تمرین اول داده کاوی - سوالات تشریحی

محمدامین محمدی - ۹۴۳۱۰۲۰

سوال ۱

a)

$\bar{g_1} = 0\,,\ \bar{g_2} = 11$
$S_{g_1 g_2} = \frac{1}{9}\sum_{k=1}^{10}(c_{k_1} - \bar{g_1}) * (c_{k_2} - \bar{g_2}) = 0$
$corr(g_1, g_2) = \frac{S_{g_1 g_2}}{S_x * S_y} = 0$
این دو ژن همبستگی خطی ندارند. در نتیجه، با توجه به این معیار، در یک خوشه قرار نمی‌گیرند.

b)

$I(g_1, g_2) = H(g_1) + H(g_2) - H(g_1, g_2)$
$H(g_1) = -\sum_{k=1}^{10} 1/10 * log_2(1/10) = 3.32192809$
$H(g_2) = -\sum_{k=1}^{10} 2/10 * log_2(2/10) = 4.64385619$
$H(g_1, g_2) = -\sum_{i=1}^{10}\sum_{j=1}^{10} 2/100 * log_2(2/100) = 11.2877124$
$I(g1, g2) = -3.32192812$
با توجه به این معیار، چون MI این دو ژن صفر نیست، بین آن‌ها همبستگی وجود دارد و در نتیجه از این جهت می‌توان آن‌ها را در یک خوشه قرار داد.

c)

بله، در حالت اول همبستگی خطی بین این دو ژن صفر می‌شود و در نتیجه، با توجه به این معیار نمی‌توان این دو ژن را در یک خوشه قرار داد، زیرا این دو معیار همبستگی خطی ندارند. اما در حالت دوم به دلیل وجود همبستگی غیرخطی بین این دو ژن، مقدار MI صفر نمی‌شود، در واقع مقدار فعالیت ژن g2 تابعی (x^2) از مقدار فعالیت ژن g1 است و با هم همبستگی دارند.

سوال ۲

a)

$cosine(x,y) = \frac{8}{4*2} = 1$
$corr(x,y) = 0$
$euc(x,y) = 2$

b)

$cosine(x,y) = 0$
$corr(x,y) = -1$
$euc(x,y) = 2$
$jacc(x,y) =$

c)

$manhattan(x,y) = 2$
$corr(x,y)$:

$$S_x = \sqrt{\frac{12}{45}} = 0.51$$
$$S_y = \sqrt{\frac{12}{45}} = 0.51$$
$$S_{xy} = \frac{1}{3}$$

$\Rightarrow corr(x,y) = 0.25$
$bhattacharyya(x,y) = -ln(3) = -1.09$

سوال ۳

آ) پیوسته، کمی، نسبت (در حالت خاموش← ۰)
ب) گسسته، کیفی، ترتیبی (با استناد به حالت‌های: کم‌نور، پرنور)
ج) پیوسته، کمی، نسبت (عدد را حقیقی در نظر می‌گیرم)
چ) پیوسته، کمی، بازه (می‌تواند عدد منفی هم باشد)
ه) گسسته، کیفی، ترتیبی
خ) گسسته، کیفی، ترکیبی