• سوال اول

g_1, g_2

$c_1, c_2, \dots\dots\dots, c_{16}$

$g_1 = \{-5, -4, -3, -2, -1, 1, 2, 3, 4, 5\} \rightarrow \{-5, 5\} - \{0\}$

$g_2 = \{25, 16, 9, 4, 1, 1, 4, 9, 16, 25\} \rightarrow \{x^2 \mid x \in z\}$

a) آیا می‌توان این دو ژن را در یک خوشه قرار داد؟ (به کمک معیار شباهت همبستگی)

$$\text{correlation} = \frac{\text{covariance}(x, y)}{\text{std}(x) * \text{std}(y)} = \frac{S_{xy}}{S_x \cdot S_y}$$

$$\text{covariance} = S_{xy} = \frac{1}{n-1} \sum (x_k - \bar{x})(y_k - \bar{y})$$

$$S_x = \sqrt{\frac{1}{n-1} \sum (x_k - \bar{x})^2} = \sqrt{\frac{1}{10_5}\overbrace{(25+16+9+4+1)}^{\times 2}} = \sqrt{11}$$

$$\begin{cases} \text{mean}(x) = 0 \\ \text{mean}(y) = \frac{55 \times 2}{10} = 11 \end{cases}$$

$$S_y = \sqrt{\frac{1}{n-1} \sum (y_k - \bar{y})^2} = \sqrt{\frac{1}{10_5}(25^2 + 16^2 + 9^2 + 4^2 + 1)} = \sqrt{195{,}8}$$

$$\text{Correlation} = \frac{\sum (x_i - \bar{x})(y_i - \bar{x})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{(-5)(25-11) + (-4)(16-11) + \dots + (-1)(1-11)}{\sqrt{\Delta}}$$

$$\text{correlation} = \frac{0}{\Delta}$$

$$\boxed{\text{correlation} = 0} \longrightarrow$$

وقتی معیار correlation صفر شود، یعنی این که دو متغیر هیچ رابطه خطی ای با هم ندارند.

پس طبق این نتیجه گیری، مشابه هم نیستند و نمی‌توان آن دو را در یک خوشه قرار داد.

ط) مقدار شباهت (MI) را برای دو ژن محاسبه کنید :

$$MI = \sum \sum p(x,y) \log \frac{p(x,y)}{p(x) \cdot p(y)}$$

$\begin{cases} p(x) \\ p(y) \end{cases}$

$p(x,y) \rightarrow$ احتمال توام

$$p(g_1) = \{ \underbrace{\tfrac{1}{10}, \tfrac{1}{10}, \tfrac{1}{10}, \ldots, \tfrac{1}{10}}_{\text{۱۰ تا}} \}$$

$$p(g_2) = \{ \tfrac{2}{10}, \tfrac{2}{10}, \tfrac{2}{10}, \tfrac{2}{10} \}$$

$p(g_1, g_2)$

	-5	-4	-3	-2	-1	1	2	3	4	5
25	$\frac{1}{10}$	0	0	0	0	0	0	0	0	$\frac{1}{10}$
16	0	$\frac{1}{10}$	0	0	0	0	0	0	$\frac{1}{10}$	0
9	0	0	$\frac{1}{10}$	0	0	0	0	$\frac{1}{10}$	0	0
4	0	0	0	$\frac{1}{10}$	0	0	$\frac{1}{10}$	0	0	0
1	0	0	0	0	$\frac{1}{10}$	$\frac{1}{10}$	0	0	0	0

محاسبه‌ی MI :

$$= \left[p(-5,25) \cdot \log \frac{p(-5,25)}{p(-5)\, p(25)} \right] + p(-5,16) \cdot \log \frac{p(-5,16)}{p(-5) \cdot p(16)} +$$

$$p(-5,9) \cdot \log \frac{p(-5,9)}{p(-5)\, p(9)}] + p(-5,4) \cdot \log \frac{p(-5,4)}{p(-5) \cdot p(4)} +$$

$$p(-5,1) \cdot \log \frac{p(-5,1)}{p(-5)\, p(1)} + p(-4,25) \cdot \frac{\log(-4,25)}{p(-4)\, p(25)} +$$

$$p(-4,16) \cdot \log \frac{p(-4,16)}{p(-4)\, p(16)} + \ldots + p(-2,1) \cdot \log \frac{p(-2,1)}{p(-2)\, p(1)}$$

۴۰ ترم این عبارت صفر می‌شود ← به خاطر احتمال توام

Subject

Date

$$\rightarrow MI = 0{,}1 * \log \frac{0{,}1}{0{,}1 * 0{,}2} + 0{,}1 * \log\left(\frac{0{,}1}{0{,}1 * 0{,}2}\right) +$$

$$0{,}1 * \log \frac{(0{,}1)}{0{,}1 * 0{,}2} + 0{,}1 * \log \frac{0{,}1}{0{,}1 * 0{,}2} + 0{,}1 * \log \frac{0{,}1}{0{,}1 * 0{,}2}$$

$$+ 0{,}1 * \log \frac{0{,}1}{0{,}1 * 0{,}2} + 0{,}1 * \log \frac{0{,}1}{0{,}1 * 0{,}2} + 0{,}1 * \log \frac{0{,}1}{0{,}1 * 0{,}2}$$

$$+ 0{,}1 * \log \frac{0{,}1}{0{,}1 * 0{,}2} + 0{,}1 * \log \frac{0{,}1}{0{,}1 * 0{,}2}$$

$$MI = 10 * \left(0{,}1 * \log \frac{0{,}1}{0{,}1 * 0{,}2}\right) = \log \frac{0{,}1}{0{,}1 * 0{,}2}$$

$$MI = \log \frac{0{,}1}{0{,}02}$$

$$MI = \log 5$$

$$MI = 0{,}698$$

Normalization $\longrightarrow$ $$\frac{MI}{\sum p(x) \log \frac{p(x)}{N} \sum p(y) \log \frac{p(y)}{N}}$$

$$\rightarrow \frac{0{,}698}{\left(0{,}1 * \log \frac{0{,}1}{10} + \text{---} + 0{,}1 * \log \frac{0{,}1}{10}\right) * \left(0{,}2 * \log \frac{0{,}2}{10} + \text{---} + 0{,}2 * \log \frac{0{,}2}{10}\right)}$$

$$\rightarrow \frac{MI = 0{,}698}{10 * 0{,}1 * (-2) * 10 * 0{,}2 * -1{,}7} = \frac{MI}{-2 * 2 * (-1{,}7)} = \frac{0{,}698}{6{,}8} \simeq 0{,}1$$

Subject
Date

چون مقدارهای MI نزدیک به صفر شده است ← پس این دو بردار شباهت خیلی کمی دارند.

$0 < MI < 1$

$\downarrow$ شباهت خیلی کم

$\downarrow$ شباهت خیلی زیاد

c) چه تفاوتی در نتیجه مشاهده می‌کنید؟ چرا؟

correlation این دو ژن صفر شد (عدم شباهت)
اما در معیار MI، مقدار کمی شباهت برای آن‌ها در نظر گرفته شده است.

Subject
Date

• سؤال دوم

$x = (1, 1, 1, 1)$
$y = (2, 2, 2, 2)$

a) cosine
correlation
euclidean

A) $\cos(x, y) = \frac{x \cdot y}{|x| \cdot |y|} = \frac{2\times1 + 2\times1 + 2\times1 + 2\times1}{|\sqrt{1^2+1^2+1^2+1^2}| \cdot |\sqrt{2^2+2^2+2^2+2^2}|}$

$\cos(x, y) = \frac{8}{\sqrt{4} \cdot \sqrt{16}} = \frac{8}{2\times4} = 1 \longrightarrow \boxed{\text{cosine}(x, y) = 1}$

B) $\text{correlation} = \frac{\text{covariance}(x, y)}{\text{std}(x) * \text{std}(y)} = \frac{S_{xy}}{S_x \cdot S_y} = \frac{0}{0\times0} =$ تعریف نشده؟ $= \cdot/\cdot$

$\boxed{\text{correlation} = \text{unknown}}$ (undefined)

$\text{correlation} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$

$\text{mean}(x) = \frac{4}{4} = 1$
$\text{mean}(y) = \frac{8}{4} = 2$

$\text{correlation} = \frac{[(2-2)\cdot(1-1)] \cdot [(2-2)\cdot(1-1)] \cdot [(2-2)\cdot(1-1)] \cdot [(2-2)\cdot(1-1)]}{\sqrt{\quad 0 \quad}}$

c) euclidean

$d = \sqrt{(2-1)^2 + (2-1)^2 + (2-1)^2 + (2-1)^2} = \sqrt{4} = 2$

$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

$\boxed{\text{euclidean} = 2}$

PAPCO

Subject
Date

$x = (0, 1, 0, 1)$

$y = (1, 0, 1, 0)$

b)
- cosine
- corelation
- euclidean
- jaccard

A) cosine

$$\text{cosine}(x,y) = \frac{x \cdot y}{|x| \cdot |y|} = \frac{\overbrace{(1\times0)+(0\times1)+(1\times0)+(0\times1)}^{0}}{\sqrt{1^2+1^2} \cdot \sqrt{1^2+1^2}}$$

$$\boxed{\text{cosine}(x,y) = 0}$$

B) correlation

$$= \frac{\text{covariance}(x,y)}{\text{std}(x) * \text{std}(y)} = \frac{s_{xy}}{s_x \cdot s_y} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2}} = \frac{4\times\frac{25}{100}}{4\times\frac{25}{100}}$$

$$\boxed{\text{correlation} = 1}$$

$\text{mean}(x) = \frac{1}{4}(2) = \frac{1}{2}$

$\text{mean}(y) = \frac{1}{4}(2) = \frac{1}{2}$

$$\text{covariance}(x,y) = s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y}) = \frac{1}{3}\times\left[(0)(1-\tfrac{1}{2}) + (1)(0-\tfrac{1}{2}) + 0(1-\tfrac{1}{2}) + (1)(1-\tfrac{1}{2})\right]$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2} = \sqrt{\frac{1}{3}\Big[\underbrace{(0-0.5)^2}_{\frac{1}{4}} + \underbrace{(1-0.5)^2}_{\frac{1}{4}} + \underbrace{(0-0.5)^2}_{\frac{1}{4}} + \underbrace{(1-0.5)^2}_{\frac{1}{4}}\Big]} = \sqrt{\frac{1}{3}}$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k-\bar{y})^2} = \sqrt{\frac{1}{3}\Big[\underbrace{(1-0.5)^2}_{\frac{1}{4}} + \underbrace{(0-0.5)^2}_{\frac{1}{4}} + \underbrace{(1-0.5)^2}_{\frac{1}{4}} + \underbrace{(0-0.5)^2}_{\frac{1}{4}}} = \sqrt{\frac{1}{3}}$$

c) Euclidean :

$$d = \sqrt{(1-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2} = \sqrt{4} = 2$$

$$\boxed{\text{Euclidean} = 2}$$

d) Jaccard

$$= \frac{\text{number of 11 matches}}{\text{number of non zero attributes}} = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} = \frac{0}{2+2+0} = 0$$

$$\boxed{\text{Jaccard} = 0}$$

Subject
Date

$x=(1,1,0,1,0,1)$
$y=(1,1,1,0,0,1)$

- manhattan
- correlation
- bhattacharyya

(C

A) manhattan : $M=|1-1|+|1-1|+|1-0|+|0-1|+(0-0)+(1-1)=2$

$\boxed{\text{manhattan}=2}$

B) correlation $=\dfrac{\text{covariance}(x,y)}{\text{std}(x)*\text{std}(y)}=\dfrac{s_{xy}}{s_x\cdot s_y}=$

$\text{mean}(x)=\dfrac{4}{6}=\dfrac{2}{3}$ $\quad \text{std}(x)=\sqrt{\dfrac{1}{5}\left[(\frac{1}{3})^2+(\frac{1}{3})^2+(\frac{2}{3})^2+(\frac{1}{3})^2+(\frac{2}{3})^2+(\frac{1}{3})^2\right]}$

$\text{mean}(y)=\dfrac{4}{6}=\dfrac{2}{3}$ $\quad \text{std}(y)=\text{std}(x)=\sqrt{\dfrac{1}{5}\left[4\times\frac{1}{9}+2\times\frac{4}{9}\right]}=\sqrt{\frac{1}{5}\times\frac{12}{9}}=\sqrt{\frac{4}{15}}$

$\text{covariance}(x,y)=S_{xy}=\dfrac{1}{n-1}\sum(x_k-\bar{x})(y_k-\bar{y})=\dfrac{1}{5}\times\left[(1)(\frac{1}{3})+(1)(\frac{1}{3})+1(\frac{-2}{3})+1(\frac{1}{3})\right]$
$=\dfrac{1}{15}$

$\text{correlation}=\dfrac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\dfrac{(1-\frac{2}{3})^2+(1-\frac{2}{3})^2+(\frac{2}{3})^1(\frac{1}{3})^1+(\frac{1}{3})^2+(\frac{2}{3})^2+(\frac{1}{3})^2}{(\frac{1}{3})^2+(\frac{1}{3})^2+(\frac{1}{3})^2+(\frac{2}{3})^2+(\frac{2}{3})^2+(\frac{1}{3})^2}$

$\rightarrow \boxed{\text{correlation}=\frac{1}{4}}$

C) $\text{Bhattacharyya}=-\ln\left(\sum_{i=1}^{n}\sqrt{x_iy_i}\right)$

$=-\ln(\sqrt{1\times1}+\sqrt{1\times1}+\sqrt{1\times0}+\sqrt{0\times1}+\sqrt{0\times0}+\sqrt{1\times1})$

$=-\ln(1+1+1)$

$=-\ln(3)$

$=-0.47 \longrightarrow \boxed{\text{Bhattacharyya}=-0.47}$

هر متغیر کیفی ترتیبی، اسمی هم هست.
هر متغیر کمّی نسبی، بازه‌ای هم هست. ← نکته‌ی مهم

Subject
Date

سوال سوم

مهم *

تعریف مجموعه
. binary
. discrete
. continuous

. categorical کیفی → Nominal اسمی / Ordinal ترتیبی
. Numerical کمّی → Interval بازه / Ration نسبت → True zero point

a) روشنایی اندازه‌گیری شده توسط نورسنج

نوع متغیر: پیوسته، کمّی، بازه‌ای / نسبت

پیوسته است چون نورسنج مقادیر صحیح ثبت می‌دهد.
مقادیر عددی اند ← کمّی است
برای مقادیر، جمع و تفریق و نسبت معنا دارد ← نسبی است

← اگر روشنایی a باشد و روشنایی کمی دیگر 2a باشد نسبت معنا دارد.

b) روشنایی با قضاوت ناظر انسانی

نوع متغیر: گسسته، کیفی، ترتیبی / اسمی

چون ناظرهای انسانی، کیفیت روشنایی را به چند دسته تقسیم می‌کنند که توصیف کیفی می‌دهد.
و چون ranking برای آنها معنا دارد، متغیر کیفی ترتیبی است.
(با این فرض که ناظر در حال قضاوت لفظی است نه عددی، اگر قضاوت عددی کند مثل حالت قبل است)

c) زاویه اندازه‌گیری شده بین $0°$ و $360°$

نوع متغیر: پیوسته، کمّی، بازه‌ای / نسبی

چون یک fixed size بین مقادیر مختلف خود دارد، بازه‌ای است.
و چون نسبت برای آن معنا دارد، نسبی هم می‌تواند باشد.

$\frac{60}{30} = 2$ ←

P4PCO

Subject
Date

d) ارتفاع از سطح اقیانوس (متر، کیلومتر)   نوع متغیر: پیوسته، کمی، بازه‌ای ← نوع و تفریق فقط معنا دارد.   fixed size
نسبی ← نسبت معنا دارد

e) درجه نظامی   نوع متغیر: گسسته، کیفی، ترتیبی
اسمی

چون یک سری کلاس با ranking است، پس کیفی گسسته ترتیبی است.

درجه‌های نظامی یک سری سلسله مراتب از پیش تعیین شده است ← گسسته ی کیفی
قابل مقایسه‌اند ← ترتیبی

f) مدال‌های طلا و نقره و برنز در المپیک   نوع متغیر: گسسته، کیفی، ترتیبی
اسمی

چون یک سری کلاس با ranking است، کیفی گسسته ترتیبی است.

نکته: هر متغیر کمی نسبی، بازه‌ای نیز هست.
هر متغیر کیفی ترتیبی، اسمی نیز هست.