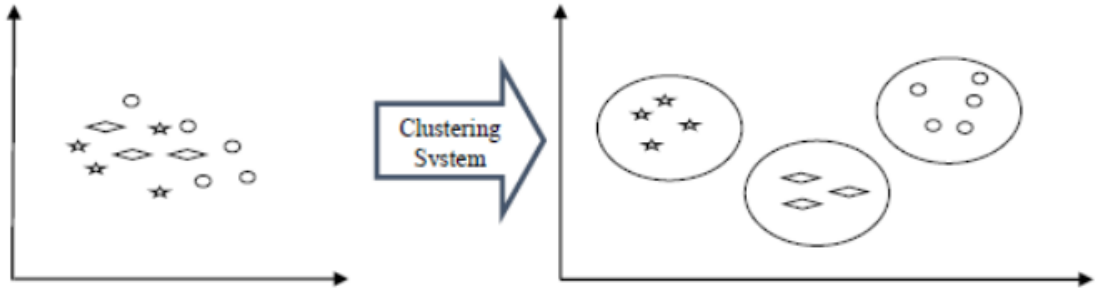
مقدمه ای بر خوشه بندی (Clustering) :

متدهای خوشه بندی یکی از کاربردی ترین متدهای یادگیری ماشین بدون نظارت هستند. از این متدها برای پیدا کردن شباهت و همچنین الگوهای ارتباطی بین نمونه های داده استفاده می شود. سپس، آن نمونه ها را درون گروه هایی که بر اساس ویژگی ها، مشابه هستند خوشه بندی می کند.

از آنجایی که خوشه بندی، گروه بندی طبیعی بین داده های بدون برچسب موجود را مشخص می کند، مهم است. آنها اساسا با در نظر گرفتن برخی فرضیات درباره نقاط داده ، تشابهات آنها را تشکیل می دهند. هر فرضیه خوشه های متفاوت اما با درجه اعتبار برابر خواهد ساخت.

برای مثال، شکل زیر یک سیستم خوشه بندی که داده ها از نوع مشابه را در خوشه های متفاوت قرار میدهد نشان می دهد.



متدهای شکل گیری خوشه: الزامی برای شکل گیری خوشه ها در قالب کروی وجود ندارد. برخی از متدهای شکل گیری خوشه ها در ادامه معرفی می شود.

مبتنی بر چگالی (Density-Based) : در این متدها، خوشه ها به عنوان نواحی چگال تشکیل می شوند. مزیت این متدها این است که از دقت و همچنین توانایی خوبی در ادغام دو خوشه برخوردار هستند. به عنان مثال، خوشه بندی فضایی مبتنی بر تراکم برنامه های کاربردی به همراه نویز (DBSCAN) و نقاط مرتب شده برای شناسایی ساختار خوشه بندی(OPTICS) .

مبتنی بر سلسله مراتب (hierarchical-based) : در این متدها، خوشه ها در قالب ساختار درختی مبتنی بر سلسله مراتب تشکیل می شوند و دارای دو خوشه به نام های جمع کننده(Agglomerative) (راهکار پایین به بالا) و تقسیم کننده(Divisive) (راهکار بالا به پایین) هستند. به عنوان مثال، خوشه بندی با استفاده نمایندگان (CURE) و خوشه بندی کاهش تکرار متعادل با استفاده از سلسله مراتب (BIRCH).

پارتیشن بندی: در این متدها، خوشه بندی به واسطه تقسیم اشیا در k تعداد خوشه تشکیل می شود. تعداد خوشه ها با تعداد پارتیشن ها برابر خواهد بود. مانند، K-means و خوشه بندی برنامه های کاربردی بزرگ مبتنی بر جستجو تصادفی (CLARANS).

شبکه توری (grid) : در این متدها، خوشه ها در قالب ساختار شبکه توری تشکیل می شوند. مزیت این متدها این است که عمل خوشه بندی انجام شده روی این شبکه های توری، مستقل از تعداد اشیای داده است. مانند، شبکه اطلاعاتی آماری (STING) و خوشه بندی در تحقیق (CLIQUE).

اندازه گیری کارایی خوشه بندی: یکی از مهمترین ملاحظات مربوط به مدل یادگیری ماشین، ارزیابی کارایی آن و یا به عبارتی کیفیت مدل است. در صورت استفاده از الگوریتم های یادگیری تحت نظارت، ارزیابی کیفیت مدل آسان خواهد بود زیرا برای همه مثال ها بر چسب داریم.

از طرف دیگر، در صورت استفاده از الگوریتم های یادگیری بدون نظارت، کار به آن سادگی ها نخواهد بود زیرا با داده های بدون برچسب سر و کار داریم. اما همچنان معیار هایی را در اختیار داریم که به تقسیم کننده درکی درباره وقوع تغییرات در خوشه ها براساس الگوریتم می دهد.

پیش از ورود به این معیار ها، باید متوجه باشیم که این معیار ها فقط کارایی مقایسه ای مدل ها را نسبت به یکدیگر ارزیابی میکند، تا اندازه گیری اعتبار پیش بینی مدل. در ادامه برخی از معیارهایی که می توان برای اندازه گیری کیفیت مدل، روی الگوریتم های خوشه بندی اعمال کرد را بررسی می کنیم.

تجزیه و تحلیل Silhouette : از تجزیه و تحلیل Silhouette برای بررسی کیفیت مدل خوشه بندی به واسطه اندازه گیری فاصله بین خوشه ها استفاده می شود. اساسا روشی برای ارزیابی پارامترهایی مانند تعداد خوشه ها را با کمک **Silhouette score** فراهم می کند. این score اندازه گیری میکند که هر نقطه در یک خوشه چقدر به نقاط خوشه های همسایه نزدیک است.

تحلیل Silhouette score : بازه **Silhouette score** بین 1- و 1 است.

انواع الگوریتم های خوشه بندی ML : در ادامه مهمترین و کاربردی ترین الگوریتم های خوشه بندی ML معرفی می شوند.

خوشه بندی K-means : این الگوریتم خوشه بندی به محاسبه نقاط مرکزی می پردازد و به تکرار آن ادامه می دهد تا نقطه مرکزی بهینه را پیدا کند. این الگوریتم فرض می کند که در حال حاضر تعداد خوشه ها مشخص است. همچنین به آن الگوریتم خوشه بندی flat نیز گفته می شود. تعداد خوشه های معین شده در داده توسط الگوریتم، با k در K-means نمایش داده می شود.

الگوریتم mean – shift : این الگوریتم یکی دیگر از الگوریتم های قدرتمند خوشه بندی است که در یادگیری بدون نظارت از آن استفاده می شود. برخلاف خوشه بندی K-means ، از آنجایی که mean – shift یک الگوریتم بدون پارامتر است، هیچ فرضی را در نظر نمی گیرد.

خوشه بندی سلسله مراتبی: این الگوریتم یکی دیگر از الگوریتم های یادگیری بدون نظارت است که از آن برای گروه بندی نقاط داده بدون برچسب که ویژگی های مشابه دارند استفاده می شود.

در بخش بعدی درباره همه این الگوریتم ها با جزییات صحبت خواهیم کرد.

برنامه های کاربردی خوشه بندی: در زمینه های زیر خوشه بندی می تواند مفید باشد.

خلاصه سازی و فشرده سازی داده: خوشه بندی در زمینه هایی که نیازمند خلاصه سازی، فشرده سازی و کاهش داده هستیم، به طور گسترده استفاده می شود. مثال هایی از آن پردازش تصویر و کمی سازی بردار است.

سیستم های مشارکتی و طبقه بندی مشتری: از آنجایی که می توان از خوشه بندی برای پیدا کردن محصولات مشابه یا کاربران هم نوع استفاده کرد، می تواند در زمینه سیستم های مشارکتی و طبقه بندی مشتری مفید باشد.

عمل به عنوان یک مرحله میانی اصلی برای سایر کارهای داده کاوی: تجزیه و تحلیل خوشه می تواند یک خلاصه فشرده از داده برای طبقه بندی، تست و تولید فرضیه، تولید کند. اگرچه به عنوان یک مرحله میانی اصلی برای سایر کارهای داده کاوی نیز کار میکند.

تشخیص روند در داده پویا: می توان با ساخت خوشه های متنوع از روند های مشابه، از خوشه بندی برای تشخیص روند در داده پویا استفاده کرد.

تجزیه و تحلیل شبکه های اجتماعی: می توان از خوشه بندی در تجزیه و تحلیل شبکه های اجتماعی استفاده کرد. مثال هایی از آن عبارتند از تولید توالی در تصاویر، ویدئو و صوت.

تجزیه و تحلیل داده های بیولوژیکی: از آنجایی که خوشه بندی به طور موفقیت آمیزی می تواند داده های بیولوژیکی را تجزیه و تحلیل کند، می توان از آن برای ساخت خوشه های تصویر و ویدئو استفاده کرد.