# تمارین دیتا ساینس

## هفته چهارم

پاییز ۱۴۰۴

**قبل از این که به بحث شیرین تمارین برسیم، لازمه که چند تا نکته رو در رابطه با تمارین بدونید:**

- در این تمرین، تعدادی سـوال در اختیار شـما قرار گرفته است.
- اولویت اول ما و شما، یادگیری شماست. پس تلاش کنید به جای حل سوالاتی که چالشی برای شما ندارد به سراغ سوالات چالشی‌تر برای خودتان بروید تا بتوانید یادگیری خودتان را افزایش دهید.
- **زمان تحویل جمعه ۱۸ مهر ماه میباشد.(ساعت ۹ شب)**
- ددلاین تمرین به هیچ عنوان تغییر نخواهد کرد.
- گروه بندی در انتهای تمرین انجام شده است.
- تمرینا رو برای منتور(نوا) ارسال کنید. (فایل های ارسالی رو با اسم خودتون بفرستید.)

Telegram: @n80a90zh94

- در صورت داشتن سوال در رابطه با تمرین یا نیاز به راهنمایی، می توانید به آدرس تلگرام زیر پیام بدهید.

Telegram: @mahdavian_a
Telegram: @n80a90zh94
Telegram: @alistelegramid

## سوالات مفهومی:

۱. سؤالات Unsupervised:

- سوالات اجباری:(حداقل ۵ سوال حل کنید.)
  - چگونه مقادیر گمشده (Missing Values)را در زمینه یادگیری بدون‌ناظر مدیریت می‌کنید؟
  - چرا در کشف ناهنجاری، معیارهایی مانند Precisionو Recall از Accuracy مهم‌تر هستند؟
  - چگونه می‌توانید عملکرد یک الگوریتم خوشه‌بندی را ارزیابی کنید؟
  - دو تکنیک را برای انتخاب تعداد خوشه‌های مناسب هنگام استفاده از K-Means توضیح دهید. چرا در بعضی داده‌ها انتخاب تعداد خوشه بهینه (k) دشوار است؟ چه عواملی باعث می‌شوند روش‌هایی مانند Elbow یاSilhouette نتیجه‌ی مبهم بدهند؟ مفهوم شاخص‌های اعتبار خوشه (Cluster Validity Indices)را توضیح دهید.
  - در یک دیتاست بانکی، تراکنش‌های تقلبی بسیار کمیاب هستند. چرا در این سناریو استفاده از کشف ناهنجاری بهتر از یک مدل طبقه‌بندی معمولی است؟
  - چه تفاوتی بین ناهنجاری (Anomaly) و نویز (Noise) وجود دارد؟ آیا همیشه باید نویز را به عنوان ناهنجاری در نظر گرفت؟
  - انتشار برچسب (Label Propagation) چیست؟ چرا آن را پیاده‌سازی می‌کنیم و چگونه؟
  - الگوریتم (t-SNE) t-Distributed Stochastic Neighbor Embeddingرا توصیف کنید و کاربردهای آن را توضیح دهید.
- سوالات اختیاری:
  - مراحلی را که برای مقیاس‌بندی (Scaling) و نرمال‌سازی (Normalization) داده‌ها پیش از خوشه‌بندی طی می‌کنید، توضیح دهید.
  - اگر داده‌هایتان شامل ویژگی‌های بسیار زیاد (High-dimensional) باشد، چرا خوشه‌بندی مستقیم روی داده خام ممکن است نتایج ضعیفی بدهد؟ چه تکنیک‌هایی برای رفع این مشکل وجود دارد؟ منظور از نفرین بُعد (Curse of Dimensionality)چیست و چگونه بر مدل‌های یادگیری ماشین تأثیر می‌گذارد؟
  - اهمیت انتخاب ویژگی‌ها (Feature Selection) در یادگیری بدون‌ناظر را توضیح دهید.
  - در DBSCAN مفهوم چگالی (Density) چه نقشی در شناسایی خوشه‌ها و نقاط ناهنجار دارد؟
  - در چه شرایطی استفاده از (GMM) Gaussian Mixture Modelمی‌تواند بهتر از K-Meansباشد؟

۲. سؤالات SVM:(همه سوالات اجباری هستند.)

- ایده‌ی اصلی پشت ماشین‌های بردار پشتیبان (Support Vector Machines) چیست؟و بردار پشتیبان (Support Vector)چیست؟
- چرا هنگام استفاده از SVM مهم است که ورودی‌ها را مقیاس‌بندی (Scaling) کنیم؟
- فرض کنید یک دسته‌بند SVM با کرنل RBF آموزش داده‌اید، اما به نظر می‌رسد که روی داده‌ی آموزشی دچار Underfitting شده است. آیا باید مقدار γ (گاما) را افزایش دهید یا کاهش؟ در مورد C چه‌طور؟

## سوالات پیاده سازی:

- سوالات اجباری:(این قسمت گروهی است.)

۱. کلاسترینگ روی دیتاست MNIST_CSV.zip

دیتاست : دیتاست MNIST یکی از معروف‌ترین و پرکاربردترین مجموعه‌داده‌ها در یادگیری ماشین است. این دیتاست معمولاً برای آموزش و ارزیابی الگوریتم‌های تشخیص دست‌نوشته مورد استفاده قرار می‌گیرد.

- نوع داده‌ها: شامل تصاویر سیاه و سفید از ارقام دست‌نویس (۰ تا ۹). هر تصویر یک رقم واحد است.
- ابعاد تصاویر: هر تصویر دارای اندازه‌ی ۲۸×۲۸ پیکسل است. بنابراین هر تصویر ۷۸۴ ویژگی (pixel features)دارد.
- مقیاس خاکستری(Grayscale) : مقادیر پیکسل‌ها بین ۰ تا ۲۵۵ هستند.

مراحل:

دیتاست را دانلود کنید و روی آن پیش‌پردازش انجام دهید.(نرمال‌سازی و کاهش ابعاد و ...)

اجرای الگوریتم خوشه‌بندی:

۱. *خوشه‌بندی با K-Means(اختیاری)*

- اجرای K-Means
- ارزیابی نتایج
- بررسی اعداد مشابه که با هم اشتباه گرفته شده‌اند.

۲. *خوشه‌بندی با DBSCAN(اجباری)*

- انتخاب مقادیر مناسب برای پارامترهای eps و min_samples.
- اجرای DBSCAN و بررسی تعداد خوشه‌های پیدا شده.
- ارزیابی نتایج

۳. *خوشه‌بندی با Gaussian Mixture Models (GMM)(اجباری)*

- اجرای GMM
- استفاده از معیارهای AIC و BIC برای انتخاب تعداد خوشه بهینه.
- ارزیابی نتایج

۴. *تحلیل و مقایسه*

- نتایج سه الگوریتم را مقایسه کنید.
- کدام الگوریتم برای MNIST عملکرد بهتری دارد و چرا؟
- در کدام حالت خوشه‌ها بهتر با برچسب‌های واقعی هم‌پوشانی دارند؟

۲. تشخیص آنومالی روی دیتاست mnist_with_anomalies.zip

دیتاست: دیتاست این قسمت بخشی از دیتاست MNIST می باشد. که شامل تمامی صفر ها(حدود ۷۰۰۰ تا) و ۶۰ نمونه آنومالی می باشد. شما باید بدون استفاده از لیبل های آنومالی ، آنومالی ها را تشخیص دهید. (۲۶ نمونه از ۶۰ آنومالی به شما داده خواهد شد.)

- از PCA (اجباری) و Autoencoder(اختیاری) استفاده کنید تا آنومالی ها را تشخیص دهید.
- نمودارها و تحلیل های ارائه دهید که نشان دهد چگونه نمونه های آنومالی از سایر داده متمایز شده اند.
- از شاخص هایی مانند خطای بازسازی، نمرات احتمال، یا ویژگی های فضای نهان برای تحلیل بهره ببرید.
- پس از شناسایی آنومالی ها، بررسی کنید که چه تعداد از نمونه های شناسایی شده با لیست اولیه آنومالی (۲۶ نمونه ای که از ۶۰ نمونه به شما داده شده است.) مطابقت دارند.
- در PCAواریانس ۹۰و ۹۵ و۹۹ درصد را هم بررسی کنید.
- این تمرین باید با روش های بدون نظارت انجام شود. از ایندکس های آنومالی تنها برای ارزیابی و تحلیل استفاده کنید، نه برای یادگیری مدل.
- روش های مختلف را آزمایش کرده و مقایسه ای از عملکرد آنها ارائه دهید. سپس گزارش کارهایی که انجام داده اید را آماده کنید.
- در صورت نیاز از تکنیک های پیش پردازش داده مانند نرمال سازی یا کاهش ابعاد استفاده کنید.
- نمونه لیست ایندکس های آنومالی:

anomaly = [81,143 ,6857, 6535, 6235, 6229, 6065, 3295, 3203, 3093, 2781, 2765, 2726 ,2658, 2609, 2597,863,790, 712,701,663, 401, 352,329,223 , 2704]

- سوال اختیاری:
  - روی تمرین گروهی هفته قبل SVM را پیاده سازی کنید و خروجی آن را تحلیل کنید.

<u>گروه بندی:</u>

گروه ۱:لیلی و علی

گروه ۲:جواد و نازنین

گروه ۳:داوود و پرین و آرش

گروه ۴:نگین و محمدامین

گروه ۵: فاضل و سهیل

KELAASOR