بنام خدا

الگوشناسی آماری (CE-725)

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف

تمرینات سری اول (معرفی مفاهیم، استخراج و کاهش ویژگی)

بهار ۱۳۹۱

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@dml.ir 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 متلب نوشته شده ضمیمه گزارش شده و تمامی خروجی‌های برنامه‌ها در گزارش شما ذکر شوند.

سوال ۱) پس از افزایش شدید حقوق کارگران، شرکت بسته‌بندی میوه عباس آقا و برادران، تصمیم به دسته‌بندی خودکار میوه‌های ورودی به انبار کرد. ورودی انبار چهار نوع میوه سیب، توت فرنگی، موز و انگور است که بر روی نواره ای 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 را بیان کنید.

پ) چه پیش پردازش‌هایی (توسط انسان یا به صورت 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 چالش ها 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 مساله دسته‌بندی را انتخاب کرده و بر اساس شهود خود از ویژگی‌های میوه‌ها، مرزهای تصمیم‌گیری را در این فضا را رسم کنید.

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

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

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 ۳) دو متغییر تصادفی X و Y را در نظر بگیرید. میانگین و انحراف از معیار را به ترتیب با μ و δ نشان می‌دهیم. رابطه بین X و Y به طرق مختلفی می‌تواند بیان شود. در این مساله از سه مورد زیر استفاده می‌کنیم:

- همبستگی: $\text{cov}(X,Y)=E((X-\mu_x)(Y-\mu_Y))=E(XY)-\mu_X\mu_Y$
- ضریب همبستگی: $\rho_{XY}=\dfrac{\text{cov}(X,Y)}{\delta_X\delta_Y}$
- اطلاعات متقابل: $I(X,Y)=H(X)-H(X|Y)=KL(P(X,Y)||P(X)P(Y))$ که در آن داریم:

$$D_{KL}(P_1(x)||P_2(x))=\int P_1(x)\ln\frac{P_1(x)}{P_2(x)}dx$$

الف) ثابت کنید قدر مطلق ضریب همبستگی بین دو متغییر تصادفی کمتر، مساوی یک است (یادآوری: نامساوی کوشی–شوارتز بیان می‌کند که $E(XY)^2\le E(X^2)E(Y^2)$).

ب) تحت چه شرایطی $\rho_{XY}=1$ می‌شود؟ در چه شرایطی $\rho_{XY}=-1$ می‌شود؟

ث) اگر مقدار $I(X,Y)=0$، آیا می توان نتیجه گرفت $\rho_{XY}=0$؟ اگر بله، ادعای خود را ثابت کنید. در غیر این صورت دو متغییر تصادفی ارائه کنید که برای آنها $I(X,Y)=0$ باشد، اما $\rho_{XY}=0$ نباشد.

پ) اگر $\rho_{XY}=0$ باشد، ایا می توان گفت $I(X,Y)=0$؟ اگر می توان، ادعای خود را ثابت کنید. در غیر این صورت دو متغییر تصادفی ارایه کنید که $\rho_{XY}=0$ است اما $I(X,Y)\neq 0$.

سوال ٤) 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 $\mu^T = [5 \quad 0 \quad 3]$ 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 منظور ارائه شده است، استفاده کنید).

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 $Y_i = V^T \times (X_i - \mu)$ 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 ۵) 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 کدام از این گاوسی‌ها با دو پارامتر میانگین و کوواریانس‌اش شناخته می‌شود.

الف) یک تابع مطلب بنویسید که پارامترهای این دو گاوسی را گرفته و در یک تصویر، دو توزیع گاوسی و مرز بین دو کلاس را با استفاده از فاصله ماهالانوبیس نمایش دهد.

نکته: اگر فاصله هر نقطه از فضا را از مرکز دو کلاس محاسبه کنیم و لیبل کلاسی را به آن نقطه بدهیم که آن نقطه دارای فاصله کمتری از مرکز آن کلاس می‌باشد، می‌توان تمام نقاط فضا را لیبل‌گذاری کرد. معمولا تعداد زیادی از نقاط هم‌لیبل در مجاورت همدیگر قرار دارند و یک زیر فضا را تشکیل می‌دهند. یا بعبارتی دیگر، در یک مساله 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 به هر صورتی 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 (برای هر تصویر سعی کنید درک کنید که چرا مرزها به این صورت در آمده‌اند).

1	$\mu_1=[10\ 10]^T$ $\mu_2=[10\ 3]^T$ $\sum_1=[1\ 0;\ 0\ 1]$ $\sum_2=[1\ 0;\ 0\ 1]$	6	$\mu_1=[10\ 10]^T$ $\mu_2=[4\ 7.5]^T$ $\sum_1=[6\ 2;\ 2\ 2]$ $\sum_2=[6\ 2;\ 2\ 2]$	11	$\mu_1=[10\ 13]^T$ $\mu_2=[10\ 3]^T$ $\sum_1=[1\ 0;\ 0\ 6]$ $\sum_2=[6\ 0;\ 0\ 1]$	16	$\mu_1=[10\ 10]^T$ $\mu_2=[10\ 10]$ $\sum_1=[1\ 0;\ 0\ 8]$ $\sum_2=[2\ 0;\ 0\ 1]$
2	$\mu_1=[10\ 10]^T$ $\mu_2=[10\ 3]^T$ $\sum_1=[2\ 0;\ 0\ 2]$ $\sum_2=[3\ 0;\ 0\ 3]$	7	$\mu_1=[10\ 10]^T$ $\mu_2=[10\ 3]^T$ $\sum_1=[6\ 2;\ 2\ 2]$ $\sum_2=[6\ 2;\ 2\ 2]$	12	$\mu_1=[13\ 5]^T$ $\mu_2=[5\ 13]^T$ $\sum_1=[1\ 0;\ 0\ 6]$ $\sum_2=[6\ 0;\ 0\ 1]$	17	$\mu_1=[10\ 10]^T$ $\mu_2=[10\ 10]^T$ $\sum_1=[1\ 0;\ 0\ 8]$ $\sum_2=[4\ 0;\ 0\ 4]$
3	$\mu_1=[10\ 10]^T$ $\mu_2=[10\ 3]^T$ $\sum_1=[1\ 0;\ 0\ 1]$ $\sum_2=[2\ 0;\ 0\ 2]$	8	$\mu_1=[10\ 10]^T$ $\mu_2=[4\ 7.5]^T$ $\sum_1=[3\ 1;\ 1\ 1]$ $\sum_2=[6\ 2;\ 2\ 2]$	13	$\mu_1=[10\ 5]^T$ $\mu_2=[5\ 13]^T$ $\sum_1=[1\ 0;\ 0\ 6]$ $\sum_2=[6\ 0;\ 0\ 1]$	18	$\mu_1=[10\ 10]^T$ $\mu_2=[10\ 10]^T$ $\sum_1=[2\ 0;\ 0\ 16]$ $\sum_2=[2\ 0;\ 0\ 2]$
4	$\mu_1=[10\ 10]^T$ $\mu_2=[10\ 3]^T$ $\sum_1=[6\ 0;\ 0\ 3]$ $\sum_2=[2\ 0;\ 0\ 2]$	9	$\mu_1=[10\ 10]^T$ $\mu_2=[10\ 3]^T$ $\sum_1=[16\ 0;\ 0\ 4]$ $\sum_2=[3\ 0;\ 0\ 1]$	14	$\mu_1=[5\ 13]^T$ $\mu_2=[10\ 5]^T$ $\sum_1=[4\ 2;\ 2\ 2]$ $\sum_2=[1\ 0;\ 0\ 6]$		
5	$\mu_1=[10\ 10]^T$ $\mu_2=[10\ 3]^T$ $\sum_1=[6\ 0;\ 0\ 2]$ $\sum_2=[6\ 0;\ 0\ 2]$	10	$\mu_1=[10\ 10]^T$ $\mu_2=[10\ 3]^T$ $\sum_1=[6\ 0;\ 0\ 2]$ $\sum_2=[3\ 0;\ 0\ 1]$	15	$\mu_1=[10\ 10]^T$ $\mu_2=[10\ 10]^T$ $\sum_1=[1\ 0;\ 0\ 8]$ $\sum_2=[8\ 0;\ 0\ 1]$		

سوال ٦) هدف از این سوال نشان دادن میزان انرژی موجود در تعداد کمی از مولفه‌های اصلی PCA است. برای این منظور از تصاویر خاکستری بهره می‌گیریم؛ که در آنها هر پیکسل دارای مقداری بین صفر تا ۲۵۵ است. اگر M سطر یک تصویر MxN را پشت سر هم قرار دهیم، یک بردار ویژگی با M*N مولفه به دست می‌آید، که هر مولفه دارای مقداری عددی (بین صفر تا ۲۵۵) است. این روش بدیهی‌ترین روش تبدیل یک تصویر به یک بردار ویژگی است که ما نیز در این تمرین از آن استفاده می‌کنیم. از آنجایی که ابعاد بردارهای ویژگی به دست آمده، حتی برای تصاویر کوچک نیز بسیار بالا خواهند بود، استفاده از روش‌های کاهش ابعاد اجتناب ناپذیر است.

مجموعه تصاویر صورت داده شده را از سایت درس بگیرید. ابتدا هر کدام از تصاویر را به یک بردار ویژگی تبدیل کنید. سپس میانگین این بردارهای ویژگی را به دست آورده و این میانگین را از همه بردارهای ویژگی کم کنید (به این کار «نرمال‌سازی» تصاویر گفته می‌شود).

پس از نرمال‌سازی بردارهای ویژگی، ماتریس کوواریانس این بردارها را به دست آورده و مقادیر و بردارهای ویژه آن را محاسبه کنید. بردارهای ویژه به دست آمده یک فضای جدیدی را می‌سازند که تعداد ابعاد آن با فضای قبلی برابر است. هر بردار ویژگی در فضای قبلی (هر تصویر صورت) معادل یک بردار ویژگی در فضای جدید است.

فرض کنید به جای استفاده از تمامی بردارهای ویژگی فقط از k مولفه با مقادیر ویژه بزرگ‌تر استفاده کرده و فضای جدیدی با k بعد بسازیم. هر تصویر در این فضا با k بعد نمایش داده می‌شود. از آن جایی که مقدار k بسیار کمتر از تعداد ابعاد اصلی است،

نگهداری هر تصویر با این k عدد، فضای بسیار کمتری را برای نگهداری لازم دارد (تصاویر فشرده می‌شوند). البته بدیهی است که اگر از این بردارهای k بعدی برای بازیابی تصاویر اصلی استفاده کنیم، تصاویر بازسازی شده دارای کیفیتی پایین‌تر از تصاویر اصلی خواهند شد.

در این رابطه به سوالات زیر پاسخ دهید:

الف) برای بازیابی تصاویر از بردارهای k بعدی، چه اطلاعاتی به غیر از خود بردارهای ویژگی فضای جدید می‌بایست ذخیره شوند؟

ب) کمترین مقدار k چقدر باید باشد که تصاویر بازسازی شده دارای کیفیتی عینا برابر تصاویر اصلی باشند؟

ج) کمترین مقدار k چقدر باید باشد که چشم شما متوجه کاهش کیفیت تصاویر بازسازی شده، نسبت به تصاویر اصلی نشود؟

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).