سلام بر دانشجویان عزیز، چند نکته مهم:

1. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.
2. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخ‌ها می‌باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.
3. کدهای ارسال شده **بدون گزارش** فاقد نمره می‌باشند.
4. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.
5. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW5_StudentNumber داشته باشد.
6. از بین سوالات **شبیه سازی** حتما به هر دو مورد پاسخ داده شود.
7. نمره تمرین ۱۰۰ نمره هست.
8. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین **صفر** خواهد شد.
9. در صورت داشتن سوال، از طریق ایمیل melikasadeghi16@mail.com سوال خود را مطرح کنید.

سوال ۱: (۱۰ نمره)

فرض کنید از روش PCA جهش کاهش بعد داده‌های نشان داده شده در شکل زیر استفاده می‌کنیم. اگر داده‌ها به دو کلاس (+ و –) تعلق داشته باشند. مطلوب است:

الف) نمایش محور اول و دوم PCA برای داده‌هایی که در شکل نمایش داده شده‌اند. آیا نوع برچسب داده‌ها در انتخاب محورهای PCA موثر است؟ چرا؟ (۵ نمره)

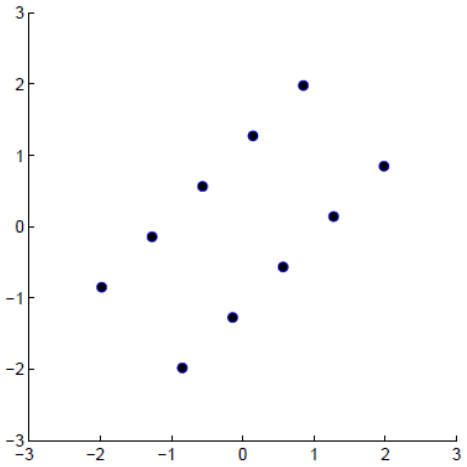
ب) برای داده‌های نشان داده شده در شکل برچسب‌گذاری دودویی (+ و -) را به گونه‌ای انجام دهید که طبقه‌بند نزدیک‌ترین همسایه در فضای اصلی (دو بعد) و کاهش بعد یافته (بعد اول PCA) خطاهای زیر را داشته باشد. اندازه‌گیری خطا به کمک روش leave-one-out cross validation انجام شده است. (۵ نمره)

ب.1)

2D data: 100% error
1D data from PCA: 0% error

ب.2)

2D data 0% error
1D data from PCA: 100% error



سوال ۲: (۱۰ نمره)

عبارت

$$J = \frac{1}{n_1 n_2} \sum_{y_i \in Y_1} \sum_{y_j \in Y_2} (y_i - y_j)^2$$

پراکندگی کل درون گروهی(within group scatter) را اندازه می‌گیرد. نشان دهید که این عبارت را می‌توان به صورت زیر هم نوشت: (۱۰ نمره)

$$J = (m_1 - m_2)^2 + \frac{1}{n_1} s_1^2 + \frac{1}{n_2} s_2^2$$

که در آن m_1 و m_2 بیانگر میانگین و مقادیر S_2 و S_1 نیز به ترتیب بیانگر پراکندگی Y_1 و Y_2 هستند.

سوال ۳: (۲۰ نمره)

به سوالات زیر پاسخ دهید.

الف) روش EM را برای توزیع پواسن به دست آورید.

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

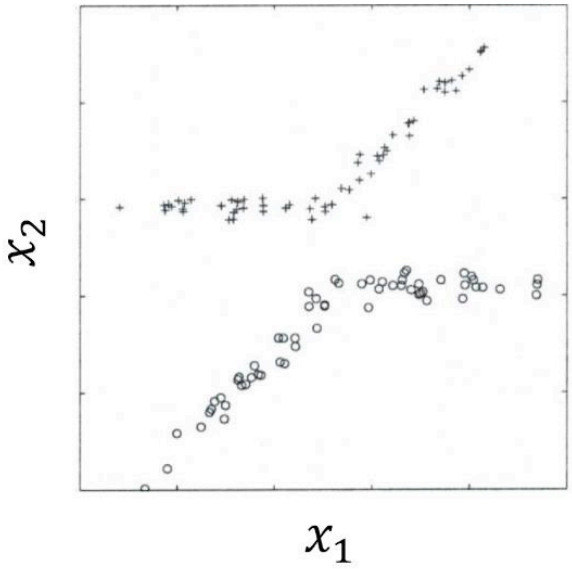
ب) در PCA ماتریس کواریانس دیتا $C = X^T X$ بصورت جمع وزن‌دار مقادیر و بردارهای ویژه‌ی خود نوشته می‌شود (λ مقدار ویژه و p بردار ویژه ماتریس C است)

$$C = \sum \lambda_i \boldsymbol{p_i} \boldsymbol{p_i^T}$$

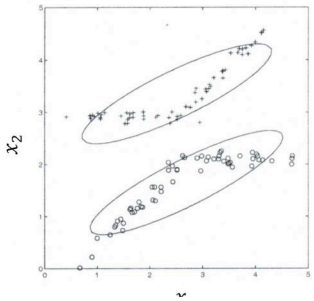
بصورت ریاضی نشان دهید که مقدار ویژه‌ی اول ($\lambda 1$) برابر با واریانس تصویر (projection) داده‌ها بر روی اولین principal component است.

سوال ۴: (۱۰ نمره)

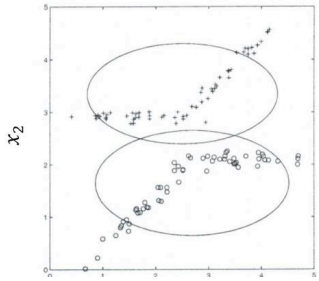
شکل زیر داده‌های مربوط به دو کلاس + و o را نشان می‌دهد.



الف) فرض کنید دو مدل زیر نتیجه اعمال الگوریتم EM و استفاده از Gaussian Mixture Model باشد، با ذکر دلیل توضیح دهید کدام یک از مدل‌های زیر مناسب‌تر است.

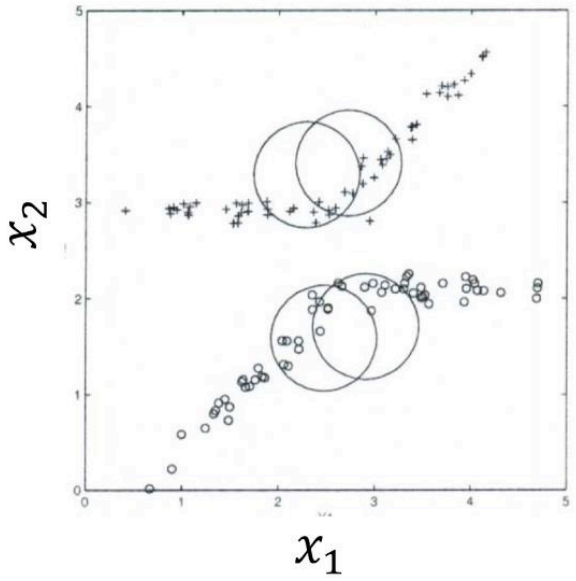


الف

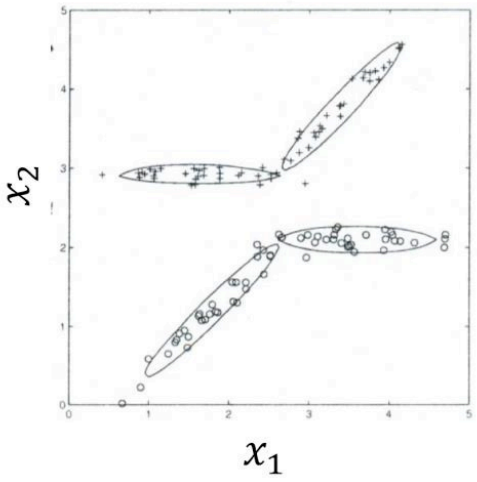


ب

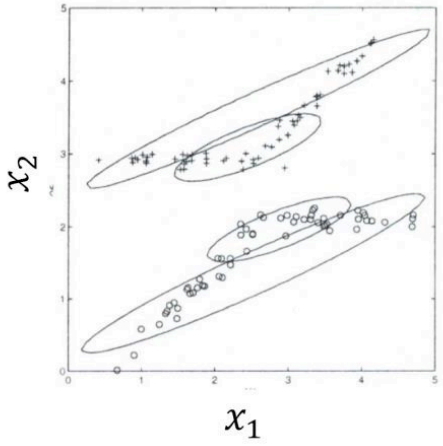
ب) فرض کنید دو تابع گوسی برای مدل کردن + ها و دو تابع گوسی برای مدل کردن o ها استفاده شود. شرایط اولیه چهار تابع گوسی در شکل زیر نمایش داده شده است.



با ذکر دلیل توضیح دهید کدام یک از موارد زیر خروجی اولین گام الگوریتم EM و استفاده از Gaussian Mixture Modelخواهد بود.



الف



ب

سوال ۵: (**شبیه‌سازی**، ۳۰ نمره)

یکی از تکنیک‌های متداول در فشرده‌سازی تصاویر PCA می‌باشد که تعداد principle components (PCs) در کیفیت تصویر و نرخ فشرده‌سازی تصویر تاثیرگذار است. حال ما برآنیم که با استفاده از PCA تصاویر را فشرده کنیم و به فضایی با تعداد ویژگی‌های کم‌تر منتقل شویم تا عملیات تشخیص تصویر بهتر انجام شود. در این سوال شما از شامل ۲۱۳ تصویر که دارای ۶ حالت (Happy, Fear, Angry, Disgust, Surprise and sad) استفاده می‌کنید. برای بارگذاری تصاویر از image_loader.py که با تصاویر پیوست شده است استفاده نمایید.

الف) مقادیر ویژه از PCA را به ترتیب کاهشی رسم نمایید و بیان نمایید که چگونه می‌توان تعداد کامپوننت مناسب را در فرآیند فشرده‌سازی تشخیص داد.

ب) ۴ مقدار ویژه اول و ۴ مقدار ویژه نهایی (eigenfaces) را نشان داده و تحلیل کنید.

حال لیبل کلاس‌ها را در نظر گرفته و ماتریس پراکندگی درون کلاسی و بین کلاسی را محاسبه نمایید تا روش LDA را پیاده‌سازی کنیم

د) از LDA کمک بگیرید و مقادیر ویژه را مرتب نمایید و مقادیر ویژه ماتریس جداپذیری را در قالب نزولی رسم نمایید.

ه) در یک نمودار مقدار $\text{trace}\left(S_W^{-1}S_B\right)$ (seperability measure) نسبت به تعداد ویژگی‌ها رسم نمایید و در مورد تاثیر تعداد ویژگی‌ها بر آن بحث کنید.

سوال ۶: (**شبیه‌سازی**، ۲۰ نمره)

در این سوال قصد داریم با استفاده از پیاده سازی الگوریتم EM و تخمین مدل GMM به طبقه بندی تصاویر بپردازیم. برای سادگی داده های دو کلاس (تیم فوتبال منچستر و چلسی) را بررسی می کنیم که بتوانیم از دو فیچر R, B از (RGB) به عنوان فیچر ها بهره بگیریم.

الف) با در نظر گرفتن K=2 به عنوان تعداد مولفه های(component) ، الگوریتم EM را برای تخمین پارامتر های توزیع های GMM مربوط به هر یک از دو کلاس پیاده سازی کنید. پارامتر های به دست آمده برای GMM مربوط به هر کلاس را در گزارش خود ذکر کنید. نمودار های داده های هر دو کلاس و کانتور های مدل های GMM فیت شده به آن ها را رسم کنید .

ب) مقدار بهینه k را با یکی از روش های cross validation مانند (p leave out, k-fold) پیدا کنید. توجه کنید که رسم نمودار های لازم برای نتیجه گیری k بهینه در گزارش لازم است.