بنام خدا

الگوشناسی آماری (CE-725)

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف

تمرینات سری اول – بهار ۱۳۸۹

به نکات زیر توجه فرمائید:

۱. زمان تحویل تمرینات در سایت درس مشخص شده است. دقت نمائید که زمانبندی‌های تعیین شده قابل تغییر نیستند.

۲. تمرینات را با عنوان SPR-HWx-8xxxxxxx (مثلا SPR-HW1-88300785) و در یک فایل فشرده با همین نام به آدرس Muhammadi@ce.sharif.edu ایمیل زده و در اولین جلسه بعد از زمان تحویل، بصورت پرینت شده تحویل استاد درس دهید.

۳. گزارش شما باید مختصر و مفید باشد. برای تمرینات پیاده‌سازی که با لوگوی مشخص شده‌اند باید کد مطلب نوشته شده ضمیمه گزارش شده و تمامی خروجی‌های برنامه‌ها در گزارش شما ذکر شوند.

سوال ۱) یک مجموعه داده (data set) با ماتریس کوواریانس زیر را در نظر بگیرید:

$$\Sigma = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}$$

الف) با توجه به ماتریس کوواریانس داده شده به سوالات زیر پاسخ دهید:

۱. این مجموعه داده چند بعدی است (هر نمونه شامل چند ویژگی می‌باشد)؟

۲. تعداد نمونه‌های مجموعه داده چه تعداد بوده است؟

۳. چه وابستگی‌ها و همبستگی‌هایی بین ابعاد مختلف داده‌ها وجود دارد؟

۴. پراکندگی داده‌های بر روی کدام یک از ابعاد بیشتر است؟

ب) آیا یک ماتریس کوواریانس همیشه متقارن است؟ چرا؟

پ) مقادیر ویژه و بردارهای ویژه ماتریس فوق را بدست آورید و با توجه به آنها به سوالات زیر پاسخ دهید:

۱. این ماتریس کوواریانس چند مقدار ویژه غیر صفر دارد؟

۲. صفر شدن یک مقدار ویژه چه مفهومی می‌تواند داشته باشد؟

۳. زاویه بین هر دو جفت بردارهای ویژه بدست آمده را محاسبه کنید (سه حالت مختلف). به چه نتیجه‌ای می‌رسید؟ آیا نتیجه بدست آمده برای هر ماتریس کوواریانس دیگری نیز برقرار است؟ چرا؟

ت) با فرض اینکه میانگین مجموعه داده فوق $\mu = [5 \quad 0 \quad 3]$ باشد، فعالیت‌های زیر را در مطلب انجام دهید:

۱. یک نمونه تصادفی ۱۰۰تایی مطابق این توزیع گاوسی تولید کنید (از تابع randn مطلب که برای این منظـور ارائـه شده است، استفاده کنید).

۲. فرض کنید که V ماتریسی باشد که هر ستون آن یکی از بردارهای ویژه باشـد، سـتون اول، بـردار ویـژه متنـاظر بـا بزرگترین مقدار ویژه، ستون دوم بردار ویژه متناظر با دومین بزرگترین مقدار ویژه و .... هر کدام از ۱۰۰ نمونه تولیدی بخش قبل را با رابطه تبدیل $Y_i = (X_i - \mu) \times V$ به فضای جدیدی که آن را S' می‌نامیم ببرید.

۳. داده‌ها ها را در هر دو فضای قبلی و جدید plot کرده و آنها را با هم مقایسه کنید.

۴. بردار کوواریانس داده‌های تبدیل یافته را بدست آورده و در مورد آن به سوالات زیر پاسخ دهید:

a. چه وابستگی‌ها و همبستگی‌هایی بین ابعاد مختلف داده‌ها وجود دارد؟

b. این ماتریس کوواریانس چند مقدار ویژه غیر صفر دارد؟

c. بردارهای ویژه این ماتریس را بدست آورید.

سوال ۲) داده‌های دو بعدی متعلق به یک کلاس را در نظر بگیرید، که دارای یک فرم گاوسی با پارامترهای زیر می‌باشند.

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad p(X|\omega) \sim N(\mu, \Sigma), \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \ \& \ \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}; \sigma_{12} = \sigma_{21} = \sigma$$

الف) رابطه فاصله اقلیدسی بین نقطه x و مرکز گاوسی (μ) را بنویسید.

ب) رابطه فاصله ماهالانوبیس بین نقطه x و مرکز گاوسی (μ) را بنویسید (با ضرب ماتریس‌ها، رابطه را ساده کنید).

پ) رابطه بدست آمده در بخش قبل برای حالتی که ماتریس کوواریانس قطری باشد، به چه صورتی ساده خواهد شد؟ آیـا ارتباطی بین رابطه بدست آمده با رابطه فاصله اقلیدسی وجود دارد؟

ت) رابطه‌های بدست آمده در بخش‌های الف و ب را با هم مقایسه کنید. در چه شرایط دو فاصله با هم برابر خواهند شد؟

ث) با توجه به نتایج مقایسه‌های بخش ت، در چه شرایطی بهتر است که از فاصله ماهالانوبیس استفاده شود؟ آیا شـرایطی وجود دارد که در آن استفاده از فاصله اقلیدسی منطقی‌تر از فاصله ماهالانوبیس باشد (به جز در شرایطی که دو فاصله نتایج یکسانی را بر می‌گردانند)؟

ج) چه راهی برای محاسبه فاصله ماهالانوبیس بین دو نمونه از یک توزیع گاوسی پیـشنهاد مـی‌کنیـد. بـا اسـتفاده از ایـن راهکار در چه حالتی فاصله ماهالانوبیس بین دو نقطه با فاصله اقلیدسی بین آنها برابر خواهد بود؟

سوال ۳) یک مساله کلاسه‌بندی دو کلاسه در یک فضای دو بعدی را در نظر بگیرید. فرض کنید که نمونه‌های آزمایشی داده شده برای هر کدام از کلاس‌ها، برای مدل‌سازی با یک توزیع گاوسی مناسب باشد. هر کـدام از ایـن گاوسـی‌هـا بـا دو پـارامتر میانگین و کوواریانس‌اش شناخته می‌شود.

الف) یک تابع مطلب بنویسید که پارامترهای این دو گاوسی را گرفته و در دو تصویر موارد زیر را نمایش دهد:

تصویر اول: دو توزیع گاوسی، مرز بین دو کلاس با استفاده از فاصله اقلیدسی و لیبل‌گذاری زیر فضاهای بدست آمده

تصویر دوم: دو توزیع گاوسی، مرز بین دو کلاس با استفاده از فاصله ماهالانوبیس و لیبل‌گذاری زیر فضاهای بدست آمده

نحوه بدست آوردن مرز بین دو کلاس: اگر فاصله هر نقطه از فضا را از مرکز دو کلاس محاسبه کنیم و لیبل کلاسـی را بـه آن نقطه بدهیم که آن نقطه دارای فاصله کمتری از مرکز آن کلاس می‌باشد، می‌توان تمام نقاط فضا را لیبل‌گـذاری کـرد. معمـولا تعداد زیادی از نقاط هم‌لیبل در مجاورت همدیگر قرار دارند و یک زیر فضا را تشکیل می‌دهند. یا بعبارتی دیگر، در یـک مـساله

کلاسه‌بندی فضا به چند زیر فضا با لیبل‌های مشخص افراز می‌شود. البته در عمل لازم نیست که برای تمامی نقاط فضا این عمل انجام شود، کافی است نقاط مرزی استخراجی شود (نقاطی که فاصله آنها از مراکز هر دو کلاس یکسان است).

ب) پارامترهای زیر را در نظر بگیرید که از مجموعه داده‌های آموزشی مختلف استخراج شده‌اند. برای هر کدام از حالت‌های داده شده، پارامترها را به تابع فوق داده و خروجی‌ها را در گزارش خود بیاورید:

1	μ_1=[10 10] μ_2=[10 3] Σ_1 = [1 0; 0 1] Σ_2 = [1 0; 0 1]	6	μ_1=[10 10] μ_2=[4 7.5] Σ_1 = [6 2; 2 2] Σ_2 = [6 2; 2 2]	11	μ_1=[10 13] μ_2=[10 3] Σ_1 = [1 0; 0 6] Σ_2 = [6 0; 0 1]	16	μ_1=[10 10] μ_2=[10 10] Σ_1 = [1 0; 0 8] Σ_2 = [2 0; 0 1]
2	μ_1=[10 10] μ_2=[10 3] Σ_1 = [2 0; 0 2] Σ_2 = [3 0; 0 3]	7	μ_1=[10 10] μ_2=[10 3] Σ_1 = [6 2; 2 2] Σ_2 = [6 2; 2 2]	12	μ_1=[13 5] μ_2=[5 13] Σ_1 = [1 0; 0 6] Σ_2 = [6 0; 0 1]	17	μ_1=[10 10] μ_2=[10 10] Σ_1 = [1 0; 0 8] Σ_2 = [4 0; 0 4]
3	μ_1=[10 10] μ_2=[10 3] Σ_1 = [1 0; 0 1] Σ_2 = [2 0; 0 2]	8	μ_1=[10 10] μ_2=[4 7.5] Σ_1 = [3 1; 1 1] Σ_2 = [6 2; 2 2]	13	μ_1=[10 5] μ_2=[5 13] Σ_1 = [1 0; 0 6] Σ_2 = [6 0; 0 1]	18	μ_1=[10 10] μ_2=[10 10] Σ_1 = [2 0; 0 16] Σ_2 = [2 0; 0 2]
4	μ_1=[10 10] μ_2=[10 3] Σ_1 = [6 0; 0 3] Σ_2 = [2 0; 0 2]	9	μ_1=[10 10] μ_2=[10 3] Σ_1 = [16 0; 0 4] Σ_2 = [3 0; 0 1]	14	μ_1=[5 13] μ_2=[10 5] Σ_1 = [4 2; 2 2] Σ_2 = [1 0; 0 6]		
5	μ_1=[10 10] μ_2=[10 3] Σ_1 = [6 0; 0 2] Σ_2 = [6 0; 0 2]	10	μ_1=[10 10] μ_2=[10 3] Σ_1 = [6 0; 0 2] Σ_2 = [3 0; 0 1]	15	μ_1=[10 10] μ_2=[10 10] Σ_1 = [1 0; 0 8] Σ_2 = [8 0; 0 1]		

لطفا نتایج این بخش را با دقت بررسی نمائید. حالت‌های ارائه شده تقریبا حالت‌های کاملی برای این مساله می‌باشند. در بررسی‌های خود سعی کنید دلیل بیاورید که چرا مرزها به این صورت در آمده‌اند (از لحاظ موقعیت مکانی یا انحنا و عدم انحنای مرز یا ...).

سوال نمره اضافه: یک گاوسی دوبعدی را در فضای سه بعدی در نظر بگیرید که سطح مقطع آن بر روی صفحه xy (با z=0) واقع شده و محور z از مرکز آن می‌گذرد. اگر یک ابر صفحه موازی صفحه xy این گاوسی را قطع کند، سطح مقطع حاصل از این تقاطع، یک بیضی خواهد شد. بسته به اینکه ارتفاع این ابر صفحه چقدر باشد، اندازه بیضی حاصله نیز متفاوت خواهد بود.

برای یکی بیضی تشکیل شده به صورت فوق، صرفنظر از مقدار z، x درصد داده‌های گاوسی تشکیل شده، داخل بیضی قرار خواهد گرفت (یا بعبارتی دیگر، x درصد از حجم گاوسی بالای ابر صفحه قطع کننده قرار خواهد گرفت).

الف) اگر مقدار x مشخص باشد، ابر صفحه باید در چه ارتفاعی با گاوسی قطع داده شود.

ب) پارامترهای بیضی تشکیل شده را بدست آورید (دو مرکز و قطرهای اصلی و فرعی بیضی).