

## تمرین دوم درس هوش محاسباتی - بهار ۱۴۰۲

در این تمرین، شما قصد دارید با خوشه بندی<sup>۱</sup> در یک دیتاست تشکیل شده از چند دامین<sup>۲</sup>، 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 دیتابیس نیز مشخص نیست. (داده های فولدر ۱)

### بخش اول - خوشه بندی

در بخش اول، شما باید یک الگوریتم بخش بندی را پیاده سازی کنید که بر روی داده ها به خوبی عمل کند. انتخاب الگوریتم به عهده شماست، اما باید نشان دهید که این انتخاب، انتخاب مناسبی بوده است. برای این مهم، ابتدا یک راهکار بصری را بررسی کنید (کد این روش در ادامه قرار گرفته است) 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 مشخص کنید و بعد از آن، بررسی کنید که آیا خوشه بندی با الگوریتم مناسبی انجام شده است یا خیر، و اینکه آیا پارامتر مناسبی انتخاب شده است یا خیر.

### بخش دوم - پیدا کردن تعداد دامین

پیدا کردن تعداد دامین، یک عمل نسبتاً 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 روش هایی که در درس یاد گرفته اید، یک راهکار برای استفاده از این تعداد خوشه برای رسیدن به تعداد دامین ها بدهید (میتوانید خوشه ها را با هم ترکیب کنید - بر اساس معیار شباهت - و یا راهکار های ابتکاری دیگر). نهایتاً خروجی شما، باید تعداد دامین ها، و اینکه هر خوشه چه دامینی را نشان می دهد باشد (طبیعتاً یعنی اگر دو خوشه هر دو با خروجی الگوریتم شما دامین ۱ دارند، در داده های واقعی هم دامین آنها یکی باشد)

---

<sup>۱</sup> clustering

<sup>۲</sup> domain

### **بخش سوم - ارزیابی با برچسب های خوشه ها**

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