بنام خدا

الگوشناسی آماری (CE-725)

دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شریف

تمرینات سری دوم – 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-HW1-89300785) 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) 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 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 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 نقاط مرزی استخراجی شود (نقاطی که فاصله آنها از مراکز هر دو کلاس یکسان است).

ب) پارامترهای صفحه‌ی بعد 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.

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 ۲) ماتریس FV (فایل mat. حاوی این ماتریس ضمیمه تمرین شده است) را در نظر بگیرید. این ماتریس حاوی ۵۱۹ داده ۱۳۹۹۶ بعدی می‌باشد. می‌خواهیم با استفاده از تحلیل مولفه اصلی به کاهش ابعاد این داده‌ها اقدام کنیم (یک نمونه کد PCA ضمیمه تمرین شده است. از princomp مطلب هم می‌توانید استفاده کنید).

1	μ_1=[10 10] μ_2=[10 3] $\sum_1$ = [1 0; 0 1] $\sum_2$ = [1 0; 0 1]	6	μ_1=[10 10] μ_2=[4 7.5] $\sum_1$ = [6 2; 2 2] $\sum_2$ = [6 2; 2 2]	11	μ_1=[10 13] μ_2=[10 3] $\sum_1$ = [1 0; 0 6] $\sum_2$ = [6 0; 0 1]	16	μ_1=[10 10] μ_2=[10 10] $\sum_1$ = [1 0; 0 8] $\sum_2$ = [2 0; 0 1]
2	μ_1=[10 10] μ_2=[10 3] $\sum_1$ = [2 0; 0 2] $\sum_2$ = [3 0; 0 3]	7	μ_1=[10 10] μ_2=[10 3] $\sum_1$ = [6 2; 2 2] $\sum_2$ = [6 2; 2 2]	12	μ_1=[13 5] μ_2=[5 13] $\sum_1$ = [1 0; 0 6] $\sum_2$ = [6 0; 0 1]	17	μ_1=[10 10] μ_2=[10 10] $\sum_1$ = [1 0; 0 8] $\sum_2$ = [4 0; 0 4]
3	μ_1=[10 10] μ_2=[10 3] $\sum_1$ = [1 0; 0 1] $\sum_2$ = [2 0; 0 2]	8	μ_1=[10 10] μ_2=[4 7.5] $\sum_1$ = [3 1; 1 1] $\sum_2$ = [6 2; 2 2]	13	μ_1=[10 5] μ_2=[5 13] $\sum_1$ = [1 0; 0 6] $\sum_2$ = [6 0; 0 1]	18	μ_1=[10 10] μ_2=[10 10] $\sum_1$ = [2 0; 0 16] $\sum_2$ = [2 0; 0 2]
4	μ_1=[10 10] μ_2=[10 3] $\sum_1$ = [6 0; 0 3] $\sum_2$ = [2 0; 0 2]	9	μ_1=[10 10] μ_2=[10 3] $\sum_1$ = [16 0; 0 4] $\sum_2$ = [3 0; 0 1]	14	μ_1=[5 13] μ_2=[10 5] $\sum_1$ = [4 2; 2 2] $\sum_2$ = [1 0; 0 6]		
5	μ_1=[10 10] μ_2=[10 3] $\sum_1$ = [6 0; 0 2] $\sum_2$ = [6 0; 0 2]	10	μ_1=[10 10] μ_2=[10 3] $\sum_1$ = [6 0; 0 2] $\sum_2$ = [3 0; 0 1]	15	μ_1=[10 10] μ_2=[10 10] $\sum_1$ = [1 0; 0 8] $\sum_2$ = [8 0; 0 1]		

جدول ۱: پارامترهای مربوط به سوال اول

الف) با استفاده از PCA بدون اینکه هیچ data loss‌ی داشته باشیم، می‌توانیم این ۱۳۹۹۶ بعد را به چند بعد کاهش دهیم؟

ب) با استفاده از PCA برای اینکه حداکثر ۱ درصد data loss داشته باشیم، می‌توانیم این ۱۳۹۹۶ بعد را به حداقل چند بعد کاهش دهیم؟

پ) با استفاده از PCA برای اینکه حداکثر ۵ درصد data loss داشته باشیم، می‌توانیم این ۱۳۹۹۶ بعد را به حداقل چند بعد کاهش دهیم؟

ت) اگر یکصد مولفه اصلی حاصل از PCA را به عنوان فضای یکصدبعدی جدید انتخاب کنیم، بردن داده‌ها به این فضای جدید، باعث چه مقدار data loss خواهد شد؟ (یا چه مقدار انرژی موجود در داده‌ها حفظ خواهد شد؟

سوال ۳) روی مجموعه داده ۱۵۰تایی iris (فایل mat. حاوی این ماتریس ضمیمه تمرین شده است) PCA و MDA را اعمال کرده و تعداد ابعاد را از چهار بعد به دو بعد کاهش دهید (دو نمونه کد برای PCA و MDA، ضمیمه تمرین شده‌اند). در یک plot دو بعدی، داده‌های کاهش بعد یافته با PCA و در plot دیگر داده‌های کاهش بعد یافته با MDA را نمایش دهید. در هر دو مورد، داده‌های کلاس‌های مختلف را با رنگ‌ها یا علایم متفاوت نمایش دهید (در فایل mat. ضمیمه، لیبل کلاس هر داده نیز به شما داده شده است). با توجه به دو plot بدست آمده کدام یک دو مجموعه داده تغییر یافته را برای کلاسه‌بندی مناسب‌تر می‌دانید؟ آیا نتایج مشاهده شده با انتظارات شما از عملکرد PCA و MDA مطابقت دارند؟ چرا؟

سوال ٤) با یک مجموعه داده یک کلاسه‌بند را دو بار آموزش می‌دهیم. بار اول داده‌ها را به صورت تصادفی به نسبت ۵۰–۵۰ به دو دسته آموزش و تست تقسیم کرده و عملیات آموزش کلاسه‌بند را انجام می‌دهیم و بار دوم داده‌ها را به صورت تصادفی به نسبت ۲۰–۸۰ به دو دسته آموزش و تست تقسیم می‌کنیم (مجموعه آموزشی، بزرگتر از مجموعه تست) و عملیات آموزش کلاسه‌بند را انجام می‌دهیم. کارایی کلاسه‌بند اول بر روی مجموعه آموزشی‌اش ۸۰ درصد و کارایی کلاسه‌بند دوم بر روی مجموعه آموزشی‌اش ۹۰ درصد بوده است. آیا برای یک مجموعه داده جدید کلاسه‌بند دوم بهتر از کلاسه‌بند اول عمل می‌کند؟ نظر خود را توضیح دهید.

سوال **۵**) فرض کنید در فضای یک بعدی، تابع چگالی برای دو کلاس w_1 و w_2 مطابق توزیع زیر است:

$$f(x) = Ae^{-\frac{|x-a_i|}{b_i}}$$

که در آن $i=1,2$ می‌باشد.

الف) A را برحسب پارامترهای دیگر بیابید.

ب) نسبت تشابه (likelihood ratio) را به دست آورید.

پ) نمودار نسبت تشابه را برای $a_1=1,b_1=2,a_2=0,b_2=1$ رسم نمایید.

سوال نمره اضافه) در کاربردهایی نظیر پردازش صورت تعداد ابعاد داده‌ها بسیار بالا خواهد شد. مثلا هر تصویر کوچک شده ۲۵۶*۲۵۶ به یک بردار ۶۵۵۳۶ بعدی تبدیل خواهد شد. اعمال PCA بر روی این داده‌ها مشکلات عملی‌ای را بدنبال خواهد داشت (ماتریس کوواریانس فضای بسیار بالایی را نیاز خواهد داشت). چه راهکاری برای حل این مشکل می‌توانید ارائه دهید؟ دلیل اعتبار راهکار خود را توضیح دهید.

راهنمایی: برای مشاهده یک راهکار مناسب می‌توانید به مقاله Turk و Pentland تحت عنوان Face Recognition Using Eigenfaces مراجعه نمایید.