Pattern Recognition

HW8 : clustering

Spring 99

---

۱. L^p Norm

نورم p ام یک بردار در حالت کلی به شکل زیر تعریف می‌شود که در آن p یک عدد طبیعی است.

$$||x||_p = \left(\sum_{i=۱}^{n} |x_i|^p\right)^{\frac{۱}{p}} \qquad p \epsilon N.$$

با استفاده از تعریف بالا موارد زیر را اثبات کنید.

(آ) $||x||_\infty = max_j(|x_j|)$

(ب) $||x||_\infty \leq ||x||_۲ \leq \sqrt{n}||x||_\infty$

(ج) $||x||_\infty \leq ||x||_۲ \leq ||x||_۱$

** در سوالات ۲ تا ۵ از مجموعه داده DS1 استفاده کنید.

۲. Agglomerative Hierarchical Clustering

در این سوال الگوریتم خوشه بندی سلسله مراتبی تجمعی را بر روی مجموعه داده DS1 که همراه تمرین قرار گرفته است، اعمال می‌کنیم. در این سوال می‌توانید از کتاب‌خانه های آماده استفاده کنید. در مجموعه داده علاوه بر نمونه‌ها، شماره کلاستر هر نمونه نیز جدا گانه آورده شده است. برای مقایسه عملکرد الگوریتم از این شماره‌ها استفاده می‌کنیم.

(آ) با استفاده از الگوریتم خوشه بندی تجمعی داده ها را به ۳ خوشه دسته‌بندی کنید.

(ب) برای هر یک از خوشه‌ها متوسط فاصله از میانگین دسته را محاسبه کنید. با تحلیل اعداد بدست‌آمده انسجام درونی خوشه ها را بایک دیگر مقایسه کید.

$$Mean\ Dist\ Mean(S_i) = \frac{۱}{Q_i} \sum_{x \epsilon S_i} d(x, \mu_i)$$

در رابطه بالا منظور از $d(x,y)$ فاصله اقلیدسی دو بردار است، همچنین S_i نشانه خوشه i ام می‌باشد.

(ج) برای این خوشه‌بند، درخت Dendogram را رسم نمایید. سپس با استفاده از مفهوم life time تعیین کنید که چه تعداد خوشه برای این داده ها مناسب است.

(د) با استفاده از شماره خوشه هایی که در اختیار شما قرار گرفته است، دقت خوشه بندی را بدست آورید. توجه کنید که الگوریتم خوشه بندی تناظر یک‌به‌یک میان خوشه‌ها و شماره‌هایی که در اختیار شما قرار داده شده است بر قرار نمی‌کند و این نگاشت را خود شما باید ایجاد کنید. در جفت کردن خوشه ها به اعداد حالتی را در نظر بگیرید که دقت خوشه‌بندی بالاتر می‌شود.

۳. KMeans

الگوریتم KMeans از جمله الگوریتم هایی است که بر روی داده‌هایی با خوشه های compact بسیار موثر واقع می‌شود. در این الگوریتم هر یک از نمونه‌ها به یکی از خوشه های $S = \{S_۱, S_۲, ..., S_k\}$ تخصیص داده می‌شوند که در آن $k \leq Q$ است. تابع هدفی که در این الگوریتم بهینه می‌شود within-cluster sum of squares (WCSS) نام دارد که رابطه آن در زیر آورده شده است.

$$\min_s \left(\sum_{i=۱}^{k} \sum_{x \epsilon S_i} ||x - \mu_i||^۲\right)$$

مراحل الگوریتم:

۱. مقداردهی اولیه n ، Q ، k و μ_i. برای مقداردهی اولیه نشان‌دسته‌ها می‌توانید k نمونه از مجموعه داده را انتخاب کنید.

۲. تا زمانی که مقادیر نشان‌دسته ها ثابت نشده اند مراحل زیر را تکرار می‌کنیم:

۱.۲ هر یک از نمونه‌ها را به نزدیک ترین نشان‌دسته اختصاص می‌دهیم.

۲.۲ مقدار نشان‌دسته را به میانگین نمونه‌های اختصاص یافته به آن، بروزرسانی می‌کنیم.

۳. نشان دسته ها را باز می‌گردانیم.

برای اطلاعات بیشتر به اینجا مراجعه کنید.

(آ) با استفاده از الگوریتم بالا داده‌ها را به ۳ خوشه تقسیم کنید. (در این قسمت مجاز به استفاده از کتاب‌خانه های آماده نیستید.)

(ب) با استفاده از کتاب خانه **sklearn** مرحله قبل را تکرار کنید.

(ج) برای هر یک از خوشه‌بند های بالا مقدار **Mean Dist Mean** را برای هر یک از خوشه ها محاسبه کنید. با تحلیل این اعداد عملکرد پیاده سازی خود و کتاب‌خانه را مقایسه کنید.

(د) دقت هر یک از خوشه بند های بالا را محاسبه کنید. توجه کنید که الگوریتم خوشه بندی تناظر یک‌به‌یک میان خوشه‌ها و شماره‌هایی که در اختیار شما قرار داده شده است بر قرار نمی‌کند و این نگاشت را خود شما باید ایجاد کنید. در جفت کردن خوشه ها به اعداد حالتی را در نظر بگیرید که دقت خوشه‌بندی بالاتر می‌شود.

۴. Separation Index

شاخص **SI** یک متریک برای **cluster validity** می‌باشد. این شاخص متناسب با نسبت فاصله‌برون خوشه‌ای بر فاصله درون خوشه‌ای می‌باشد، بنابراین هرچه خوشه ها از یک دیگر تمییز پذیرتر باشند و انسجام درونی بیشتری داشته باشند، مقدار **SI** بزرگتری دارند. رابطه این شاخص در زیر آورده شده است.

$$SI = \min_{j} \left\{ \min_{i(i \neq j)} \left\{ \frac{d(S_i, S_j)}{\max_l d(S_l, S_l)} \right\} \right\}$$

$$d(S_i, S_j) = \min_{x,y} \{d(x, y | x \epsilon S_i, y \epsilon S_j)\}$$

$$d(S_l, S_l) = \min_{x,y} \{d(x, y | x, y \epsilon S_l)\}$$

شاخص **SI** را برای هر یک از خوشه‌بند های سوال ۲و۳ محاسبه کنید. کدام یک از خوشه بند ها اعتبار بیشتری دارد؟

۵. Fisher's Discriminant Index

یکی دیگر از شاخص های جدایی پذیری **FDI** می‌باشد که با آن در مبحث **multi-class LDA** آشنا شدید. رابطه این شاخص در زیر آمده است.

$$FDI = trace\left(S_W^{-1} S_B\right)$$

$$S_W = \sum_{i=1}^{k} S_i \qquad S_B = \sum_{i=1}^{k} Q_i \left(\hat{\mu}_i - \hat{\mu}\right)\left(\hat{\mu}_i - \hat{\mu}\right)^T \qquad S_i = \sum_{q=1}^{Q_i} \left(x^q - \hat{\mu}_i\right)\left(x^q - \hat{\mu}_i\right)^T$$

$$\hat{\mu}_i = \sum_{x^q \epsilon S_i} x^q \qquad \mu = \sum_{q=1}^{Q} x^q$$

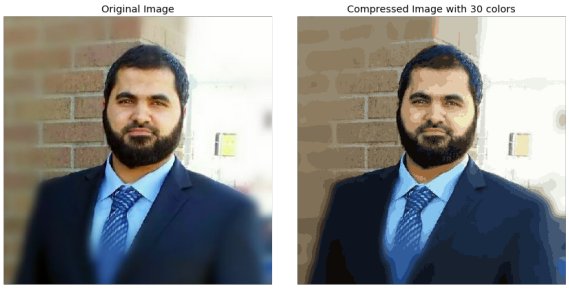
شاخص **FDI** را برای هر یک از خوشه‌بند های سوال ۲و۳ محاسبه کنید. کدام یک از خوشه بند ها اعتبار بیشتری دارد؟

۶. Image Compression using KMeans

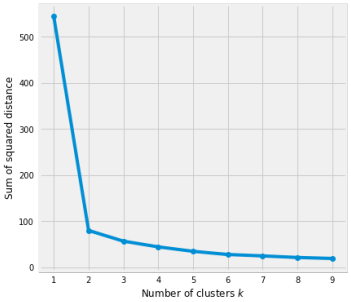
از الگوریتم KMeans میتوان برای فشرده سازی تصاویر استفاده کرد. به این صورت که k رنگ از رنگ های تصویر انتخاب می‌شود و هر یک از رنگ‌های تصویر به نزدیک ترین رنگ از k رنگ تغییر پیدا می‌کند. در حالت معمولی برای ذخیره سازی تصویر به $rows \times columns \times ۳ \times ۸$ بیت نیاز است (هر پیکسل ۳ کانال رنگی دارد که هر رنگ با ۸ بیت نمایش داده می‌شود.) ، در حالی که با استفاده از KMeans می‌توان تصویر را با اندازه $rows \times columns \times \log_۲ k + k \times ۳ \times ۸$ بیت نمایش داد.

با استفاده از الگوریتم KMeans تصویر همراه تمرین را فشرده کنید و تصویر حاصل را در گزارش خود بیاورید.

برای انتخاب k مناسب از Elbow Method استفاده می‌کنیم. در این روش خطای MSE میان تصویر فشرده شده و تصویر اصلی را برای k های مختلف رسم می‌کنیم. نقطه شکستگی نمودار را به عنوان k مناسب انتخاب می‌کنیم.



(آ) نمونه از فشرده سازی تصویر با استفاده از خوشه‌بندی



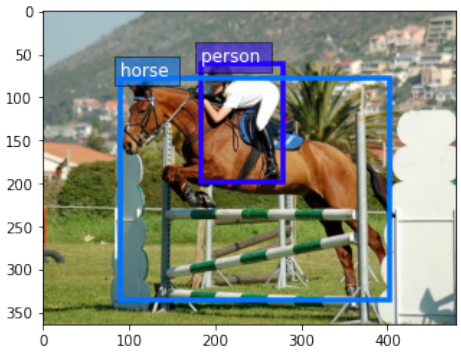
(ب) روش elbow برای انتخاب k مناسب

۷. Bounding Box Clustering

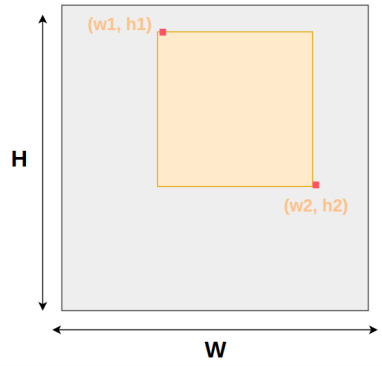
استفاده اصلی bounding box ها در سیستم های object detection می باشد. این باکس ها حدود یک شی را در تصویر مشخص می‌کنند. اهمیت دسته بندی این باکس‌ها در آنجا است که سیستم هایی همانند YOLO مشخصاتی از جندین باکس به عنوان هایپر پارامتر دارند. در این تمرین روشی بر مبنای KMeans Clustering برای تعیین این هایپر پارامتر ها معرفی می‌شود.

در این سوال از مجموعه داده PASCAL Visual Object Classes (VOC) استفاده می‌شود که برای سادگی بیشتر نشان‌گذاری های این مجموعه داده در فایل bboxes.csv در اختیار شما قرار گرفته است.

یک باکس در حالت کلی با ۴ عدد $(w_۱, h_۱, w_۲, h_۲)$ نشان داده می شود که در آن $(w_۱, h_۱)$ مختصات گوشه بالا-چپ و $(w_۲, h_۲)$ مختصات گوشه پایین-راست می‌باشد. از آنجایی که موقعیت و اندازه نسبی باکس در خوشه بندی اهمیتی ندارد، هر باکس را به صور نرمالیزه شده $\left(\frac{w_۲ - w_۱}{W}, \frac{h_۲ - h_۱}{H}\right)$ نشان می‌دهیم که در آن H و W ، طول و عرض تصویر می‌باشد.



(آ) نمونه تصویر از مجموعه داده



(ب) شخصات یک باکس در تصویر

(آ) با استفاده از الگوریتم KMeans و فاصله اقلیدسی، داده ها را یک‌بار به ۵ خوشه و بار دیگر به ۹ خوشه تقسیم کنید. با استفاده از scatter plot چینش خوشه ها را نمایش دهید.

متریکی که در قسمت قبل برای فاصله سنجی استفاده شد، متریک مناسبی نیست چرا که به باکس هایی با اندازه بزرگ تر ارزش بیشتری می‌دهد. (اگر از پارامتر alpha در satter plot استفاده کنید، چگالی نقاط بهتر دیده می‌شوند و این ارزش دهی نادرست فاصله اقلیدسی مشهود تر است.) به همین دلیل از متریک Intersection over Union برای شباهت سنجی دو باکس استفاده خواهیم کرد. برای اطلاعات بیشتر به اینجا مراجعه کنید.

$$IoU(b_1, b_2) = \frac{\min(x_1, x_2) \times \min(y_1, y_2)}{x_1 y_1 + x_2 y_2 - \min(x_1, x_2) \times \min(y_1, y_2)}$$

$$b_1 = (x_1, y_1) \qquad b_2 = (x_2, y_2) \qquad d(b_1, b_2) = 1 - IoU(b_1, b_2)$$

(ب) این بار با استفاده از متریک IoU داده هارا یک‌بار به ۵ خوشه و بار دیگر به ۹ خوشه تقسیم کنید. همچنین با استفاده از scatter plot تقسیم بندی داده ها را نشان دهید.

شکل ۳: متریک IoU برای چند باکس

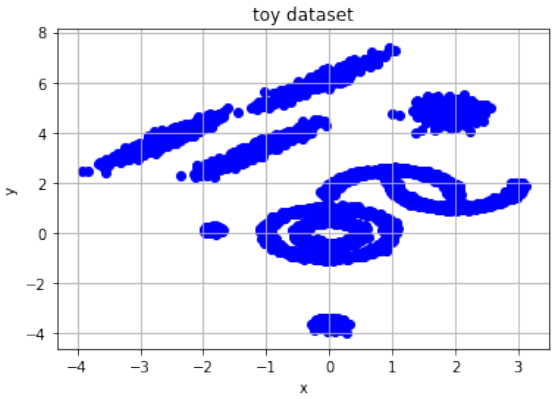
۸. DBSCAN vs KMeans

دسته‌دیگری از الگوریتم های خوشه‌بندی بر اساس چگالی داده ها خوشه ها را انتخاب می‌کنند، از این دسته از الگوریتم ها می‌توان به روش خوشه‌بندی Density-Based Spatial Clustering of Applications with Noise اشاره کرد. برای آشنایی بیشتر به اینجا و اینجا مراجعه کنید.

(آ) به صورت مختصر عملکرد این الگوریتم را توضیح دهید

(ب) در این قسمت می‌خواهیم عملکرد DBSCAN و KMeans را بر روی یک مجموعه داده ساختگی (ToyDataSet.csv) با یک‌دیگر مقایسه کنیم. این مجموعه داده دارای ۱۰ خوشه مجزا است. با استفاده از هریک از روش ها داده ها را خوشه بندی کنید و با استفاده از Scatter plot خوشه ها را نمایش دهید.

(ج) عملکرد DBSCAN بر چه نوع خوشه هایی بهتر از عملکرد KMeans می‌باشد؟



شکل ۴: مجموعه داده ساختگی

نکات پایانی:

۱. شما باید پاسخ های خود را با الگو PATREC-HW8-SID.zip در محل تعیین شده آپلود کنید

۲. گزارش شما معیار اصلی ارزیابی خواهد بود، در نتیجه دقت کنید کیفیت عکس ها مناسب باشند.

۳. کدهای خود را به تفکیک سوال ارسال کنید و استفاده از دیگر زبان های برنامه‌نویسی ممانعتی ندارد.

۴. هدف از انجام تمارین یادگیری مباحث درس می‌باشد، بنابر این تمارین را خودتان انجام دهید. در صورت کشف مشابهت بلاتوجیه، با توجه به قوانین درس عمل خواهد شد.

۵. شما میتوانید سوالات خود را از طریق ایمیل sj.pakdaman@ut.ac.ir بپرسید

- موفق باشید