دانشگاه تهران

به نام خدا

دانشکده مهندسی برق و کامپیوتر

تمرین سری پنجم یادگیری ماشین

پردیس دانشکده های فنی

سلام بر دانشجویان عزیز، چند نکته مهم:

1. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.
2. نکته ی مهم در گزارش نویسی روشن بودن پاسخ ها می باشد، 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 می باشند.
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. در نظر داشته باشید که می‌بایست به حداقل ۷۰ نمره از سوالات تحلیلی و ۶۰ نمره از سوالات شبیه‌سازی پاسخ دهید.
7. نمره تمرین ۱۳۰ نمره می باشد و حداکثر تا نمره ۱۵۰ (**۲۰ نمره امتیازی**) می توانید کسب کنید.
8. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می باشد و کل تمرین برای طرفین **صفر** خواهد شد.
9. در صورت داشتن سوال، از طریق ایمیل maedehtoosi@gmail.com یا mesbahamirhossein@gmail.com سوال خود را مطرح کنید.

سوال ۱: (۱۵ نمره)

با توجه به Probabilistic principal component analysis (PPCA) به بخش های زیر پاسخ دهید.

الف : اگر برای PCA احتمالی (PPCA) فضای توزیع latent زیر را در نظر بگیریم:

$$p(z) = N(z|0, I)$$

و توزیع شرطی برای متغیر مشاهده شده $x \in R^d$ برابر باشد با:

$$p(x|z) = N(x|Wz + \mu, \sigma^2 I)$$

بررسی کنید که کواریانس توزیع marginal : $p(x) = N(x|\mu, C)$ برابر است با $C = WW^T + \sigma^2 I$ و نتیجه را تفسیر کنید.

ب: عبارتی برای posterior متغیرهای latent : $p(z|x)$! استخراج کنید.

سوال ۲: (۱۰ نمره)

الف: مشکلات محاسباتی و عددی که استفاده از PCA و LDA را در داده هایی با ابعاد بالا وجود دارد را به همراه راه حل آنها ذکر کنید.

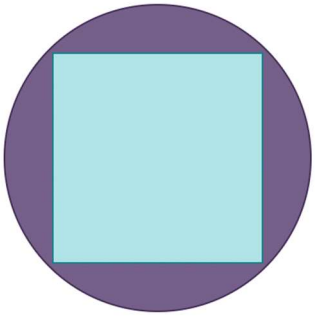
ب : مقدار LDA را برای مجموعه داده های زیر محاسبه کنید:

$$X_1 = \{(4.2,1.3), (2,3.5), (2,4), (3,5.8), (4,5)\}$$

$$X_2 = \{(8,10), (6,8.2), (9,5), (8,7), (10,9)\}$$

سوال ۳: (۱۰ نمره)

کره ای با شعاع یک که بیش از سه بعد دارد را در نظر بگیرید[1]. درون این کره یک hypercube قرار گرفته است. برای مثال شکل زیر نمایش این کره در دو بعد را نشان می دهد. حال به سوالات زیر پاسخ دهید:



الف: برای بعد خاص d ، یک عبارت برای حجم hypercube داده شده بدست آورید. عبارت را برای یک ، دو و سه بعد استخراج کنید و سپس به ابعاد بالاتر تعمیم دهید.

ب: اگر نسبت حجم hypercube به حجم hypersphere به سمت بی نهایت میل کند چه رخ می دهد؟ این نسبت را در یک ، دو و سه بعد ارائه دهید و سپس به ابعاد بالاتر تعمیم دهید.

سوال ۴: (۱۵ نمره)

با توجه به جدول زیر به موارد زیر پاسخ دهید:

الف: مقدار μ_{+1} و μ_{-1} را به همراه ماتریس B(between-class scatter matrix) محاسبه کنید.

ب : مقدار S_{+1} و S_{-1} را به همراه ماتریس S(within-class scatter matrix) محاسبه کنید.

ج : داده ها را به همراه بهترین جهت که آن ها را از هم جدا می کند با استفاده از بردارهای ویژه رسم کنید.

i	X_i^T	y_i
X_1^T	(1,1)	1
X_2^T	(2,4)	1
X_3^T	(2.5,1)	-1
X_4^T	(3,2)	-1

[1] Hypersphere

سوال ۵ : (۱۰ نمره)

با توجه به نقاط داده شده در جدول زیر ، فرض کنید که K = 2 می باشد ، و در ابتدا نقاط به خوشه ها به شرح زیر اختصاص می یابد:

$C_1 = \{x_1 \ ، x_2 \ ، x_4\}$

$C_2 = \{x_3 \ ، x_5\}$

الگوریتم k-means تا جایی که خوشه ها تغییر نکنند با فرض های زیر اعمال کنید:

الف : فاصله معمول اقلیدسی یا L2-norm به عنوان فاصله بین نقاط

ب: فاصله منهتن یا L1-norm

	X_1	X_2
X_1^T	0	2
X_2^T	0	0
X_3^T	1.5	0
X_4^T	5	0
X_5^T	5	2

سوال ۶: (**شبیه سازی**، ۲۰ نمره)

کاهش ابعاد با استفاده از PCA تکنیک متداولی برای فشرده کردن تصاویر است. تعداد کامپوننت های مورد استفاده بر نرخ فشردگی (compression rate) و کیفیت تصویر تاثیرگذار است. در این سوال شما از دیتاست FER2013 که ضمیمه شده است استفاده میکنید.

الف : مقادیر ویژه از PCA را به ترتیب کاهشی رسم نمایید و بیان نمایید که چگونه میتوان تعداد کامپوننت مناسب را در فرآیند فشرده سازی تشخیص داد؟

ب: 4 مقدار ویژه اول و 4 مقدار ویژه نهایی(eigenfaces) را (برای یک کلاس دلخواه) نشان دهید و تحلیل کنید که این تصاویر بیانگر چه می باشند؟

ج: حال طبقه بند K-NN را با k= 1, 2 را یک بار برا داده های کاهش بعد یافته و یک بار بر داده های خالص اعمال کنید و CCR و ماتریس کانفیوژن راگزارش نمایید و مقایسه نمایید.

د: اکنون مقدار کامپوننت تابع PCA را متغیر گرفته و CCR (مربوط به طبقه بند نزدیکترین همسایه) را برحسب تعداد کامپوننت PCA رسم نمایید و تحلیل کنید.

سوال ۷: (**شبیه سازی**، ۲۰ نمره)

در این سوال از طبقه بند بیز به عنوان معیاری برای مقایسه ی خطای هر مرحله از مجموع مربعات خطا استفاده کنید و قسمت الف و ب را با توجه به داده های دیتاست Madelon انجام دهید.

الف : الگوریتم Sequential Forward Selection را برای داده های ذکرشده پیاده سازی نمایید

ب : الگوریتم Sequential Backward Eliminationرا برای داده های ذکرشده پیاده سازی نمایید.

ج: مزیت Backward Elimination را نسبت به Forward Selection به صورت کامل شرح دهید .

د: برای هر دو الگوریتم بالا بهترین ویژگی ها و همچنین نمودار خطا بر حسب تعداد ویژگی ها را رسم نمایید. سپس دقت و سرعت دو الگوریتم را مقایسه نمایید.

*** در این سوال الگوریتم خواسته شده را باید خودتان پیاده سازی کنید و مجاز به استفاده از کتابخانه های آماده نیستید، اما در مورد طبقه بند میتوانید از توابع آماده استفاده کنید.

سوال ۸: (**شبیه سازی**، ۲۰ نمره)

در این سوال میخواهیم از Gaussian mixture model به عنوان یک مدل generative استفاده کنیم. برای این کاربرد نیاز به کار با دیتاست اعداد دست نوشته داریم. این دیتاست را میتوانید از این لینک تهیه کنید.

الف) پس از لود کردن و بررسی دیتاست، به کمک نمودار AIC تعداد componentهای مناسب برای fit کردن یک مدل GMM را برای این دیتا به دست بیاورید. نمودار AIC را رسم کرده و نتیجه به دست آمده را تحلیل کنید.

ب) پس از fit کردن یک مدل gmm با تعداد component به دست آمده در مرحله قبل، 100 دیتاپوینت از این مدل به دست آمده سمپل بگیرید و این دیتاپوینت‌ها را نمایش دهید. پس از fit کردن مدل بررسی کنید که آیا همگرایی رخ داده است یا خیر.

ج) مراحل الف و ب را پس از اعمال کاهش بعد بر روی دیتا با روش pca انجام دهید. لازم است برای نمایش دیتا پس از سمپل‌گیری عکس این تبدیل را بر روی دیتای تولید شده انجام دهید.

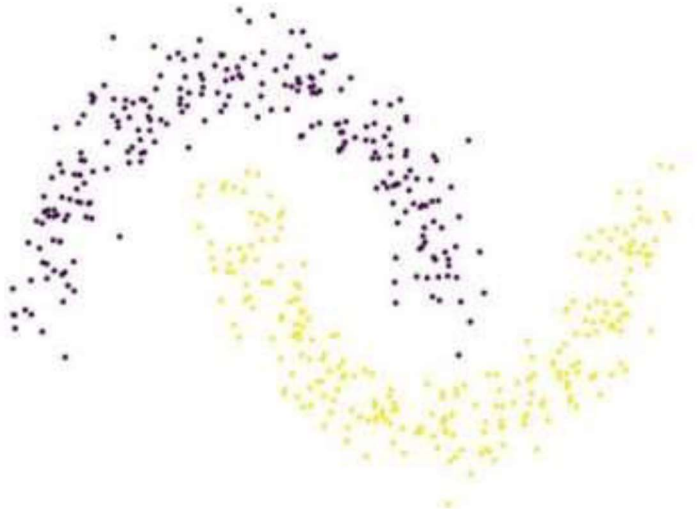
د) نتایج تولید شده برای حالت بدون کاهش بعد و با کاهش بعد را مقایسه کنید و در صورت بهتر شدن نتایج در بخش ج دلیل آن را توضیح دهید.

سوال ۹: (**شبیه سازی**، ۲۰ نمره)

در این سوال مجاز به استفاده از کتابخانه*ای آماده نیستید*.

ابتدا دیتاست شکل زیر را با استفاده از قطعه کد زیر ایجاد کنید.

```
from sklearn import cluster, datasets, mixture
noisy_moons=datasets.make_moons(n_samples=500, noise=0.11)
```



شکل 1 – دیتاپوینت‌های دیتاست *moon*

الف) یک بار هر کلاس را با توزیع نرمال تقریب بزنید و پارامترهای آن را به دست آورده و کانتورهای مربوطه را رسم نمایید.

ب) این بار از روش GMM استفاده کنید. روش GMM را با تعداد مولفه‌های 1 تا 16 تست کنید و شکل داده‌ها و کانتورها را برای تعداد مولفه برابر با 3، 8 و 16 به دست بیاورید.

ج) تعداد مولفه‌های بهینه را با توجه به متریک‌های *AIC* و *BIC* به دست بیاورید.

سوال ۱۰: (۱۵ نمره)

روش *EM* را برای توزیع پواسون به دست آورید.

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

سوال ۱۱: (۱۵ نمره)

به سوالات زیر پاسخ دهید.

الف) منظور از *VC dimension* برای یک طبقه چیست؟

ب) تعریف *Vapnik* و *Chervonenkis* از *VC dimension* را بیان کنید.

ج) نشان دهید که برای یک طبقه بند خطی *vc dimsnesion* حداقل برابر با 3 و کمتر از 4 است.