به نام خدا

تمرین سری چهارم یادگیری ماشین

محمد میرزانژاد – 830596027

1)

1-1)

آ)عکس های رنگی با مقادیر مختلف K

K=2

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

ب) عکس های سیاه سفید با مقادیر مختلف K

پ) عکس های رنگی و مختصات هر پیکسل با مقادیر مختلف K

1-2)

آ) با توجه به اینکه در روش K-means ،هر بار K نقطه تصادفی انتخاب میشود و در پیاده سازی نیز این مورد لحاظ شده است، در ابتدا نزدیک ترین پیکسل ها به نقطه تصادفی انتخاب شده با توجه به مولفه RGB ، تشکیل خوشه میدهند ، سپس میانگین هر خوشه به عنوان centroid جدید انتخاب میشود و خوشه بندی با توجه به نزدیکی پیکسل ها به centroid های جدید انجام میشود ، لذا در نهایت نزدیکترین پیکسل ها به هر مرکز خوشه مشخص شده و در همان خوشه قرار میگیرند ، بنابراین انتخاب نفطه اولیه تاثیری بر روی خوشه بندی ندارد.

ب) یک روش برای تعیین بهترین K جهت خوشه بندی روش Silhouette است. در این روش به ازای هر داده ( در اینجا پیکسل ) میانگین فاصله هر داده با سایر داده ها در همان کلاستری که در آن قرار دارد، محاسبه میشود ( a(i) ) . سپس میانگین فاصله هر داده با داده های موجود در تمام کلاسترهای دیگر به غیر از کلاستری که در آن قرار دارد محاسبه میشود و کوچکترین آن به عنوان b(i) انتخاب میشود. سپس برای هر داده مقدار زیر محاسبه میشود :



مقدار s(i) همیشه بین 1و 1- قرار میگیرد و هر چه این مقدار به 1 نزدیکتر باشد ایده آل تر است. در واقع مقدار 1 نشان دهنده این است که داده در خوشه مناسب قرار دارد ، مقدار 0 نشان میدهد که داده در منطقه مرزی قرار دارد که میتواند متعلق به خوشه مجاور باشد و مقدار 1- نشان دهنده این است که داده متعلق به خوشه فعلی نیست و متعلق به دورترین خوشه نسبت به خوشه فعلی است. البته مقادیر بدست آمده بین 1 و 1- هستند و نزدیک بودن بیشتر به 1 نشان دهنده مناسب ترین خوشه است و فاصله بیشتر نشان دهنده فاصله بیشتر از خوشه فعلی و تعلق به خوشه هاس دورتر به لحاظ مجاورت می باشد. در این روش به ازای k های مختلف الگوریتم k-means اجرا میشود و سپس میانگین s(i) فوق برای تمامی پیکسل ها اجرا و میانگین آن محاسبه میشود ، و آن دور از اجرا یا K که بیشترین میانگین s(i) را حاصل میدهد ، نشان دهنده بهترین K است . که در برنامه ما بهترین K برای عکس های مختلف متفاوت بود که در ادامه آمده است.

پ) تغییر مقدار آستانه اگر از حد زیادی بیشتر باشد باعث میشود الگوریتم قبل از همگرا شدن خاتمه یابد، یعنی باعث میشود مرکز خوشه تعیین شده دارای فاصله مناسب با سایر پیکسلها نباشد و مرکز خوشه و خوشه های موجود در آن دارای میانگین فاصله مناسب نباشند و داده هایی در خوشه قرار داشته باشند که مربوط به آن خوشه نباشند . از طرف دیگر اگر مقدار آستانه را صفر در نظر بگیریم ، ممکن است الگوریتم هیچ وقت همگرا نشود زیرا در مراحل پایانی ممکن است مراکز خوشه بین یک مقدار خاص نوسان پیدا کنند و هیچگاه فیکس نشوند.

در زیر نمونه هایی آمده اند که مقدار آستانه بیش از حد زیاد در نظر گرفته شده است : ( مقدار : )

2) از بین feature ها تنها PS ID میتواند جهت خوشه بندی مناسب باشد چرا که شناسه هر دستگاه را نشان میدهد. از الگوریتم خوشه بندی سلسله مراتبی پایین به بالا ( Bottom To Up یا Agglomerative ) استقاده میکنیم. از فاصله منهتن جهت تشکیل کلاستر های اولیه استفاده میکنیم . در واقع هر نمونه داده در اینجا یک cluster است که با توجه به نزدیکس یه نمونه دیگر به صورت دو به دو با هم تشکیل یک cluster جدید میدهند. برای تجمیع کلاسترها با یکدیگر و تعیین فاصله آنها از یکدیگر از یکی از معیارهای Linkage Criteria میتوان بهره برد که در اینجا میتوان از Minimum or single linkage Clustering استفاده کرد ، یعنی فاصله نزدیکترین نقاط دو ملاستر نسبت به هم را ملاک قرار داد. الیته روشی که به نظر جواب بهتری میتواند بدهد روش Centroid linkage clustering که میتوان برای هر خوشه یک میانگین حساب کرد و سپس فاصله حاصل از میانگین دو خوشه را به عنوان ملاک تجمیع دو کلاستر قرار داد. به این روش  nearest neighbour clustering نیز گفته میشود.

ابندا از یک کلاستر که همان یک نمونه داده است شروع میکنیم و فاصله آن را با بقیه کلاسترها محاسبه میکنیم ، کوتاهترین فاصله را ملاک تجمیع کلاستر قرار میدهیم. یعنی کلاستری که کمترین فاصله را دارد تجمیع میشود .

در خصوص این مثال ما میتوانیم یک Threshold قائل شویم که اگر فاصله منهتن چند نمونه داده از یکدیگر ( مقایسات دو به دو ) از یک میزان آستانه ای کمتر بود آنها را داخل یک کلاستر تجمیع کنیم و و این کلاستر را شهر بنامیم ، سپس همین کار را برای استان انجام دهیم و اگر فاصله میانگین داده های کلاستر ( جمع آنها تقسیم بر تعداد ) از میانگین داده های دیگر از یک حد آستانه ای کمتر بود آن کلاسترها با هم ادعام میشوند و تشکیل کلاستر سطح بالاتری به نام استان میدهند.

پس در این روش ما به دو سطح از clustering نیاز داریم ، ابتدا تشکیل کلاسترهای شهر و سپس استان. برای شهر میتوانیم آستانه 5 را قائل شویم و هر گاه فاصله منهتن آنها کمتر از 5 بود آنها را در یک کلاستر شهر قرار دهیم. سپس برای استان حد آستانه 10 قائل شویم که هرگاه اختلاف میانگین هر خوشه با خوشه دیگر کمتر از 10بود آن دو خوشه یا بیشتر ، تشکیل یک خوشه استان دهند . به عنوان مثال نمونه های با مقدار PS ID ، 1019، 1020 و 1022 تشکیل یک خوشه شهر میدهند چرا که فاصله شان از هر کدام کمتر از 5 است . سپس این خوشه با خوشه که فقط حاوی نمونه داده 1012 است تشکیل یک خوشه استان میدهند چرا که میانگین خوشه اول 1020 و فاصله آن با میانگین خوشه دوم کمتر از 10 است . این روال ابتدا برای خوشه بندی شهرها صورت میگیرد و سپس خوشه بندی استان تا جایی ادامه پیدا میکند که هیچ خوشه ای وجود نداشته باشد که فاصله آن با خوشه دیگر از آستانه تعیین شده ( در اینجا 10 ) کمتر باشد.

برای ارزیابی مدل ساده ترین روش مراجعه به دیتای برچسب خورده و سنجش صحت عملکرد خوشه بندی است که در بسیاری از مواقع این امکان وجود ندارد. یکی از روشهای ارزیابی خوشه بندی ، تخمین تعداد کلاسترهای تشکیل شده است. برای تخمین تعداد کلاسترها سه روش شناخته شده وجود دارد که در تمرین 1 روش Silhoette استفاده شده و توضیح داده شده است. دو روش دیگر Elbow Method و GAP Statistic می باشند.