(b

معیار شباهت MI را برای دو ژن محاسبه کنید و به این سوال پاسخ دهید.

رابطه MI به صورت زیر است

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

که در این رابطه به $p(x)$ و $p(y)$ احتمالات حاشیه‌ای و به $p(x,y)$ احتمال توأم میگویند. برای این بخش ابتدا نیاز است که احتمالات حاشیه‌ای و احتمالات توأم را محاسبه کنیم.

ابتدا احتمال حاشیه‌ای را برای g1 محاسبه می‌کنیم که برابر است با تعداد رخداد هر مشاهده در مجموعه نمونه‌ها

c	-5	-4	-3	-2	-1	1	2	3	4	5
P(c)	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10	1/10

سپس احتمال حاشیه‌ای را برای g2 محاسبه می‌کنیم برابر با تعداد رخداد هر مشاهده در نمونه‌ها

c	25	16	9	4	1
P(c)	2/10	2/10	2/10	2/10	2/10

سپس احتمال توأم را محاسبه می‌کنیم

P(g1,g2)	-5	-4	-3	-2	-1	1	2	3	4	5
25	1/10	0	0	0	0	0	0	0	0	1/10
16	0	1/10	0	0	0	0	0	0	1/10	0
9	0	0	1/10	0	0	0	0	1/10	0	0
4	0	0	0	1/10	0	0	1/10	0	0	0
1	0	0	0	0	1/10	1/10	0	0	0	0

حال محاسبه مقدار MI

$$MI = \sum\sum p(x,y)\log(\frac{p(x,y)}{p(x)p(y)})$$
$$= [p(-5,25).\log(\frac{p(-5,25)}{p(-5)p(25)})] + [p(-5,16).\log(\frac{p(-5,16)}{p(-5)p(16)})] +$$
$$[p(-5,9).\log(\frac{p(-5,9)}{p(-5)p(9)})] + [p(-5,4).\log(\frac{p(-5,4)}{p(-5)p(4)})] +$$
$$[p(-5,1).\log(\frac{p(-5,1)}{p(-5)p(1)})] + [p(-4,25).\log(\frac{p(-4,25)}{p(-4)p(25)})] +$$
$$[p(-4,16).\log(\frac{p(-4,16)}{p(-4)p(16)})] + [p(-4,9).\log(\frac{p(-4,9)}{p(-4)p(9)})] +$$
$$[p(-4,4).\log(\frac{p(-4,4)}{p(-4)p(4)})] + [p(-4,9).\log(\frac{p(-4,1)}{p(-4)p(1)})] +$$
$$[p(-3,25).\log(\frac{p(-3,25)}{p(-3)p(25)})] + [p(-3,16).\log(\frac{p(-3,16)}{p(-3)p(16)})] +$$
$$[p(-3,9).\log(\frac{p(-3,9)}{p(-3)p(9)})] + [p(-3,4).\log(\frac{p(-3,4)}{p(-3)p(4)})]$$
$$+ [p(-3,1).\log(\frac{p(-3,1)}{p(-3)p(1)})] + \ldots +$$
$$[p(5,25).\log(\frac{p(5,25)}{p(5)p(25)})] + [p(5,16).\log(\frac{p(5,16)}{p(5)p(16)})]$$
$$+ [p(5,9).\log(\frac{p(5,9)}{p(5)p(9)})] + [p(5,4).\log(\frac{p(5,4)}{p(5)p(4)})]$$
$$+ [p(-2,1).\log(\frac{p(-2,1)}{p(-2)p(1)})]$$

رابطه بالا دارای ۵۰ ترم جمع است که از این ۵۰ جمع، ۴۰ مورد آن صفر می‌شود (با توجه به جدول احتمال توأم) و تنها ده مورد از آن‌ها می‌توانند مقدار غیرصفر داشته باشند که به شرح زیر است.

$$MI = \sum\sum p(x,y)\log(\frac{p(x,y)}{p(x)p(y)})$$
$$= [p(-5,25).\log(\frac{p(-5,25)}{p(-5)p(25)})] + [p(5,25).\log(\frac{p(5,25)}{p(5)p(25)})] +$$
$$[p(-4,16).\log(\frac{p(-4,16)}{p(-4)p(16)})] + [p(4,16).\log(\frac{p(4,16)}{p(4)p(16)})] +$$
$$[p(-3,9).\log(\frac{p(-3,9)}{p(-3)p(9)})] + [p(3,9).\log(\frac{p(3,9)}{p(3)p(9)})] +$$
$$[p(-2,4).\log(\frac{p(-2,4)}{p(-2)p(4)})] + [p(2,4).\log(\frac{p(2,4)}{p(2)p(4)})] +$$
$$[p(-1,1).\log(\frac{p(-1,1)}{p(-1)p(1)})] + [p(1,1).\log(\frac{p(1,1)}{p(1)p(1)})]$$

سپس مقادیر را از جداول بالا در رابطه پایانی جایگذاری می‌کنیم

$$MI = [0.1*\log(\frac{0.1}{0.1*0.2})]+[0.1*\log(\frac{0.1}{0.1*0.2})]+$$
$$[0.1*\log(\frac{0.1}{0.1*0.2})]+[0.1*\log(\frac{0.1}{0.1*0.2})]+$$
$$[0.1*\log(\frac{0.1}{0.1*0.2})]+[0.1*\log(\frac{0.1}{0.1*0.2})]+$$
$$[0.1*\log(\frac{0.1}{0.1*0.2})]+[0.1*\log(\frac{0.1}{0.1*0.2})]+$$
$$[0.1*\log(\frac{0.1}{0.1*0.2})]+[0.1*\log(\frac{0.1}{0.1*0.2})]$$

سپس ساده‌سازی می‌کنیم

$$MI = 10*[0.1*\log(\frac{0.1}{0.1*0.2})] = \log(\frac{0.1}{0.02}) = \log(5) = 0.698$$

برای اینکه مشخص کنیم شباهت این دو بردار تا چه حد است، باید مقدار MI را نرمال‌سازی کنیم

$$NMI = \frac{MI}{\sum p(x)\log(\frac{p(x)}{N})\sum p(y)\log(\frac{p(y)}{N})}$$

پس نرمال‌سازی را انجام می‌دهیم

$$NMI = \frac{MI}{\sum p(x)\log(\frac{p(x)}{N})\sum p(y)\log(\frac{p(y)}{N})}$$
$$= \frac{MI}{[0.1*\log(\frac{0.1}{10})+...+0.1*\log(\frac{0.1}{10})]*[0.2*\log(\frac{0.2}{10})+...+0.2*\log(\frac{0.2}{10})]}$$
$$= \frac{MI}{[10*0.1*-2][10*0.2*-1.7]} = \frac{MI}{[-2][2*-1.7]} = \frac{MI}{6.8} = \frac{0.698}{6.8} \sim 0.1$$

مقدار به دست آمده حدود ۰٫۱ است. طبق شاخص MI زمانی که مقدار این شاخص نزدیک به یک باشد یعنی دو نمونه دارای شباهت بیشتری هستند تا زمانی که مقدار این شاخص نزدیک به صفر باشد. پس با توجه به مقدار ۰٫۱ می‌توان گفت که این دو بردار شباهت خوبی با هم ندارند و به نظر نمی‌رسد که در یک خوشه قرار بگیرند.

c)

آیا تفاوتی درنتیجه مشاهده می‌کنید؟ اگر بله چرا؟

بله. در معیار همبستگی به طور قطع گفته شده است که این دو نمونه کاملا از هم متمایز هستند ولی در شاخص MI ، تا حدودی شباهت برای آن‌ها در نظر گرفته شده است.

دلیل : در همبستگی ضرب وزن‌دار دو متغیر تصادفی مطرح است اما در شاخص MI، ضرب وزن‌دار احتمال توأم دو متغیر تصادفی را در نظر می‌گیریم. پس در شاخص MI به شباهت توزیع نگاه می‌کند ولی در همبستگی به ارتباطات بین دو متغیر تصادفی. از طرفی در شاخص MI به ارتباط خطی بین دو متغیر تصادفی توجهی ندارد در صورتی که در ضرایب همبستگی این ارتباط خطی را می‌توانیم متوجه شویم. پس این احتمال وجود دارد که با توجه به این همبستگی بین دو متغیر خطی نباشد و در نتیجه همبستگی صفر را ارائه دهد ، اما شاخص MI این دو متغیر تصادفی را همبسته اعلام می‌کند.

سوال دوم

برای وکتورهای داده شده، موارد مطلوب را به دست آورید:

(a

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

Cosine, Correlation , Euclidean

$$\text{Cosine} = \frac{\sum_{i=1}^{n} x_i . y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} = \frac{(2*1)+(2*1)+(2*1)+(2*1)}{\sqrt{(2^2+2^2+2^2+2^2)}\sqrt{(1^2+1^2+1^2+1^2)}} = \frac{8}{\sqrt{16}\sqrt{4}} = \frac{8}{4*2} = 1$$

$$Euclidean = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(2-1)^2+(2-1)^2+(2-1)^2+(2-1)^2} = \sqrt{4} = 2$$

$$Correlation = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{[(2-2)(1-1)].[(2-2)(1-1)].[(2-2)(1-1)].[(2-2)(1-1)]}{\sqrt{[(2-2)^2+(2-2)^2+(2-2)^2+(2-2)^2].[(1-1)^2+(1-1)^2+(1-1)^2+(1-1)^2]}}$$

$$= \frac{0}{0} = nan$$

(b

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

Cosine, Correlation , Euclidean, Jaccard

$$\text{Cosine} = \frac{\sum_{i=1}^{n} x_i . y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \frac{(1*0)+(0*1)+(1*0)+(0*1)}{\sqrt{(1^2+0^2+1^2+0^2)}\sqrt{(0^2+1^2+0^2+1^2)}} = \frac{0}{\sqrt{2}\sqrt{2}} = 0$$

$$Euclidean = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(1-0)^2+(0-1)^2+(1-0)^2+(0-1)^2} = \sqrt{4} = 2$$

$$Correlation = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{[(1-0.5)(0-0.5)].[(0-0.5)(1-0.5)].[(1-0.5)(0-0.5)].[(1-0.5)(0-0.5)]}{\sqrt{[(1-0.5)^2+(0-0.5)^2+(1-0.5)^2+(0-0.5)^2].[(0-0.5)^2+(1-0.5)^2+(0-0.5)^2+(1-0.5)^2]}}$$

$$= \frac{[-0.25].[-0.25].[-0.25].[-0.25]}{\sqrt{[0.25+0.25+0.25+0.25].[0.25+0.25+0.25+0.25]}}$$

$$= \frac{[4*0.25]}{\sqrt{[4*0.25].[4*0.25]}} = \frac{[4*0.25]}{[4*0.25]} = 1$$

$$Jaccard = \frac{\sum_{i=1}^{n}\min(x_i, y_i)}{\sum_{i=1}^{n}\max(x_i, y_i)} = \frac{\min(1,0)+\min(0,1)+\min(1,0)+\min(0,1)}{\max(1,0)+\max(0,1)+\max(1,0)+\max(0,1)} = \frac{0}{4} = 0$$

(c

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

Manhattan, Correlation , Bhattacharyya

$$\text{Manhattan} = \sum_{i=1}^{n}|x_i - y_i|$$

$$= (|1-1|+|1-1|+|1-0|+|0-1|+|0-0|+|1-1|) = (0+0+1+1+0+1) = 3$$

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

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

$$= -\ln(1+1+0+0+0+1) = -\ln(3) = -0.47$$

$$Correlation = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

$$= \frac{[(1-(4/6))(1-(4/6))].[(1-(4/6))(1-(4/6))].[(1-(4/6))(0-(4/6))].[(0-(4/6))(1-(4/6))].[(0-(4/6))(0-(4/6))].[(1-(4/6))(1-(4/6))]}{\sqrt{[(1-(4/6))^2+(1-(4/6))^2+(1-(4/6))^2+(0-(4/6))^2+(0-(4/6))^2+(1-(4/6))^2].[(1-(4/6))^2+(1-(4/6))^2+(0-(4/6))^2+(1-(4/6))^2+(0-(4/6))^2+(1-(4/6))^2]}}$$

$$= \frac{(2/6)^2.(2/6)^2.(2/6*4/6).(2/6*4/6).(2/6)^2.(2/6)^2}{\sqrt{[4*(2/6)^2+2*(4/6)^2].[4*(2/6)^2+2*(4/6)^2]}}$$

$$= \frac{(2)^2.(2)^2.(2*4).(2*4).(2)^2.(2)^2}{\sqrt{[4*(2)^2+2*(4)^2].[4*(2)^2+2*(4)^2]}} = \frac{2^{14}}{[4*(2)^2+2*(4)^2]} = \frac{2^{14}}{3*2^4} = \frac{2^{10}}{3}$$

سوال سوم

ویژگی‌های زیر را به صورت دوتایی (Binary)، گسسته (Discrete) یا پیوسته (Continuous) تعریف کنید. همچنین آن‌ها را به عنوان کیفی (اسمی یا ترتیبی) یا کمی (بازه یا نسبت) دسته‌بندی کنید. نکته: بعضی موارد ممکن است بیشتر از یک حالت داشته باشند. بنابراین، در صورتی که فکر می‌کنید چند حالت وجود دارد، دلیل انتخابتان را ذکر کنید.

(a

روشنایی اندازه‌گیری شده توسط نورسنج

جواب: .ویژگی از نوع پیوسته است چون نورسنج مقادیر عدد صحیح مثبت می‌دهد و چون مقادیر به صورت عددی هستند، در نتیجه ویژگی کمی محسوب می‌شود. از طرفی چون مقادیر از یک مقدار صفر مطلق شروع می‌شوند، در نتیجه از نوع کمی نسبی خواهد بود.

(b

روشنایی از روی قضاوت ناظر انسانی

جواب: این سوال را می‌توان از چند منظر نگاه کرد. اگر ناظر بخواهد مثل نورسنج مقادیر عددی بدهد، پس ویژگی می‌تواند از نوع پیوسته، کمی و نسبی باشد.

اگر ناظر بخواهد روشنایی را به صورت لفظی بیان کند مثلا بگوید نور کم، زیاد ، خیلی زیاد و غیره است ، در نتیجه ویژگی از نوع گسسته است و به صورت کیفی بیان شده است و چون این لفظها به صورت ترتیبی قابل قیاس هستند، در نتیجه از نوع ترتیبی است.

(c

زاویه اندازه‌گیری شده بین ۰ و ۳۶۰ درجه

جواب : ویژگی از نوع پیوسته به دلیل اینکه مقادیر می‌توانند عدد صحیح بین ۰ تا ۳۶۰ باشند. چون مقادیر عددی هستند پس از نوع کمی محسوب شده و به دلیل اینکه مقدار زاویه از یک صفر مطلق شروع می‌شود، پس از نوع نسبی خواهد بود.

(d

ارتفاع از سطح اقیانوس

جواب: پیوسته، کمی، نسبی. دلایل مشابه با بخش c می‌باشد

(e

درجه نظامی

جواب: ویژگی از نوع گسسته است چون بین دو درجه نظامی، مقدار بینابینی نمی‌توان یافت. از طرفی چون درجه نظامی مقادیر عددی نیستند پس از نوع کیفی محسوب می‌شوند و همچنین چون درجه‌های نظامی قابل مقایسه هستند و می‌توان در یک ترتیب اولویت‌بندی کرد پس از نوع ترتیبی هستند.

(f

مدال‌های طلا و نقره و برنز در بازی المپیک

جواب: گسسته، کیفی، ترتیبی. دلایل مشابه با قسمت e می‌باشد