دانشگاه تربیت مدرس

دانشکده مهندسی برق و کامپیوتر

تمرین اول درس یادگیری ماشین

مبحث کاهش ابعاد و خوشه‌بندی

استاد درس:

دکتر فؤاد قادری

نیمسال اول ۱۴۰۴-۱۴۰۳

مقدمه

- ✓ فایل تحویلی شما، یک فایل زیپ شده‌ی نهایی شامل گزارش کار (فایل pdf و word) و فایل کد پایتون با پسوند ipynb (Jupiter Notebook) می‌باشد. لطفا آن را به صورت زیر نام‌گذاری و ارسال نمایید.

  HW1_[LastName]_[FirstName]

- ✓ گزارش کار خود را در یک فایل pdf و word تحویل دهید و از گذاشتن صرفا اسکرین شات‌های پشت سرهم از کد در گزارش کار خودداری کنید.
- ✓ توجه داشته باشید که در فایل ارسالی پایتون، خروجی هر سلول (شامل نمودار، خروجی عددی و غیره) حتما ذخیره شده و قابل مشاهده باشد.
- ✓ لازم است حتما نتایج بدست آمده را گزارش و تحلیل کنید.
- ✓ علاوه بر مهارت حل سوالات، نوشتن پاسخ مینی پروژه‌ها در فرمت گزارش فنی (فصل بندی و صفحه بندی مناسب، رعایت اصول نگارش، درج زیرنویس برای شکل‌ها و بالانویس برای جداول و اشاره به شماره شکل یا جدول در متن و ...) برای دانشجویان تحصیلات تکمیلی اهمیت دارد، این مورد نیز در ارزشیابی لحاظ می‌شود.
- ✓ در صورت فراموشی در ارسال کد پایتون، هیچ نمره‌ای به شما تعلق نخواهد گرفت.
- ✓ در صورت مشاهده تشابه در هر بخش از انجام پروژه، نمره هر دو نفر صفر لحاظ می‌گردد.

- ✓ ایمیل دستیار طراح: mahtabmirzaee999@gmail.com

هدف از انجام این مینی‌پروژه اعمال الگوریتم‌های کاهش بعد و سپس خوشه‌بندی بر روی داده‌های کاهش یافته و ارزیابی الگوریتم خوشه‌بند به کمک معیار Silhouette است.

- **کاهش ابعاد**

برای کاهش ابعاد قصد داریم از دو الگوریتم PCA و LDA استفاده کنیم. در جدول زیر به طور خلاصه تفاوت‌های این دو الگوریتم آورده شده است.

Comparison Criteria	**LDA**	**PCA**
Type	Supervised	Unsupervised
Goal	Maximizes class separability	Maximizes variance in the data
Output	Linear discriminants (1 to C-1 components)	Principal components (up to N components)
Feature Dependency	Depends on labeled data for classes	Independent of labels; only uses feature variance

در الگوریتم کاهش بعد PCA برای پیدا کردن تعداد مناسبِ مولفه‌های اصلی می‌توان از قانون Kaiser در نمودار Scree استفاده کرد. برای مطالعه بیشتر می‌توانید به این منبع مراجعه نمایید.

- **خوشه‌بندی**

برای عمل خوشه‌بندی از دو الگوریتم k-means و Hierarchical Clustering استفاده می‌کنیم. در جدول زیر به طور خلاصه این دو الگوریتم مقایسه شده‌اند.

Comparison Criteria	**K-Means Clustering**	**Hierarchical Clustering**
Number of Clusters	Must be specified in advance	Determined from the **dendrogram**
Method	Partitions data into k clusters	Builds a hierarchy of nested clusters
Distance Metric	Typically Euclidean	Various metrics available (Euclidean, Manhattan, etc.)
Scalability	Efficient for large datasets	Computationally intensive for large datasets

نمودار dendrogram نوعی نمودار درختی است که در خوشه‌بندی سلسله‌مراتبی استفاده می‌شود. این نمودار ساختار سلسله‌مراتبی خوشه‌ها را به شکل یک درخت نشان می‌دهد و کمک می‌کند که روابط بین داده‌ها و چگونگی تجمیع خوشه‌ها را در سطوح مختلف مشاهده کنیم. برای مطالعه بیشتر و نحوه تفسیر این نمودار می‌توانید به این منبع مراجعه نمایید.

- **معیار ارزیابی**

برای ارزیابی کیفیت خوشه‌ها از معیار درونی Silhouette Score استفاده می‌کنیم. این معیار نشان می‌دهد که هر نقطه داده[1] در مقایسه با سایر خوشه‌ها (separation) تا چه میزان به خوشه خود (cohesion) شباهت دارد. امتیاز Silhouette بالاتر نشان‌دهنده آن است که خوشه‌ها به خوبی از یکدیگر جدا شده‌اند و نقاط به صورت فشرده درون خوشه‌های خود قرار دارند.

برای هر نقطه داده، این معیار به صورت زیر محاسبه می‌شود:

Cohesion: فاصله میانگین بین یک نقطه و تمامی نقاط دیگر در همان خوشه

Separation: فاصله میانگین بین یک نقطه و تمامی نقاط در نزدیک‌ترین خوشه همسایه (یعنی نزدیک‌ترین خوشه‌ای که نقطه به آن تعلق ندارد).

$$\text{Silhouette Score} = \frac{\text{Separation} - \text{Cohesion}}{\max(\text{Separation}. \text{Cohesion})}$$

هرچقدر که این مقدار به عدد ۱ نزدیک‌تر شود به این معنی است که نقطه‌داده به خوبی به خوشه خود تعلق دارد و از خوشه‌های دیگر دور است، در حالی که نزدیک شدن این مقدار به عدد ۱- به این معناست که ممکن است نقطه به خوشه نادرستی اختصاص داده شده باشد.

[1] Data point

مینی پروژه اول: کاهش بعد و خوشه بندی

- **گام اول: مجموعه داده**

مجموعه داده Ionosphere را در نظر بگیرید. این مجموعه داده شامل ۳۵۱ نمونه و ۳۴ ویژگی است و شامل سیگنال‌های راداری است که توسط آرایه‌ای از آنتن‌های پرقدرت جمع‌آوری شده و به بررسی ساختار یونوسفر[1] می‌پردازد. سیگنال‌های دریافتی، بسته به وجود ساختار در یونوسفر، به دو دسته خوب (مقدار g) و بد (مقدار b) طبقه‌بندی شده‌اند و هر نمونه با ویژگی‌های مختلطی که حاصل همبستگی سیگنال‌ها هستند، توصیف می‌شود.

برای دسترسی به این مجموعه داده می‌توانید یا از وبسایت داده شده آن را دانلود کنید و یا کد زیر را اجرا کنید:

```
# Install the ucimlrepo package dataset
!pip install ucimlrepo
```

```
# Import the dataset into your code

from ucimlrepo import fetch_ucirepo

# fetch dataset
ionosphere = fetch_ucirepo(id=52)

# data (as pandas dataframes)
X = ionosphere.data.features
y = ionosphere.data.targets

# metadata
print(ionosphere.metadata)

# variable information
print(ionosphere.variables)
```

توجه: قبل از اعمال الگوریتم‌های مربوط به کاهش بعد، لازم است در صورت لزوم بر روی داده، پیش‌پردازش (مانند نرمال سازی، داده‌های پرت[2] و مقادیر گم‌شده[3] و ...) انجام شود.

[1] یونوسفر (Ionosphere) یکی از لایه‌های بالایی جو زمین است و دارای ذرات باردار الکتریکی می‌باشد. این لایه نقش مهمی در انتقال امواج رادیویی ایفا می‌کند، زیرا می‌تواند امواج رادیویی را منعکس یا جذب کند.

[2] Outliers

[3] Missing values

- **گام دوم: کاهش بعد**

الف) ابتدا نمودار Scree را بر روی تمام مولفه‌های اصلی[1] بدست آمده از الگوریتم PCA اعمال کرده و نمودار حاصل و نیز تعداد مولفه‌های اصلی طبق قانون Kaiser را گزارش نمایید. در نهایت عددی را برای تعداد نهاییِ مولفه‌های اصلی انتخاب کنید که بین ۷۰ تا ۹۰ درصد واریانس کل داده‌ها حفظ شود. این عدد و نیز واریانس تجمعی را گزارش نمایید.

ب) الگوریتم LDA را بر روی داده‌های پیش‌پردازش شده اعمال کنید. نتایج را گزارش و تحلیل نمایید.

- **گام سوم: خوشه‌بندی**

الف) در این مرحله، الگوریتم خوشه‌بندی K-means را بر روی هر دو مجموعه داده کاهش‌داده شده از مرحله قبل اعمال کنید. تعداد مراکز خوشه‌ها را برابر $k = [2, 3, 4, 5, 6]$ گرفته و به ازای هر k، عملکرد خوشه‌بند را با معیار Silhouette Score سنجیده و این مقادیر را گزارش کنید.

در نهایت با توجه به معیار ارزیابی Silhouette Score، بهترین مقدار k در هر مجموعه داده را گزارش نمایید.

داده خوشه‌بندی شده را رسم کرده و نتایج را گزارش و تحلیل نمایید.

ب) ابتدا نمودار dendrogram را برای هر دو مجموعه داده کاهش‌داده شده رسم کنید و سپس مشابه مراحل بالا، الگوریتم خوشه‌بندی سلسله مراتبی (از پایین به بالا) را بر روی هر دوی این مجموعه داده اعمال کنید. تعداد مراکز خوشه‌ها را برابر با $k = [2, 3, 4, 5, 6]$ گرفته و به ازای هر k، عملکرد خوشه‌بند را با معیار Silhouette Score سنجیده و این مقادیر را گزارش کنید.

در نهایت با توجه به معیار ارزیابی Silhouette Score، بهترین مقدار k در هر مجموعه داده را گزارش نمایید. آیا این تعداد خوشه با نمودار dendrogram حاصل هم‌خوانی دارد؟

داده خوشه‌بندی شده را رسم کرده و نتایج را گزارش و تحلیل نمایید.

- **گام چهارم: بحث و نتیجه‌گیری**

نتایج بدست آمده از هر الگوریتم خوشه‌بندی و به ازای هر یک از دو الگوریتم کاهش بعد را در یک جدول گزارش کرده و با یکدیگر مقایسه کنید. مشخص کنید در کدام مورد بهترین نتیجه حاصل شده است و علت آن چیست؟

[1] Principal Component