تمرین دوم درس هوش محاسباتی - بهار ۱۴۰۲

در این تمرین شما قصد دارید با خوشه بندی[1] در یک دیتاست تشکیل شده از چند دامین[2]، بدون ناظر دامین های مختلف در دیتاست را جدا کنید. (برای توضیحات بیشتر، به ویس جلسه حل تمرین مراجعه کنید)

برای این منظور، یک دیتاست تشکیل شده از چند دامین مختلف به شما داده شده است. دیتاست داده شده، به این صورت می باشد که به ازای هر تصویر، یک بردار ویژگی استخراج شده، و برچسب آن در اختیار شما قرار داده شده است. این برچسب، نشان دهنده برچسب محتوای تصاویر است و در خصوص برچسب دامین اطلاع خاصی نداریم. بنابراین، تعداد دامین های موجود در دیتاست نیز مشخص نیست. در داده های آموزشی که در اختیار شما قرار گرفته (داده های این فولدر)، بردار های ویژگی استخراج شده از ۲۰هزار تصویر آورده شده، به طوری که بردار ویژگی هر تصویر ۲۰۴۸ بعدی است.

بخش اول - خوشه بندی

در بخش اول، شما باید یک الگوریتم خوشه بندی را استفاده کنید که بر روی داده ها به خوبی عمل کند. انتخاب الگوریتم به عهده شماست، اما باید پس از مقایسه الگوریتم های مختلف خوشه بندی و انتخاب یکی از آنها، نشان دهید که انتخاب مناسبی بوده است. برای این مهم، ابتدا یک راهکار بصری را بررسی کنید (می توانید از روش های کاهش ابعاد[3] مثل t-SNE (که کد نمونه از این روش در ادامه قرار گرفته است) و یا PCA استفاده کنید) که راهی برای نمایش داده های دارای ابعاد بسیار در فضای دو بعدی ارائه می دهد. حال شما، بعد از خوشه بندی، با استفاده از خروجی الگوریتم خوشه بندی، رنگ داده ها را در پلات فضای کاهش یافته مشخص کنید و بعد از آن، بررسی کنید که آیا خوشه بندی با الگوریتم مناسبی انجام شده است؟ و اینکه آیا پارامترهای مناسبی استفاده شده است؟

بخش دوم - پیدا کردن تعداد دامین

پیدا کردن تعداد دامین، یک عمل نسبتا پیچیده است که نحوه انجام آن، وابسته به الگوریتم انتخابی در بخش پیشین خواهد بود. با بررسی پارامتر های مختلف، ابتدا بررسی کنید که الگوریتم شما، برای عملکرد مناسب، به چه تعداد خوشه رسیده است. حال با توجه به روش هایی که در درس یاد گرفته اید، یک راهکار برای استفاده از این تعداد خوشه برای رسیدن به تعداد دامین ها بدهید (می توانید خوشه ها را با هم ترکیب کنید - بر اساس معیار شباهت - و یا هر راهکار های ابتکاری دیگر.) نهایتا خروجی شما در این بخش باید تعداد دامین ها، و اینکه هر خوشه چه دامینی را نشان می دهد باشد. (طبیعتا یعنی اگر دو خوشه هر دو با خروجی الگوریتم شما دامین ۱ دارند، در داده های واقعی هم دامین آنها یکی باشد)

---

[1] Clustering
[2] Domain
[3] Dimensionality Reduction

بخش سوم - ارزیابی با برچسب های دامین ها

در بخش تست، دو دسته فایل قرار دارد. دسته اول، برای بخش کوچکی از داده های آموزشی (۵ درصد) برچسب دامین را مشخص کرده است، و دسته دوم داده های تست است. با استفاده از دسته اول داده ها، کیفیت الگوریتم خود را در تشخیص دامین و با دسته دوم داده ها الگوریتم خوشه بندی خود را ارزیابی کنید.

پاسخ خود را به صورت یک فایل فشرده، شامل فایل پی دی اف گزارش و کد های خود، در قالب گروه های حداکثر دو نفره ارسال کنید.