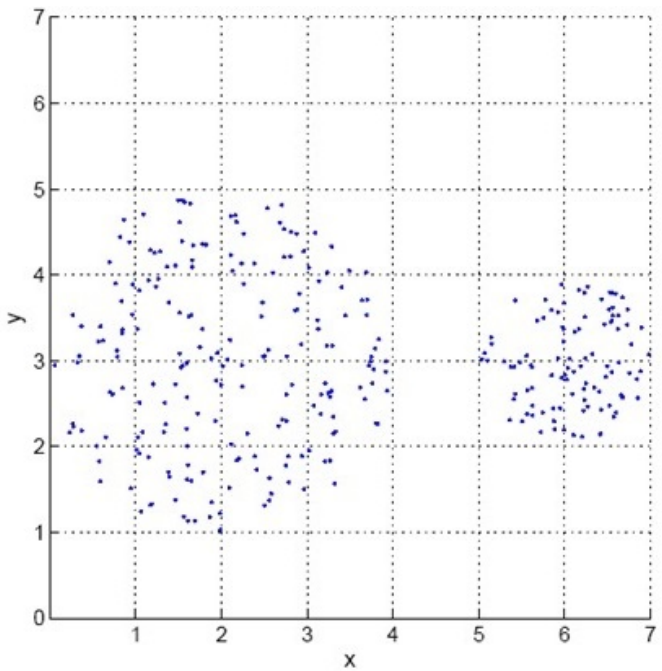
  - $H(X) \in [0, log_2 n]$ представља меру колико битова је потребно у просеку за представљање догађаја X

Мере на основу густина

- Мери се степен блискости објеката у некој области
- Концепт густине се користи у кластеровању и откривању аномалија
- Нечешће се користе
  - Еуклидска густина - број тачака по јединици површине/запремине
  - Густина вероватноће - процена дистрибуције података на основу изгледа
  - Граф засноване густине - на основу повезаности

Мере на основу густина

Пример: Еуклидска густина заснована на ћелијама - подела региона на неки број ћелија и дефинисање густине преко броја тачака у ћелијама



0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Мере на основу густина

Пример: Еуклидска густина заснована на центру - број ћелија на одређеној удаљености од централне тачке