سلام بر دانشجویان عزیز، چند نکته مهم:

۱. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.

۲. 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.

۳. کدهای ارسال شده **بدون گزارش** فاقد نمره می‌باشند.

۴. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.

۵. 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.

۶. از بین سوالات شبیه سازی حتما به هر دو مورد پاسخ داده شود.

۷. نمره تمرین ۱۰۰ نمره می‌باشد و حداکثر تا نمره ۱۱۰ (۱۰ نمره امتیازی) می توانید کسب کنید.

۸. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین صفر خواهد شد.

۹. در صورت داشتن سوال، از طریق ایمیل های زیر سوال خود را مطرح کنید.

سوال ۱، ۲، ۵: hbn.yasaman@gmail.com و jvseraj@gmail.com

سوال ۳، ۴، ۶، ۷ : Ehsan.karamii۹۷@gmail.com و erfanasgari۲۱@gmail.com

سوال ۱: (۱۰ نمره)

دو دیتاست دو بعدی زیر را در نظر بگیرید:

$$X1 = (x1, x2) = \{(4,1), (2,4), (2,3), (3,6), (4,4)\}$$

$$X2 = (x1, x2) = \{(9,10), (6,8), (9,5), (8,7), (10,8)$$

الف) ماتریس پراکندگی درون گروهی(S_w) را محاسبه کنید.

ب) ماتریس پراکندگی بین گروهی(S_B) را محاسبه کنید.

ج) بزرگترین مقدار eigen value را بیابید.

سوال ۲: (۱۰ نمره)

الف) تفاوت model selection و model assessment را بیان کنید. دلیل به کارگیری از model selection چیست؟

ب) الگوریتم‌های model selection به دو دسته Probabilistic و Resampling تقسیم می‌شوند. این دو دسته را با هم مقایسه کنید.

ج) توضیح دهید زمانی که تعداد دادگان کم باشد چه چالش‌هایی در انتخاب مدل وجود دارد و چگونه میتوان مدل را انتخاب و ارزیابی کرد.

سوال ۳: (۱۵ نمره)

برای توزیع مخلوط نمایی زیر

$$p(x) = \alpha\lambda_1 \, e^{-\lambda_1 x} + (1-\alpha)\lambda_2 e^{-\lambda_2 x}$$

با فرض داشتن مجموعه داده‌ی $\{x_1, x_2, ..., x_n\}$ ، ابتدا تابع log-complete likelihood را تشکیل دهید و سپس با استفاده از روش EM روابط بروزرسانی پارامترهای مدل $(\alpha, \lambda_1, \lambda_2)$ را بدست آورید.

سوال ۴: (۱۵ نمره)

فرض کنید می‌خواهیم از یک شبکه‌ی عصبی چند لایه برای تخمین پارامترهای توزیع مخلوط گوسی استفاده کنیم.

الف) توضیح دهید خروجی‌های شبکه به چه صورت در نظر گرفته شوند.

ب) با توجه به مقادیر قابل قبول برای هر یک از پارامترهای توزیع مخلوط گوسی، چه توابع فعالسازی برای خروجی‌های شبکه پیشنهاد می‌کنید.

پ) تابع هزینه‌ی شبکه به چه صورت خواهد بود.

سوال ۵: (شبیه سازی، ۲۰ نمره)

در این سوال قصد داریم که PCA را روی مجموعه داده fashion-MNIST به صورت گام به گام پیاده سازی کنیم.

الف) در ابتدا مجموعه داده را خوانده و یکی از تصاویر مجموعه داده را به دلخواه رسم کنید.

ب) سپس داده ها را استاندارد سازی کنید.

ج) ماتریس کواریانس را محاسبه کرده و ابعاد آن را چاپ کنید.

د) مقادیر ویژه و بردارهای ویژه را محاسبه کرده و مقادیر ویژه از PCA را به ترتیب کاهشی رسم نمایید. سپس بیان نمایید که چگونه میتوان تعداد کامپوننت مناسب را در فرآیند فشرده سازی تشخیص داد؟

ه) با تعداد کامپوننت مناسب تصاویر دیتاست را فشرده کرده و یکی از تصاویر را به دلخواه در حالت قبل از فشرده سازی و بعد از فشرده سازی رسم نمایید.

سوال ۶: (شبیه سازی، ۲۰ نمره)

مجموعه داده‌ی MNIST را در نظر بگیرید.

الف) تنها داده‌های کلاس ۰ و ۱ را در نظر بگیرید. ابتدا هر یک از تصاویر را تبدیل به یک بردار کرده، سپس با استفاده از PCA، بعد این بردارها را از ۷۸۴ به ۲ کاهش دهید. سپس تابع مخلوط گوسی با ۲ جز را بر روی داده‌ها برازش کنید.

ب) اختلاف بین مقادیر میانگین هر کدام از ۲ جز تابع مخلوط گوسی را گزارش کنید (از فاصله‌ی اقلیدسی استفاده کنید).

پ) سپس با اعمال عکس PCA، مقادیر میانگین هر کدام از ۲ جز تابع مخلوط گوسی را از ۲ بعد به فضای ۷۸۴ بعدی برگردانده و به صورت تصویر نمایش دهید. نتیجه‌ی بدست آمده را تحلیل کنید.

ت)۲ مورد از نمونه‌هایی که احتمال تعلقشان به هر یک از ۲ جز تابع مخلوط گوسی برازش شده، کمترین تفاوت را دارند پیدا کنید. سپس با اعمال عکس PCA، نمونه‌ها را از ۲ بعد به فضای ۷۸۴ بعدی برگردانده و به صورت تصویر نمایش دهید. نتیجه‌ی بدست آمده را تحلیل کنید.

ث) مشابه با بخش الف، عملیات برازش تابع مخلوط گوسی با ۲ جز بر روی داده‌های تمام جفت کلاس‌های غیرهمسان ممکن دیگر انجام دهید. سپس جفت کلاسهای غیرهمسان که بیشترین و کمترین اختلاف بین میانگینهای تابع مخلوط گوسی بدست آمده شان وجود دارد را گزارش کنید. نتیجه‌ی بدست آمده را تحلیل کنید.

سوال۷: (شبیه سازی، ۲۰ نمره)

داده مشتریان یک فروشگاه را که در فایل customers_dataset.csv ضمیمه شده است در نظر بگیرید. در این سوال می‌خواهیم مشتریان این فروشگاه را بر اساس جنسیت، سن، میانگین درآمد و امتیاز خریدشان خوشه بندی کنیم تا بتوانیم برای هرکدام از خوشه ها، بسته های تخفیفی مناسبی طراحی کنیم.

الف) در مورد روش های زیر برای تعیین تعداد خوشه مناسب در یک مسئله خوشه بندی تحقیق کنید و هرکدام را توضیح دهید:

۱- K-means Distortion و تحلیل ELBOW روی آن

۲- Silhouette Score

۳- Davies-Bouldin Index

۴- Calinski-Harabasz Index

۵- Dunn Index

ب) برای تعداد ۲ تا ۱۰ خوشه، الگوریتم k-means را بر روی دیتاست فراهم شده اجرا کنید و با رسم نمودار هرکدام از روش های بخش قبل، تعداد مناسب خوشه را بیابید.

ج) با توجه به اینکه هرکدام از داده ها یک بردار ۴ تایی می‌باشد، امکان نمایش آنها در صفحه وجود ندارد. در مورد روش های مورد استفاده برای نمایش داده های با ابعاد بالا تحقیق کنید و برای حداقل دو روش، نمودار خوشه بندی داده‌ها را فقط برای تعداد مناسب خوشه‌ها رسم کنید. هرکدام از خوشه ها باید با رنگ متفاوتی نمایش داده شوند. سعی کنید از روش هایی استفاده نمایید که تحلیل تاثیر هرکدام از ویژگی‌ها در خوشه بندی را ممکن می‌سازد.