9431035 رضا جبی

Subject

Date

1)

a) $\bar{g}_1 = 0 \rightarrow g_1 - \bar{g}_1 = (-5\ -4\ -3\ -2\ -1\ 0\ 1\ 2\ 3\ 4\ 5)$

$\bar{g}_2 = 11 \rightarrow g_2 - \bar{g}_2 = (14\ 5\ -2\ -7\ -10\ -10\ -7\ -2\ 5\ 14)$

$$Corr(g_1, g_2) = \frac{cov(g_1, g_2)}{\sigma(g_1)\sigma(g_2)}$$

$$cov(g_1, g_2) = \frac{1}{m-1} \sum_{k=1}^{m} (g_{1k} - \bar{g}_1)(g_{2k} - \bar{g}_2)$$

$$= \frac{1}{9} (14 \times -5 + 5 \times -4 + \dots + 5 \times 4 + 14 \times 5) = 0$$

$$\rightarrow Corr(g_1, g_2) = \frac{cov(g_1, g_2)}{\sigma(g_1)\sigma(g_2)} = \frac{0}{\;} = 0$$

پس با استفاده از این معیار، این ۲ داده هیچ شباهتی به هم ندارند

و در یک خوشه قرار نمی‌گیرند.

b) Mutual Information

$$I(g_1, g_2) = H(g_1) + H(g_2) - H(g_1, g_2)$$

$$H(g_1) = -\sum_{i=1}^{15} P_i \log P_i = -\sum \frac{1}{15} \log \frac{1}{15} = -\log \frac{1}{15} = \log 15$$

$$H(g_2) = -\sum_{i=1}^{5} P_i \log P_i = -\sum_{i=1}^{5} \frac{1}{5} \log \frac{1}{5} = -\log \frac{1}{5} = \log 5$$

$$H(g_1, g_2) = -\sum_{i=1}^{10} \sum_{j=1}^{10} P_{ij} \log P_{ij} =$$

۱۰۰ حالت ممکن بوده که P_{ij} برای ۹۰ مورد ۰ و برای ۱۰ مورد $\frac{1}{10}$

$$= -10 \left(\frac{1}{10} \log \frac{1}{10} \right) = -\log \frac{1}{10} = \log 10$$

$$I(g_1, g_2) = \log 15 + \log 5 - \log 15 = \log 5$$

c) بله از آن جا که داده‌ها رابطه‌ی خطی ندارند بنابراین corr آن ۲ (g_1 و g_2) صفر شده چون corr رابطه‌ی خطی را نشان می‌دهد

اما MI رابطه‌ی غیر خطی را نیز نشان می‌دهد و ۰ نشده است

پس g_1 و g_2 رابطه‌ی غیرخطی با یکدیگر دارند.

PAPCO

(۴)

a → euclidean: $\sqrt{(2-1)^2+1^2+1^2+1^2} = \sqrt{4} = \boxed{2}$

→ correlation: $corr(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{0}{\sigma_x \sigma_y} = \boxed{0}$

$$cov(x,y) = \frac{1}{m-1}\sum_{k=1}^{m}(x_k-\bar{x})(y-\bar{y})$$
$$= \frac{1}{3}(0+0+0+0) = 0$$

→ cosine: $cos(x,y) = \frac{\langle x,y\rangle}{\|x\|\,\|y\|} = \frac{2+2+2+2}{\sqrt{4+4+4+4}\cdot\sqrt{4}} = \frac{8}{4\cdot 2} = \boxed{1}$

---

b → jaccard: $J = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} = \frac{0}{\quad} = \boxed{0}$

→ euclidean: $\sqrt{1^2+1^2+1^2+1^2} = \sqrt{4} = \boxed{2}$

→ correlation: $corr(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$

$x-\bar{x} = (\frac{1}{4}, -\frac{1}{4}, \frac{1}{4}, -\frac{1}{4})$

$y-\bar{y} = (-\frac{1}{4}, \frac{1}{4}, -\frac{1}{4}, \frac{1}{4})$

$$cov(x,y) = \frac{1}{3}(-\frac{1}{4} - \frac{1}{4} - \frac{1}{4} - \frac{1}{4}) = -\frac{1}{3}$$

$$\sigma_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2} = \sqrt{\frac{1}{3}(\frac{1}{4}\times 4)} = \sqrt{\frac{1}{3}}$$

$$\sigma_y = \sqrt{\quad} = \quad = \sqrt{\frac{1}{3}}$$

$$corr(x,y) = \frac{-\frac{1}{3}}{\sqrt{\frac{1}{3}}\sqrt{\frac{1}{3}}} = \boxed{-1}$$

→ cosine: $cos(x,y) = \frac{\langle x,y\rangle}{\|x\|\,\|y\|} = \frac{0}{\quad} = \boxed{0}$

c)

→ bhattacharyya (x,y): $D_B(x,y) = -\ln\left(\sum \sqrt{x_i y_i}\right)$

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

$= \boxed{-\ln(3)}$

→ Correlation: $corr(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y}$

$\bar{x} = \frac{2}{3}$

$x - \bar{x} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, -\frac{2}{3}, -\frac{2}{3}, \frac{1}{3}\right)$

$\bar{y} = \frac{2}{3} \Rightarrow y - \bar{y} = \left(\frac{1}{3}, \frac{1}{3}, -\frac{2}{3}, \frac{1}{3}, -\frac{2}{3}, \frac{1}{3}\right)$

$cov(x,y) = \frac{1}{6}\left(\frac{1}{9} + \frac{1}{9} - \frac{2}{9} - \frac{2}{9} + \frac{4}{9} + \frac{1}{9}\right) = \frac{1}{18}$

$\sigma_x = \sqrt{\frac{1}{6}\left(\frac{1}{9} \times 4 + \frac{4}{9} \times 2\right)} = \sqrt{\frac{1}{6}\left(\frac{12}{9}\right)} = \sqrt{\frac{4}{18}}$

$\sigma_y = \sqrt{\frac{1}{6}(\quad\quad)} = \sqrt{\frac{4}{18}}$

$corr(x,y) = \frac{\frac{1}{18}}{\sqrt{\frac{4}{18}}\sqrt{\frac{4}{18}}} = \frac{\frac{1}{18}}{\frac{4}{18}} = \boxed{\frac{1}{4}}$

→ manhattan = نرم ۱

$(0+0+1+1+0+0) = \boxed{2}$

(۳)

a) پیوسته - کمی - نسبت

b) گسسته - کیفی - ترتیبی

c) پیوسته - کمی - نسبت

d) پیوسته - کمی - بازه

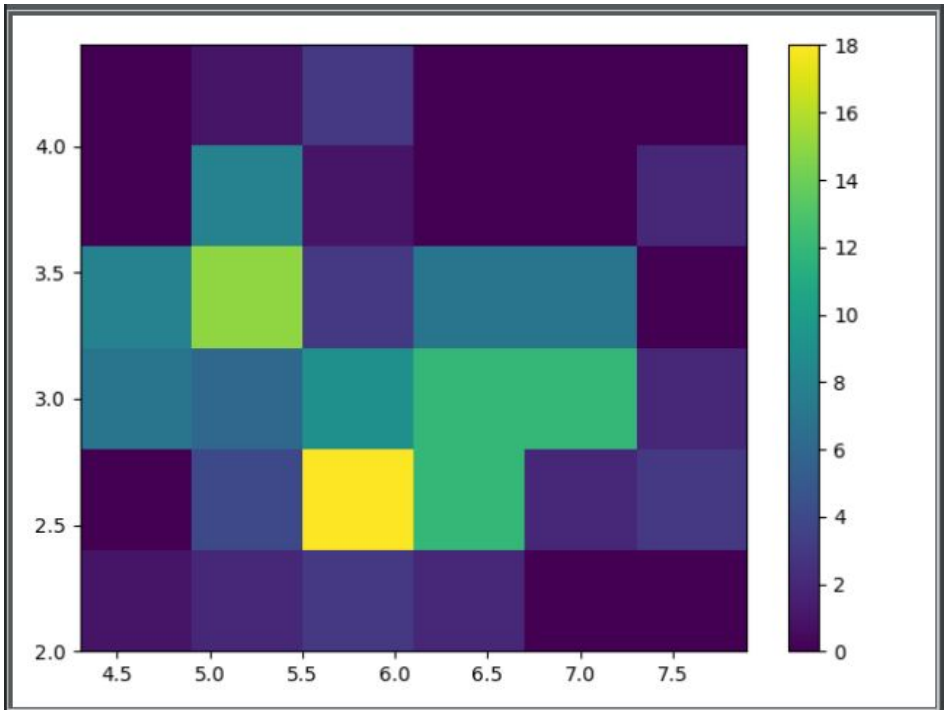
e) گسسته - کیفی - ترتیبی

f) گسسته - کیفی - ترتیبی

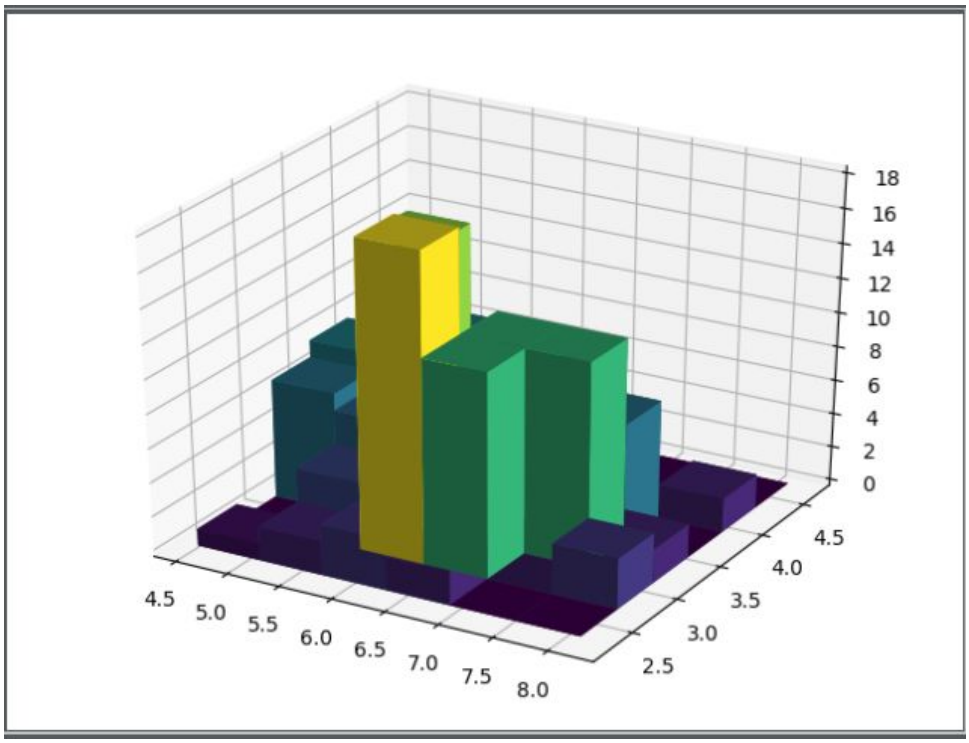
حواس ناظر انسانی تنها تاریک و روشن و متقابل را تشخیص می‌دهد

۴-

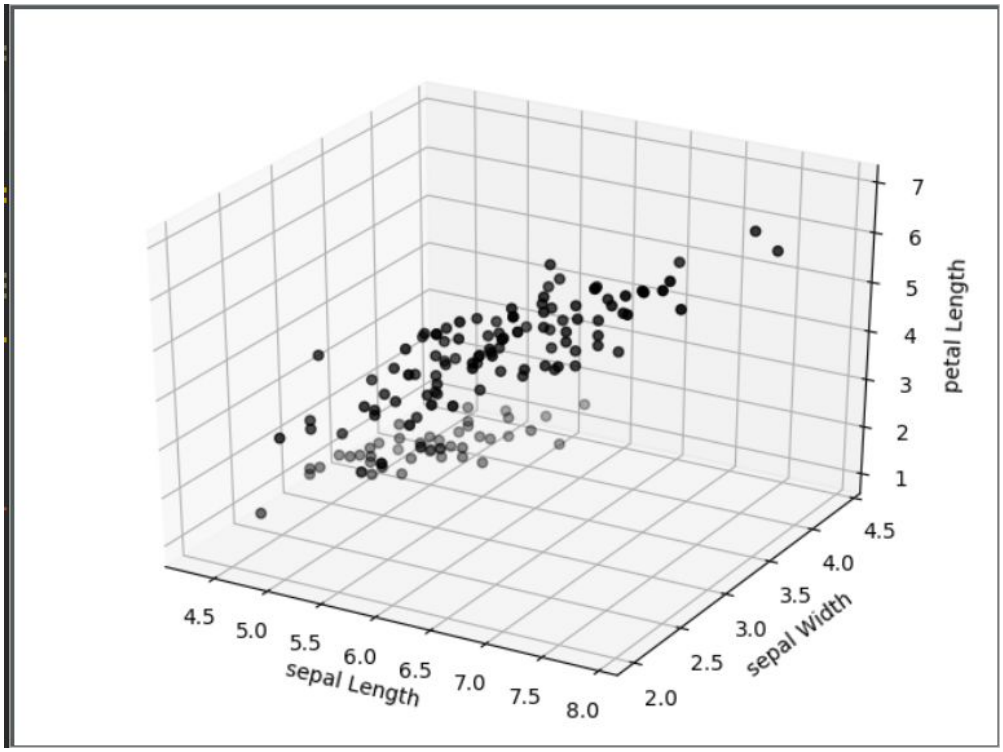
a) دو ویژگی اول انتخاب شده اند. محور x بیانگر فیچر اول و محور y بیانگر فیچر دوم است. رنگ نزدیکتر به زدر بیانگر آن است که تعداد بیشتری داده دارای فیچر اول و فیچر دوم برابر با مقدار مشخص شده وجود دارند



b) دو ویژگی اول انتخاب شده اند. محور x بیانگر فیچر اول و محور y بیانگر فیچر دوم است. محور عمودی هرچه بالاتر باشد بیانگر آن است که تعداد بیشتری داده دارای فیچر اول و فیچر دوم برابر با مقدار مشخص شده وجود دارند



c) سه ویژگی اول انتخاب شده اند. محور x بیانگر فیچر اول و محور y بیانگر فیچر دوم است و محور z بیانگر فیچر سوم است و هرنقطه بیانگر یک داده است. هر محور با استفاده از label مشخص شده مربوط به کدام اتریبیوت است.



(d

```
Mean :  [5.843333333333334, 3.0540000000000003, 3.758666666666666, 1.1986666666666668]
Variance :  [0.6811222222222223, 0.18675066666666668, 3.092424888888889, 0.5785315555555555]
```

e) دو فیچر اول انتخاب شده اند نتیجه یک آرایه ی ۲ در ۲ است که درایه ی i,j بیانگر کواریانس فیچر iام با jام است.

```
[[ 0.68569351 -0.03926846]
 [-0.03926846  0.18800403]]
```

f) همانند e

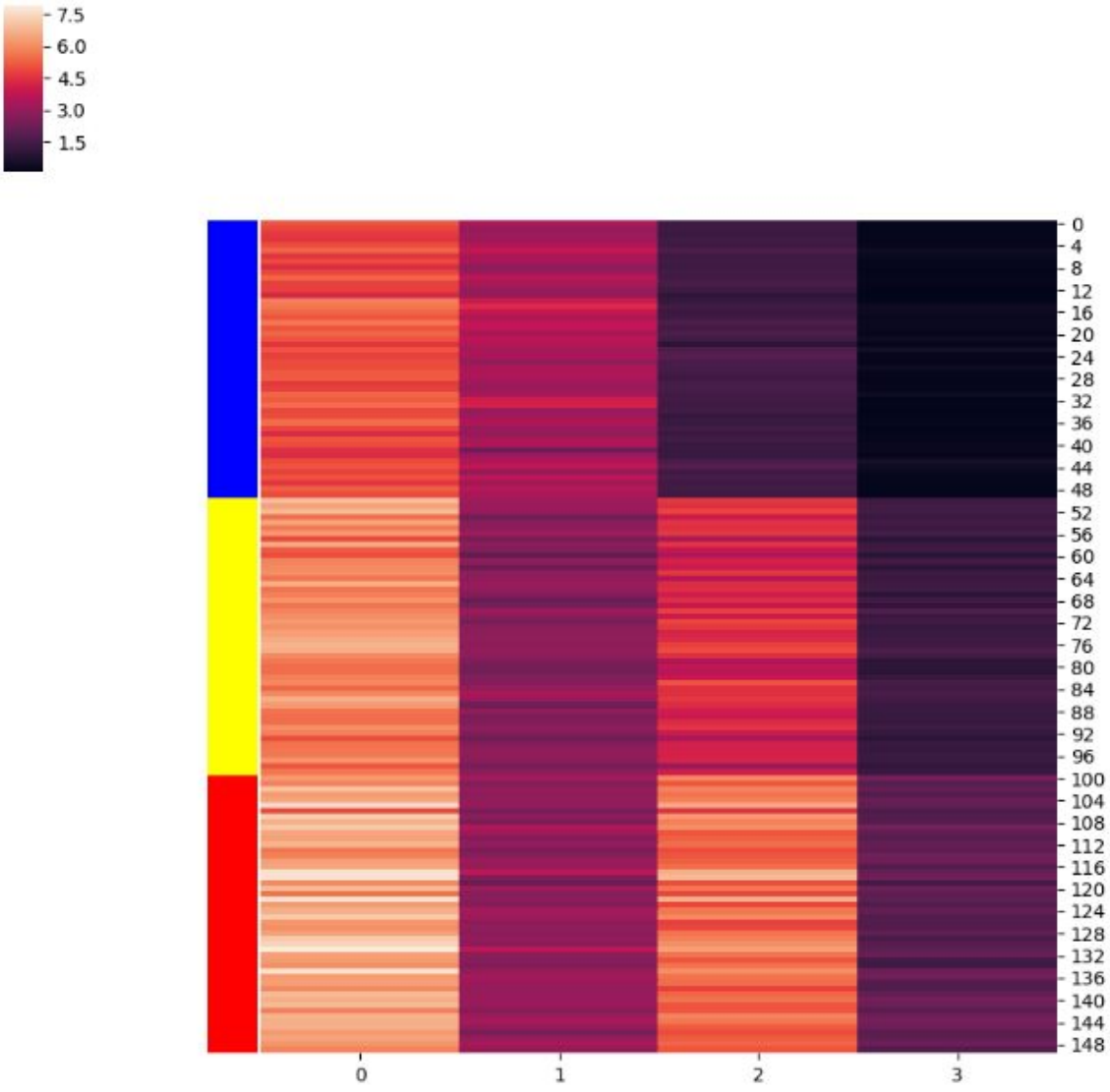
```
[[ 1.         -0.10936925]
 [-0.10936925  1.        ]]
```

g) ماتریس همبستگی فیچر ها را برای گیاه مدنظر بدست می آوریم هر درایه که عدد بیشتری داشته باشد فیچر های آن وابستگی بیشتری دارند.

```
[[1.         0.45722782 0.86422473 0.28110771]
 [0.45722782 1.         0.40104458 0.53772803]
 [0.86422473 0.40104458 1.         0.32210822]
 [0.28110771 0.53772803 0.32210822 1.        ]]
```

فیچر های اول و سوم عدد بزرگتری دارد پس بیشترین شباهت برای این ۲ فیچر است

(h



Clustermap کشیده شده نشان می‌دهد که فیچر سوم و چهارم بهتر از ۲ فیچر دیگر می‌توانند بین گیاهان مختلف تمایز قایل شوند و تفکیک کنند. همچنین به نظر می‌رسد فیچر سوم بهترین تفکیک گیاهان از یکدیگر را داشته باشد.

5- همانطور که در صورت سوال ذکر شده از مدل صورت سوال استفاده کرده ام و ضرایب آن را با train کردن بدست آورده ام که بسیار نزدیک به واقعیت شده است.

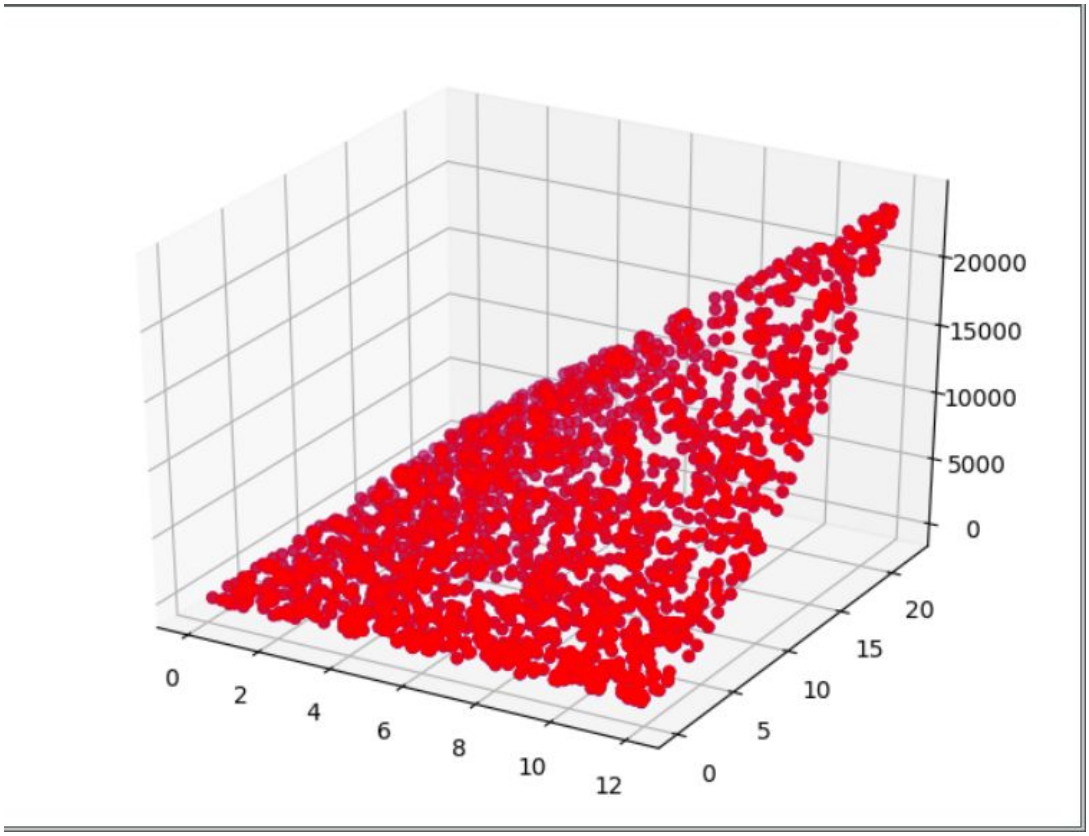
برای gradient descent :

```
Reading data ...
Data read!
Wait for training. It can take a few seconds!
Beta =  [[0.01608638]
 [0.0673921 ]
 [2.0030327 ]
 [4.00727882]]
SSE =  1371227.3654922154
SSE for train data :  1371227.3654922154
SSE for test data :  425516.4013891396

Process finished with exit code 0
```

مقایسه ی y واقعی داده ها و y پیش بینی شده (امتیازی):

Y واقعی با رنگ قرمز و y پیش بینی شده با رنگ آبی مشخص شده. مشخص است که داده ها کاملا روی هم قرار گرفته اند.



برای stochastic gradient descent :

```
Reading data ...
Data read!
Wait for training. It can take a few seconds!
Beta =  [[0.94145994]
 [3.01608337]
 [2.00105947]
 [3.99976178]]
SSE =  70.97660980508954
SSE for train data :  70.97660980508954
SSE for test data :  52.98509267995236

Process finished with exit code 0
```

مقایسه ی y واقعی داده ها و y پیش بینی شده (امتیازی):
Y واقعی با رنگ قرمز و y پیش بینی شده با رنگ آبی مشخص شده. مشخص است که داده ها کاملا روی هم قرار گرفته اند.