به نام خدا

دانشکده‌ی مهندسی برق و کامپیوتر

تمرین سری پنجم یادگیری ماشین

دانشگاه تهران

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. 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_HW#_StudentNumber داشته باشد.
6. از بین سوالات **شبیه سازی** حتما به هر دو مورد پاسخ داده شود.
7. نمره تمرین ۱۰۰ نمره می‌باشد.
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 alireza.javid84@ut.ac.ir سوال خود را مطرح کنید.

سوال ۱ : (۱۵ نمره)

دو متغیر تصادفی نرمال در نظر بگیرید که میانگین دلخواه اما واریانس یکسان دارند.

$$p\left(y|\tilde{\theta}_i\right) = \frac{1}{\sqrt{2\pi}\tilde{s}} \exp[-(y-\tilde{\mu}_i)^2/\left(2\tilde{s}^2\right)]$$

که در آن $\tilde{\theta}_i = \begin{pmatrix} \tilde{\mu}_\iota \\ \tilde{s} \end{pmatrix}$ به ازای i=1,2 نشان دهنده سپل ها پس از تبدیل (projection) $\tilde{D}$. اثبات کنید که FLD (Fisher linear discriminant) می‌تواند از منفی نسبت لگاریتمی احتمال (negative log-likelihood ratio) نتیجه شود.

راهنمایی: log-likelihood ratio به صورت $r = \frac{P(\tilde{D}|\tilde{\theta}_1)}{P(\tilde{D}|\tilde{\theta}_2)}$

سوال۲ : (۱۵ نمره)

$P_x(\boldsymbol{x}|w_i)$ یک توزیع دلخواه با میانگین $\boldsymbol{\mu_i}$ و واریانس $\boldsymbol{\Sigma_i}$ می باشد. (که این توزیع می تواند نرمال نباشد)

تبدیل $y = \boldsymbol{w^T x}$ در نظر بگیرید و توزیع یک بعدی نیز به صورت $P(y|w_i)$ فرض کنید.

نشان دهید که عبارت

$$J_1(w) = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

با انتخاب زیر ماکسیمم می شود.

$$\boldsymbol{w = (\Sigma_1 + \Sigma_2)^{-1}(\mu_1 - \mu_2)}$$

سوال ۳ : (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) به نظر شما درصورتی که به جای ماتریس کوواریانس از متریک های دیگری (مثلا اطلاعات متقابل بین داده ها و لیبل) برای تبدیل استفاده شود چه بهبود هایی می تواند حاصل شود و چه معایبی در پی دارد؟

سوال ۴ : (۱۵ نمره)

در این سوال به Probabilistic principal component analysis (PPCA) می پردازیم.

الف) اگر برای PCA احتمالی (PPCA) فضای توزیع latent نرمال با میانگین صفر و واریانس σ^2 را در نظر بگیریم و توزیع شرطی برای متغیر مشاهده شده $x \in \mathbb{R}^d$ برابر باشد با :

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

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

ب) عبارتی برای $p(z|x)$ استخراج کنید.

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 به عنوان فاصله بین نقاط

	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**، ۱۵ 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 PCAرا به ترتیب کاهشی رسم نمایید و بیان نمایید که چگونه میتوان تعداد کامپوننت مناسب را در فرآیند فشرده سازی تشخیص داد؟

ب) 5 مقدار ویژه اول و 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)

از الگوریتم **KMeans** میتوان برای فشرده سازی تصاویر استفاده کرد. به این صورت که k رنگ از رنگ های تصویر انتخاب می شود و هر یک از رنگ های تصویر به نزدیک ترین رنگ از k رنگ تغییر پیدا می کند. در حالت معمولی برای ذخیره سازی تصویر به $rows \times columns \times 3 \times 8$ بیت نیاز است (هر پیکسل ۳ کانال رنگی دارد که هر رنگ با ۸ بیت نمایش داده می شود). در حالی که با استفاده از **KMeans** می توان تصویر را با اندازه $rows \times columns \times \log_2 k + k \times 3 \times 8$ بیت فشرده کرد. استفاده از الگوریتم **KMeans**تصویر همراه تمرین را فشرده کنید و تصویر حاصل را در گزارش خود بیاورید. برای انتخاب k مناسب از **Elbow Method** استفاده می کنیم. در این روش خطای **MSE** میان تصویر فشرده شده و تصویر اصلی را برای k های مختلف رسم می کنیم. نقطه شکستگی نمودار را به عنوان k مناسب انتخاب می کنیم.