مبانی هوش محاسباتی
دانشگاه فردوسی مشهد
گروه مهندسی کامپیوتر

پروژه اول
مهلت تحویل 1403/08/2

هدف پروژه

در این پروژه از شما می‌خواهیم تا تصاویر افراد را بر اساس ویژگی‌هایی که در بخش‌های جلوتر توضیح داده می‌شود، خوشه‌بندی کنید تا در نهایت افراد با ویژگی‌های مشابه در یک خوشه قرار گیرند.

دیتاست

این دیتاست شامل 50000 تصویر از افراد معروف (Celebrity) می‌شود که به همراه یک فایل csv در سامانه ویو قرار گرفته است. هر سطر در این فایل csv به یک تصویر اختصاص داده شده و شامل یک یا چند ویژگی مختلف چهره می‌باشد که به صورت باینری (-1 یا 1) کدگذاری شده‌اند.
همچنین یک دیتاست **تست** شامل 100 تصویر دیگر در اختیار شما قرار گرفته است.

	A	B	C	D	E	F	G	H	I	J
1	image_id	5_o_Clock_Shadow	Arched_Eyebrows	Attractive	Bags_Under_Eyes	Bald	Bangs	Big_Lips	Big_Nose	Black_Hair
2	000001.jpg	-1	1	1	-1	-1	-1	-1	-1	-1
3	000002.jpg	-1	-1	-1	1	-1	-1	-1	1	-1
4	000003.jpg	-1	-1	-1	-1	-1	-1	1	-1	-1
5	000004.jpg	-1	-1	1	-1	-1	-1	-1	-1	-1
6	000005.jpg	-1	1	1	-1	-1	-1	1	-1	-1
7	000006.jpg	-1	1	1	-1	-1	-1	1	-1	-1
8	000007.jpg	1	-1	1	1	-1	-1	1	1	1
9	000008.jpg	1	1	-1	1	-1	-1	1	-1	1
10	000009.jpg	-1	1	1	-1	-1	1	1	-1	-1
11	000010.jpg	-1	-1	1	-1	-1	-1	-1	-1	-1

فاز اول: Feature Extraction

در این فاز باید با استفاده از کتابخانه های ذکر شده در کلاس، مختصات صورت افراد را استخراج کرده و با روش های توضیح داده شده در کلاس حل تمرین، تقریبی از **رنگ چشم** و **رنگ پوست** هر عکس را به دست آورید.

- پس از استخراج این دو ویژگی، برای آسان تر شدن کار می‌توانید آن‌ها را به فایل csv دیتاست اضافه کنید تا دیگر مجبور نباشید این بخش از پروژه را اجرا کنید.
- فقط برای استخراج مختصات صورت (face) می‌توانید از کتابخانه‌ها استفاده کنید. پیدا کردن رنگ چشم و پوست باید توسط خودتان پیاده‌سازی شود.

فاز دوم: Feature Selection

در این فاز شما باید از بین ویژگی‌های موجود، حداقل 6 ویژگی را انتخاب کنید تا بر اساس آن ویژگی‌ها خوشه‌بندی را انجام دهید. برای این کار باید همبستگی (correlation) بین تک تک ویژگی‌ها را حساب کرده و سپس یک correlation matrix ایجاد کنید. پس از آن باید ویژگی‌ها را به گونه‌ای انتخاب کنید که خوشه بندی بهتری داشته باشید.(توضیحات بیشتر در کلاس حل تمرین)

- محاسبه correlation بر عهدهٔ خودتان است و استفاده از کتابخانه‌ها و توابع عمومی مجاز نیست.
- دو ویژگی رنگ چشم و رنگ پوست (که در فاز قبل از تصاویر استخراج کردید) ممکن است بر اساس آستانه‌ای (threshold) که در نظر گرفته‌اید انتخاب نشوند و برای فازهای بعد باید به صورت دستی آن‌ها را اضافه کنید.
- با توجه به حداقل تعداد ویژگی ای که قید شد، بهینه‌ترین threshold را انتخاب کنید.

فاز سوم: Clustering

در این فاز باید با استفاده از سه الگوریتم KMeans و DBSCAN و MeanShift دیتاست موجود را بر اساس ویژگی‌های انتخاب شده خوشه بندی کنید.

- برای پارامترهای هر الگوریتم باید hyperparameter tuning انجام دهید و بهترین مقادیر را انتخاب کنید.
- یک معیار ارزیابی برای خوشه‌بندی خود پیدا کنید و با آن مقدار عددی میزان کارآمد بودن الگوریتم‌ها را با هم مقایسه کنید.
- این کار را یک‌بار بدون دو ویژگی رنگ پوست و رنگ چشم، و یک‌بار با وجود آن‌ها انجام دهید و تاثیر این دو ویژگی در خوشه بندی را تحلیل کنید (فقط برای الگوریتم KMeans کافی است).
- تصاویر مربوط به 10 تصویر از هر خوشه را در یک فولدر سیو کنید و خروجی‌ها را بررسی و تحلیل کنید (هر خوشه در یک فولدر مجزا).
- جهت مقایسهٔ ویژگی‌های متمایز کننده هر خوشه و پیدا کردن ویژگی‌های بارز آن‌ها، یک heatmap باتوجه به اعضای خوشه‌ها plot کنید.

فاز چهارم: Visualization

در این بخش ابتدا با استفاده از کتابخانه‌های ذکر شده در کلاس حل تمرین ابعاد دیتا را کاهش داده و سپس نتیجه خوشه‌بندی‌ها را visualize کنید و نتایج این خوشه‌بندی‌ها را در داکیومنت خود تحلیل کنید.

فاز پنجم: KNN و K-Means

در این فاز ابتدا مراکز خوشه‌های الگوریتم KMeans را پیدا کرده و سپس با استفاده از الگوریتم KNN که در کلاس حل تمرین توضیح داده شد، 50 داده نزدیک هر مرکز خوشه را پیدا کنید. سپس بررسی کنید بر اساس خوشه‌بندی‌ای که با KMeans انجام دادید، آیا این داده‌ها مربوط به همان مرکز خوشه هستند یا خیر؟

- این کار را با 3000 داده نزدیک هر مرکز کلاستر هم بررسی کنید.
- بررسی کنید که داده‌هایی که مربوط به آن مرکز خوشه نبوده‌اند مربوط به کدام خوشه هستند و علت این اتفاق چه بوده است؟

فاز ششم: Prediction

در این فاز با استفاده از داده‌هایی که به عنوان داده تست در اختیار شما قرار می‌گیرد بررسی کنید که این داده‌ها می‌تواند مربوط به کدام خوشه باشد؟ (با توجه به توضیحات کلاس حل تمرین)

- برای 10 دادهٔ تست نتایج را visualize کنید. برای هرکدام از این 10 داده، 5 داده از خوشهٔ مرتبط با آن نیز visual شود. (داده‌های تست را به گونه‌ای انتخاب کنید که حتما از هر خوشه یک داده بررسی شود)
- یک ستون برای خوشه‌ای که هر تصویر به آن تعلق گرفته (لیبل آن) به فایل csv تست اضافه کنید.
- heatmap بررسی شده در فاز سوم را برای این فاز نیز plot کرده و نتایج را باهم مقایسه و تحلیل کنید.

نکات و توضیحات تکمیلی

- انجام پروژه می‌تواند در قالب گروه‌های دو نفره و یا به صورت انفرادی صورت گیرد.
- علاوه بر سورس کد پروژه، فایل **مستندات** نیز باید آپلود شود.
- نام اعضای گروه در فایل مستندات ذکر شود و فقط یکی از اعضا پروژه را آپلود کند.
- هر گونه شباهت نامتعارف بین کد شما و کد سایر گروه ها تقلب محسوب می‌شود و نمره‌ای برای این پروژه دریافت نخواهید کرد.
- در صورت نوشتن داکیومنت تمیز (برای مثال با LATEX) نمره اضافه برای شما در نظر گرفته خواهد شد.
- فایل شامل سورس کد پروژه و مستندات را در قالب فایل zip و با نام شماره دانشجویی خود ذخیره و ارسال نمایید.
- در صورت داشتن هرگونه سوال می‌توانید با kourosh_hsz و یا fatemeh_dehbashii در ارتباط باشید و یا در گروه درسی مطرح نمایید.

موفق باشید؛ تیم حل تمرین