

سوال اول

یکی از کاربردهای خوشه‌بندی که می‌توان به آن اشاره کرد، کم‌حجم‌سازی تصاویر است. در این سوال قرار است با پیاده‌سازی این روش بتوانیم با دریافت تصویر ورودی، تصویر خروجی که معادل کم‌حجم تصویر ورودی است را نمایش دهیم. برای این کار ابتدا می‌بایست تمامی مراحل مربوط به خوشه‌بندی با استفاده از روش K-means و پیدا کردن centroidها در هر مرحله را پیاده‌سازی کرده و سپس با استفاده از توابعی که در اختیار داریم، مرحله کم‌حجم‌سازی تصویر را پیاده‌سازی کنیم. توجه داشته باشید که تنها با استفاده از کتابخانه‌هایی که در اختیار شما قرار گرفته است می‌توانید تمامی این الگوریتم را پیاده‌سازی کنید. فایل مربوط به این سوال به اسم ImageCompression.ipynb است.

- در این بخش می‌بایست که توابعی که در اختیار شما قرار گرفته است با توجه به عملکرد آن‌ها کامل کنید. دقت داشته باشید که به کدهایی که از قبل کامل شده است دست نزده و آن‌ها را تا حد امکان تغییر ندهید. در هر بخش با استفاده از کامنت‌ها توضیحات مربوط به بخشی که می‌بایست کامل کنید، داده شده است.
- در این بخش می‌خواهیم با استفاده از توابعی که در بخش قبلی آماده‌سازی شده‌اند به سراغ چند مجموعه داده برویم. روش K-means ای که قبل‌تر آماده کرده بودید را بر روی مجموعه دادگان در اختیار قرار گرفته و یک مجموعه دادگان تصادفی امتحان کرده و خروجی‌ها را مشاهده نمایید. دقت کنید که لازم نیست بخش مجموعه دادگان تصادفی را تغییر دهید.
- حال در این بخش می‌خواهیم که کم‌حجم‌سازی را بر روی یک تصویر دلخواه مشاهده کنیم. یک تصویر به دلخواه خود انتخاب کرده و کد را کامل کنید. خروجی را گزارش کرده و با توجه به تجربه‌هایی که از کم‌حجم کردن تصاویر داشته‌اید، خروجی را توصیف کنید.

سوال دوم

در این بخش می‌خواهیم به پیاده‌سازی الگوریتم DBSCAN که یکی از الگوریتم‌های معروف در خوشه‌بندی است، بپردازیم. در این بخش از مجموعه دادگانی که در لینک زیر هست استفاده می‌شود.

https://reneshbedre.github.io/assets/posts/tsne/tsne_scores.csv

همان‌طور که بعد از خواندن مجموعه دادگان متوجه می‌شوید، این مجموعه دارای دو ویژگی و ۴۴۰۶ سطر است. حال برای پیاده‌سازی این الگوریتم می‌بایست مقادیر حداقل نقاط و اپسیلون را مشخص کنیم. در بسیاری از حالت‌ها برای مساله حداقل نقاط می‌بایست که از دو برابر ابعاد داده استفاده کرد. همچنین برای تعیین اپسیلون نیز می‌توان از روش kNN استفاده کرد. نقطه‌ی زانویی شکل را می‌توان به عنوان مقدار اپسیلون پذیرفت. برای این کار می‌توانید از تابع KneeLocator() از کتابخانه kneed استفاده کنید. سپس با استفاده از کتابخانه

sklearn و تابع DBSCAN در آن به سراغ خوشه‌بندی بروید. در نهایت نمودار مربوط به خوشه‌ها را در فضای ابعاد نمایش دهید.

سوال سوم

در این سوال قصد داریم از روش خوشه‌بندی سلسله‌مراتبی بر روی مجموعه داده‌های IRIS استفاده کنیم. فایلی که در این سوال با آن قرار است به مفاهیم بپردازیم به اسم `HierarchicalClustering.ipynb` است. در ابتدا روش سلسله‌مراتبی و مفاهیم مربوط به آن آمده است. سپس در بخش بعدی که می‌بایست پیاده‌سازی را انجام دهید، ابتدا می‌بایست مجموعه داده‌های IRIS که برای گل‌های متفاوت است را دریافت کرده و سراغ مراحل بعد بروید. در هر بخش با استفاده از جاهای مشخص‌شده کد را کامل کنید. در انتها با استفاده از مجموعه داده‌ها، مقدار `silhouette score` را محاسبه کرده و نمودار نمایش هر کلاس را در این خوشه‌بندی به صورت `scatter plot` نمایش دهید که در آن هر کدام از سه کلاس مذکور با رنگ مجزا قابل تشخیص باشند.

سوال چهارم

- در این سوال قصد پیاده‌سازی روش GMM را برای خوشه‌بندی را داریم. ابتدا سه توزیع گوسی دو بعدی در فایلی `GMM.ipynb` برای شما تولید شده است. حال مراحل زیر را بر روی این مجموعه داده پیاده‌سازی کنید.
- با توجه به نمودار `scatter` نمایش داده شده و انتخاب تعداد خوشه مناسب، از روش Kmeans با استفاده از کتابخانه sklearn استفاده کرده و خروجی را نمایش دهید.
 - حال با استفاده از کتابخانه sklearn و ماژول `mixture` یک مدل `GaussianMixtureModel` بر روی این مجموعه داده سوار کنید.
 - حال ماتریس احتمالات هر نقطه برای اینکه متعلق به هر یک از توابع گوسی باشد را با استفاده از دستور `predict_proba` محاسبه کنید.
 - حال با استفاده از ماتریس احتمالات هر نقطه که در بخش قبل بدست آمد، اندازه دایره هر داده را متناسب با تعلق آن به خوشه متناسب کرده و در نمودار دوبعدی `scatter` آن را اعمال کنید.
 - در مرحله بعد تابع `draw_ellipse` و `plot_gmm` که در اختیار شما قرار گرفته شده است را کامل کرده و با استفاده از آن و تعداد خوشه‌های مناسب، داده‌ها را نمایش دهید و استنباط خود را از این نمودار ارائه کنید.