بسمه تعالی

تمرین های سری اول

۱– داده یک بعدی x={-5,23,17.6,7.23,1.11} را به روش های زیر نرمال سازی کنید.

الف) حرکت نقطه ی اعشار روی فاصله [-1,1].

ب) نرمال سازی Max–Min روی فاصله [0,1].

ج) نرمال سازی Max–Min روی فاصله [-1,1].

د) نرمال سازی بر اساس انحراف استاندارد.

ه) نتایج روش ها را مقایسه کنید و مزایا و معایب هر یک را بیان کنید.

۲–جدول زیر ضریب هوشی گروهی از افرادی را نشان می دهد که برای شرکت در مسابقه دانش عمومی شرکت کرده اند.

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
IQ	92	107	83	101	107	92	99	119	93	106	105	88	106	90	97	118	120	72	100	104

الف) آنها را به پنج <بین> با عرض مساوی تقسیم کنید.

ب) آنها را به پنج <بین> با فرکانس مساوی تقسیم کنید.

ج) برای دو بند فوق هیستوگرام <بین> ها را رسم کنید.

۳– مقادیر یک ویژگی برای دو کلاس به صورت زیر می باشد :

class ω_1:	3.5	3.7	3.9	4.1	3.4	3.5	4.1	3.8	3.6	3.7
class ω_2:	3.2	3.6	3.1	3.4	3.0	3.4	2.8	3.1	3.3	3.6

با استفاده از آزمون **T** بررسی کنید ، آیا ویژگی مناسبی است ؟

۴– برای داده $\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$:

الف) مولفه های اساسی آن و درصد واریانس کل برای هر مولفه را بدست آورید.

ب) با استفاده از روش **SVD** دو مولفه اول را به دست آورید.

۵– در طبقه بندی وضعیت سالم از ناسالم از ویژگیهای دو بعدی استفاده شده است. بردار میانگین کلاسها m_1 , m_2 و ماتریس کواریانس آنها $\sum_1, \sum_2$ است. در نظر است که ابعاد بردارهای این دو کلاس را با تابع معیار $J = tr\left([s]_w^{-1}[s]_b\right)$ که در آن $[s]_w$, $[s]_b$ به ترتیب ماتریسهای پراکندگی داخل

کلاس و ماتریس پراکندگی بین کلاسها است، کاهش دهیم. ماتریس تبدیل را برای اینکار به دست آورید.

۶–الف) نشان دهید که اگر تابع چگالی احتمال دو کلاس گوسی با ماتریس کوواریانس یکسان Σ باشد، با اضافه شدن یک ویژگی جدید x_{m+1} به بردار ویژگی قبل ،مقدار جدید دیورژانس بصورت تکراری زیر محاسبه می شود :

$$d_{ij}(x_1,\ldots,x_{m+1}) = d_{ij}(x_1,\ldots,x_m) + \frac{[(\mu_i-\mu_j)-(\underline{\mu_i}-\underline{\mu_j})^T\Sigma^{-1}r]^2}{\sigma^2 - r^T\Sigma^{-1}r}$$

که μ_i و μ_j مقادیر متوسط x_{m+1} برای دو کلاس ، σ^2 واریانس ،**r** بردار کوواریانس متقابل با عناصر دیگر **x** است و $\underline{\mu_i}$ و $\underline{\mu_j}$ بردار متوسط **x** قبل از اضافه شدن x_{m+1} است.

ب) اگر x_{m+1} با ویژگی های انتخاب شده قبلی $x_1,\ldots,x_m$ ناهمبسته باشد ، پس :

$$d_{ij}(x_1,\ldots,x_{m+1}) = d_{ij}(x_1,\ldots,x_m) + \frac{(\mu_i-\mu_j)^2}{\sigma^2}$$

ج) اگر ویژگی ها به طور آماری مستقل باشند ، پس : $d_{ij}(x_1,\ldots,x_l) = \sum_{i=1}^{l} d_{ij}(x_i)$

۷–روش پیشنهادی مقاله ضمیمه(An Algorithm for Discretization of Real Value Attributes Based on Interval Similarity) و الگوریتم**Chi2** را برای گسسته سازی داده های **Breast** اجرا نموده و مقایسه کنید.

۸–مجموعه داده های بیان پروتئین موش با ۱۰۸۰ نمونه ، ۸۲ ویژگی و ۸ کلاس را از مخزن یادگیری ماشینی **UCI** درنظربگیرید(https://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression). از **82** ویژگی فقط ۷۷ ویژگی مقدار حقیقی دارند ، که فقط از آنهااستفاده خواهیم کرد.

الف) مقادیر از دست رفته در داده های اصلی را بازیابی کنید. برای (ب) ، (ج) ، و (د) نتایج خود راهم با وهم بدون بازیابی مقادیر از دست رفته گزارش کنید.

ب) از **PCA** برای کاهش بعدداده ها به فضای ۲ و ۳ بعدی استفاده کنید. نتایج خودرا با استفاده از نمودارهای پراکندگی (scatter plots) نمایش دهید.واریانس بین کلاسی و درون کلاسی را قبل و بعداز **PCA**مقایسه کنید.مقادیر J_1 یا J_3(به درس رجوع شود) را قبل و بعداز **PCA**مقایسه کنید.

ج) بند(ب) را با استفاده از **LDA** تکرار کنید.

(د) بند(ب) را با استفاده از **ISOMAP** تکرار کنید.

(ه) نتایج خود را تفسیر کنید. چرا فکر می کنید، که این سه روش متفاوت عمل می کنند؟ کدام یک از این نگاشت ها جدایی پذیری را به بهترین شکل ممکن حفظ می کنند؟