بخش عملی:

سوال اول)

در این قسمت قصد داریم به بررسی Dimensionality reduction و Clustering بپردازیم. در ابتدا فایل MDA_P1.rar از حالت فشرده خارج کنید و فایل covtype.info جهت آشنایی با دیتاست مطالعه کنید.(توجه کنید که ستون اول صرفا شماره هر سمپل می‌باشد.)

الف) با استفاده از الگوریتم PCA، به رسم واریانس Principal components به استفاده از رابطه$\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{d}\lambda_i}$ بپردازید. در ادامه مقدار k به نحوی انتخاب کنید که حداقل ۹۰ درصد از واریانس سمپل‌ها حفظ شود و در نهایت با استفاده از eigenvector های بدست آمده ابعاد نمونه‌هارا کاهش دهید.(می‌توانید از سایر الگوریتم‌های کاهش بعد جهت بهبود نتیجه قسمت (ب) استفاده کنید.)

ب) در این قسمت داده هارا به سه چانک تقسیم کنید و با استفاده از یکی از الگوریتم هایBRF یا Cure داده‌ها را به ۷ کلاستر گروه‌بندی کنید.(همه موارد طراحی از جمله نحوه نرمالایز کردن داده‌ها، نحوه initialize کردن اولیه کلاسترها و ... در گزارش خود بنویسید.)

ج) مقدار دو متریک زیر را برای ارزیابی کلاسترینگ صورت گرفته محاسبه کنید و بر اساس الگوریتمی که برای کلاسترینگ در قسمت (ب) انتخاب کردید، توضیح دهید کدام یک جهت ارزیابی کلاسترینگ مناسب تر می‌باشد. برای هر متریک توضیح دهید چه معیاری را ارزیابی می‌کند و رنج آن را مشخص کنید.

Silhouette Score:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

$a(i)$: میانگین فاصله سمپل i ام تا سمپل‌های هم‌کلاستر

$b(i)$: میانگین فاصله سمپل i ام تا نقاط نزدیک‌ترین کلاستر

Davies-Bouldin Index:

$$DBI = \frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{dij}\right)$$

σ_i: میانگین فاصله سمپل‌ها در کلاستر i تا مرکز

dij: فاصله بین مرکز کلاستر i و j

K: تعداد کلاستر (۷)

د) در این قسمت، نتایج کلاسترینگ را در ستون‌های فایل Evaluation_P1 ذخیره کنید. (بخشی از نمره این تمرین به نتایج این بخش اختصاص دارد.)