بنام خدا

الگوشناسی آماری (CE-725)

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف

تمرینات سری اول – بهار 1390

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 بزنید.

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) پس از افزایش شدید حقوق گارگران، شرکت بسته‌بندی میوه عباس آقا و برادران، تصمیم به دسته‌بندی خودکار میوه‌های ورودی به انبار کرد. ورودی انبار چهار نوع میوه سیب، توت‌فرنگی، موز و انگور است که بر روی نوارهای نقاله به اتاقک دسته‌بندی منتقل می‌شوند. در این اتاقک هر نوع میوه باید در جعبه مربوط به خود قرار گیرد.

در این رابطه به سوال‌های زیر پاسخ دهید:

الف) آیا تصمیم این شرکت عاقلانه است؟ خوبی‌ها و بدی‌های دسته‌بندی خودکار میوه‌ها را ذکر کنید.

ب) از چه حسگرهایی می‌توان استفاده کرد؟ طرحی کلی از نحوه قرارگیری و استفاده این حسگرها بیان کنید.

پ) چه پیش‌پردازش‌هایی (توسط انسان یا بصورت خودکار / قبل یا بعد از کار حسگرها) باید انجام شود تا نحوه نمایش میوه‌ها یا استخراج مشخصه‌ها آسان‌تر شود؟

ت) چه مشخصه‌هایی برای جداسازی این چهار کلاس مناسب است؟

ث) چه چالش‌هایی در راه دسته‌بندی این میوه‌ها وجود دارد و چگونه هر یک از مشخصه‌های بخش قبل به حل هر یک از چالش‌ها کمک می‌کند؟

ج) تخمینی از بازگشت سرمایه بدهید. به بیان دیگر با بیان فرضیات مناسب، تخمینی از بازده انسان و ماشین برای این کار ارائه نمائید.

چ) دو تا از مهم‌ترین مشخصه‌های این مسئله دسته‌بندی را انتخاب کنید. پنج نمونه آموزشی از هر کلاس پیدا کنید (سایت google منبع خوبی برای این کار است!) و آنها را در فضای مشخصه‌های انتخاب شده (دو بعدی) نمایش داده و مرز تصمیم‌گیری در این فضا را رسم کنید.

سوال ۲) یک مجموعه داده (data set) با ماتریس کوواریانس زیر را در نظر بگیرید:

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

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 توابع آماده مطلب برای تولید نمونه‌های تصادفی گاوسی استفاده نمائید).

۲. فرض کنید که V ماتریسی باشد که هر ستون آن یکی از بردارهای ویژه باشد، ستون اول، بردار ویژه متناظر با 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 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 ۳) داده‌های دو بعدی متعلق به یک کلاس را در نظر بگیرید، که دارای یک فرم گاوسی با پارامترهای زیر می‌باشند.

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

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؟ آیا 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 شرایطی 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 پیشنهاد می‌کنید. با استفاده از این راهکار در چه حالتی فاصله ماهالانوبیس بین دو نقطه با فاصله اقلیدسی بین آنها برابر خواهد بود؟

سوال ٤) ماتریس FV (فایل mat. حاوی این ماتریس ضمیمه تمرین شده است) را در نظر بگیرید. این ماتریس حاوی ۵۱۹ داده ۱۳۹۹۶ بعدی می‌باشد. ابعاد این داده‌ها را با استفاده از T-test به ۱۰۰۰ بعد کاهش دهید (نتایج را در یک فایل mat ضمیمه کنید).

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 x، z 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).