هوش مصنوعی **تمرین شماره‌ی ۳**

مدرسین: **دکتر فدایی و دکتر یعقوب‌زاده**

طراحان: نیما مدیر، محمدرضا علوی، محمدامین یوسفی

مهلت تحویل: شنبه 1 اردیبهشت ۱۴۰۳، ساعت ۲۳:۵۹

بخش کتبی

Clustering

سوال اول

با استفاده از الگوریتم k-means نقاط زیر را در سه cluster قرار دهید و به سوالات پاسخ دهید (فرض کنید مرکز اولیه ی cluster ها به ترتیب نقاط A1 و A4 و A7 هستند. الگوریتم را فقط یک epoch اجرا کنید):

A1=(2,10), A2=(2,5), A3=(8,4), A4=(5,8), A5=(7,5), A6=(6,4), A7=(1,2), A8=(4,9)

الف) بعد از یک epoch برای هر نقطه cluster آن را مشخص کنید.

ب) مرکز های جدید را محاسبه کنید.

ج) نقاط را رسم کنید و در شکل cluster های آن ها را مشخص کنید.

د) چند iteration دیگر نیاز است تا الگوریتم همگرا شود؟ (دیگر cluster های نقاط تغییر نکند) برای هر iteration صرفا نقاط و cluster های آن ها را رسم کنید.

سوال دوم

نقاط سوال قبل را با الگوریتم DBSCAN دسته بندی کنید. یک بار ε را 2 و یک بار $\sqrt{10}$ در نظر بگیرید. در هر دو حالت minpoints را 2 در نظر بگیرید. شکل نقاط و cluster ها را رسم کنید. نقاط همسایه ی هر نقطه و نقاط نویز را نیز مشخص کنید.

سوال سوم

با الگوریتم agglomerative clustering و ماتریس فاصله ی زیر نقاط A تا D را (مرحله به مرحله) دسته بندی کنید و نمودار hierarchical آن ها را رسم کنید. فاصله ی دو دسته را فاصله ی نزدیک ترین نقاط آن ها در نظر بگیرید.

	A	B	C	D
A	0	1	4	5
B		0	3	6
C			0	2
D				0

بخش عملی

مقدمه

خوشه بندی یا Clustering تکنیکی است که شامل گروه بندی اشیاء مشابه بر اساس شباهت های ذاتی آن ها می شود. به عبارت دیگر، هدف آن است که نقاط داده را به خوشه های مجزا تقسیم کند، به صورتی که نقاط درون یک خوشه بیشتر به یکدیگر شباهت داشته باشند تا به خوشه های دیگر. با کشف این گروه بندی های طبیعی، الگوریتم های خوشه بندی می توانند بینش های ارزشمندی را در مورد ساختار زیربنایی داده ها ارائه دهند. خوشه بندی در حوزه های مختلفی از جمله تقسیم بندی مشتری، دسته بندی تصاویر و اسناد، تشخیص ناهنجاری و سیستم های توصیه کاربرد دارد.

تعریف مسئله

در این پروژه قصد داریم با استفاده از الگوریتم‌های Clustering، به تجزیه و تحلیل تصاویر تعدادی گل مختلف بپردازیم و سعی کنیم با استفاده از داده‌هایی که در اختیار داریم، آن‌ها را در دسته‌بندی‌های مختلف قرار دهیم، به طوری که بعد از اعمال الگوریتم خوشه‌بندی تا حد ممکن در خوشه درست خودشان قرار گرفته باشند.

آشنایی با مجموعه داده

مجموعه داده ای که در این پروژه استفاده می شود، شامل تعدادی عکس رنگی از گونه های مختلف گل می باشد که در کنار صورت پروژه در اختیارتان قرار گرفته است. همچنین یک فایل csv در کنار این تصاویر قرار دارد که حاوی لیبل های مرتبط با این تصاویر است. شما در انتها در بخش ارزیابی از این دسته بندی ها استفاده خواهید کرد.




پیش پردازش و استخراج ویژگی

در این بخش باید اطلاعات موجود در عکس‌ها را با استفاده از مدل VGG16 استخراج کنید. VGG16 یک pre-trained Convolutional Neural Network می‌باشد که می‌توانید طریقه استفاده از آن را جستجو نمایید.

برای استخراج ویژگی‌ها، ابتدا لایه‌های تماما متصل (Fully Connected) مدل VGG16 را حذف نموده و سپس استخراج را آغاز نمایید.

۱. علت استخراج ویژگی‌ها چیست ؟ چرا تنها به خواندن پیکسل‌ها بسنده نمی‌کنیم ؟ توضیح دهید.

۲. راجع به استخراج ویژگی از عکس ها تحقیق کنید و به طور خلاصه راجع به 3 تکنیک آن توضیح دهید.

۳. چه پیش پردازشی بر روی تصاویر باید انجام شود تا آماده وارد شدن به مدل شوند؟

پیاده‌سازی خوشه یابی

هدف کلی در این بخش استفاده از روش های clustering برای خوشه بندی عکس‌های دیتاست است.

روی بردارهای ویژگی استخراج شده، با استفاده از روشهای خوشه‌بندی که یاد گرفته‌اید (K-Means و DBSCAN)، داده‌هایتان را خوشه‌بندی کنید. تمامی پارامترهای مدل‌های مورد استفاده دست شماست و سعی کنید با آزمون و خطا به پارامترهای مناسبی برسید. توجه داشته باشید که در روش K-Means، انتخاب با تعداد دسته‌های گل‌ها باید تناسب داشته باشد. این مقدار مناسب برای K اهمیت بسیاری دارد و احتمالا موضوع در ارزیابی نتایج به شما کمک خواهد کرد.

۴. در مورد روش های K-Means و DBSCAN و مزایا و معایب این روشها نسبت به هم توضیح دهید.

۵. از چه روشی برای پیدا کردن مناسب ترین K در روش K-Means استفاده کرده اید؟ توضیح دهید.

۶. خروجی حاصل از دو نوع خوشه‌بندی را باهم مقایسه کنید.

کاهش بُعد

در این بخش، می‌خواهیم خوشه‌های استخراج شده در فاز قبلی را نمایش دهیم. نکته مهمی که در نمایش ابعاد بردار ویژگی زیاد بوده و همین موضوع باعث میشود که نتوان خوشه‌ها وجود دارد، این است که معمولا آن را در صفحه دو/سه‌بعدی به صورت مستقیم نمایش داد. برای حل این مشکل، از روشهای کاهش بعد مثل PCA استفاده میشود.

۷. درباره PCA تحقیق کنید و نحوه عملکرد آن را به اختصار توضیح دهید.

حال روی بردارهای ویژگی بدست آمده کاهش بعد را انجام دهید و با استفاده از بردارهای کاهش یافته، خوشه‌ها را نمایش دهید و خوشه‌های بدست آمده توسط دو الگوریتم را با یکدیگر مقایسه کنید. برای کاهش بعد میتوانید از کتابخانه sklearn استفاده کنید.

ارزیابی و تحلیل

در این بخش به ارزیابی نتایج حاصل از پیاده سازی روشها می‌پردازیم. برای ارزیابی روشهای خوشه‌بندی، می‌توان دقت خوشه‌بندی را با استفاده از دسته های واقعی دادهها و بدون استفاده از آن اندازه گیری کرد. برای مطالعه این روشها میتوانید از این لینک استفاده کنید. برای روشهای مبتنی بر label true، از معیار homogeneity و برای روشهای غیر از آن از امتیاز silhouette استفاده میکنیم.

۸. در مورد نحوه محاسبه معیار silhouette و homogeneity توضیح دهید.

۹. نتایج حاصل از معیارهای ذکر شده را برای هر یک از روشها گزارش کنید.

۱۰. راهکارهایی پیشنهاد کنید که بتوان عملکرد مدلها را بهبود داد.

نکات پایانی

- توضیحات مربوط به هر بخش از پروژه را بطور خلاصه و در عین حال مفید در گزارش خود ذکر کنید. از ابزارهای تحلیل داده مانند نمودارها استفاده کنید.
- پس از مطالعه کامل و دقیق صورت پروژه، در صورت وجود هرگونه ابهام یا سوال با طراحان پروژه در ارتباط باشید.
- نتایج، گزارش و کدهای خود را در قالب یک فایل فشرده با فرمت AI_CA3_[stdNumber].zip در سامانه ایلرن بارگذاری کنید.
- محتویات پوشه باید شامل فایل jupyter-notebook، خروجی html و فایل‌های مورد نیاز برای اجرای آن باشد. از نمایش درست خروجی‌های مورد نیاز در فایل html مطمئن شوید.
- دقت کنید که نیازی به آپلود مجموعه داده‌ها در سامانه ایلرن نیست.