به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس یادگیری ماشین

استاد ناظرفرد

تمرین چهارم

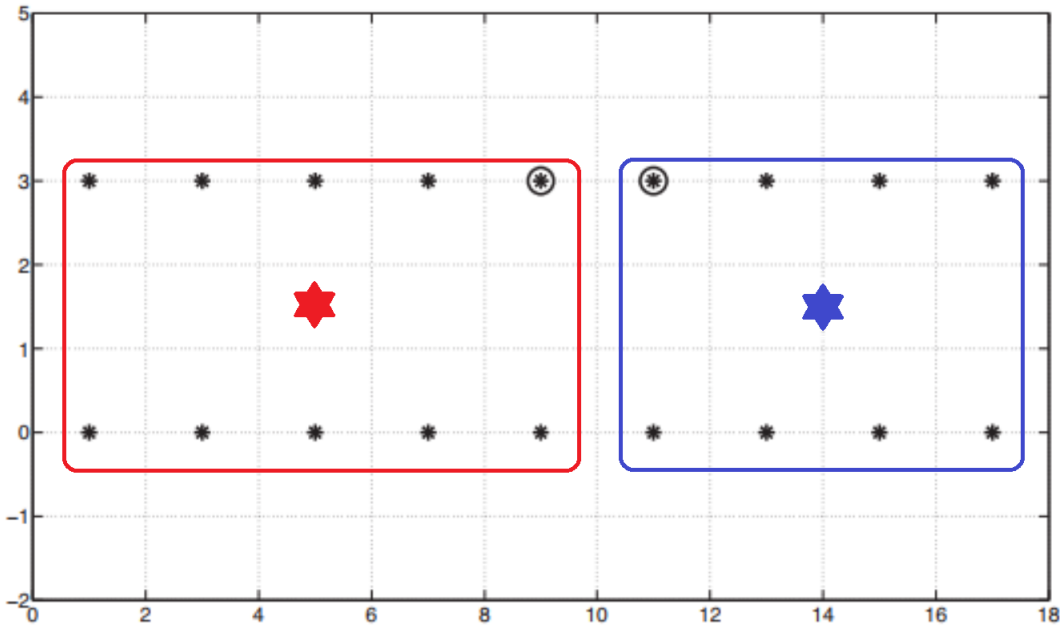
علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

بخش اول: پرسش‌های تشریحی

سوال ۱

دو خوشه آبی و قرمز به همراه مراکز که با ستاره نشان داده شده‌اند در تصویر زیر آورده شده است:



سوال ۲

الف) وقتی که یک الگوریتم خوشه‌بندی را بر روی داده‌های کاربران اعمال کنیم، کاربرانی در یک خوشه قرار خواهند گرفت که ویژگی‌های مشابه به یکدیگر داشته باشند. این همان چیزی است که در کاربرد Customer Segmentation مدنظر است.

ب) وقتی که خوشه‌بندی انجام می‌گیرد و مثلا k خوشه داشته باشیم. اگر خوشه‌بندی سخت باشد، هر داده به یک خوشه تعلق دارد و به مابقی نه. پس می‌توان برای هر داده یک بردار k-بعدی در نظر گرفت که هر بعد متعلق به یک خوشه باشد. هر بعد مقدار یک بگیرد اگه داده به آن خوشه تعلق داشته باشد و در غیر این صورت صفر. اگر خوشه‌بندی نرم باشد، هم طبیعتا برای هر بعد مقدار تعلق داده به آن خوشه نشان داده می‌شود. نهایتا اینکه داده‌ها از فضای اولیه با ابعاد احتمالی بالا به ابعاد k کاهش بعد پیدا می‌کنند.

ج) در تعدادی از الگوریتم‌های خوشه‌بندی مانند DBSCAN داده‌های پرت به طور خودکار شناسایی می‌شوند و طبیعتا می‌توان شناسایی کرد که یک داده، داده غیرعادی و پرت هست یا خیر. در سایر الگوریتم‌ها هم باز می‌توان با داشتن یک معیار فاصله محاسبه کرد که یک داده با هر خوشه چه میزان فاصله دارد؛ طبیعی است که اگر داده با تمام خوشه‌ها فاصله بالایی داشته باشد، یک داده عادی نخواهد بود.

د) مشابه با سوال پیاده‌سازی می‌توان پیکسل‌ها را خوشه‌بندی کرد و بدین ترتیب پیکسل‌های هم خوشه را در یک قطعه قرار دارد. جدای از بحث خوشه‌بندی پیکسل، در هر روشی باید قسمتی از عکس را در یک قطعه قرار داد که نوعی شباهت میان ویژگی‌های درون آن قطعه باشد و این چیزی است که خوشه‌بندی انجام می‌دهد.

سوال ۳

اگر حجم داده‌ها پایین باشد و امکان اجرای چندباره الگوریتم فراهم باشد، می‌توان با رنجی از مقادیر الگوریتم DBSCAN را اجرا کرد و یک سری از شروط را در آن چک کرد؛ مثلا تعداد خوشه در یک بازه معقول و متناسب با کاربرد باشد و یا آنکه اندازه خوشه‌ها نسبت به یکدیگر از یک آستانه‌ای کمتر باشد. بدین ترتیب حالتی که شرایط را داشته باشد مورد پذیرش است. به طور مشابه می‌توان به جای شروط باینری، به هر وضعیت یک امتیاز متناسب با آن نسبت داد و مجموعه‌ای از پارامترها که بهترین امتیاز داشت را انتخاب کرد.

نهایتا باید توجه داشت که با یک سری بررسی آماری روی داده‌ها نظیر میانگین تراکم می‌توان یک رنج معقول اولیه پیدا کرد.

سوال ۴

در حالت value iteration یک فرآیند تکراری طی می‌شود تا برای هر وضعیت مقدار value یا امتیاز آن محاسبه شود. این فرآیند وقتی متوقف می‌شود که value ها همگرا شوند. پس از اتمام و پیداشدن این مقادیر، نوبت به پیداکردن policy بهینه است.

policy بهینه بر اساس جدول مقادیر بهینه بدست می‌آید. لذا این فرآیند یک بار بیشتر انجام نمی‌شود. این درحالی است در policy iteration ابتدا یک policy اولیه درنظر گرفته و سپس به صورت تکرارشونده ابتدا بر اساس policy مقادیر وضعیت‌ها مشخص می‌شوند و سپس بر اساس مقادیر وضعیت، policy بهبود پیدا می‌کند. یعنی در این حالت در هر گام policy و value با هم بروز می‌شوند و زمانی که policy همگرا شود الگوریتم متوقف می‌شود.

سوال ۵

الف) هنگامی که قصد ادغام دو خوشه را در الگوریتم سلسله مراتبی را داشته باشیم بسته به معیار فاصله‌های متفاوتی برای خوشه‌ها حاصل می‌شود:

- در Complete Link بیشترین فاصله میان یک عضو از خوشه اول و یک عضو از خوشه دوم به عنوان فاصله دو خوشه درنظر گرفته می‌شود.
- در Single Link کمترین فاصله میان یک عضو از خوشه اول و یک عضو از خوشه دوم به عنوان فاصله دو خوشه درنظر گرفته می‌شود.
- در Average Link میانگین فاصله تمام جفت داده‌ها که یکی از خوشه اول و دیگری از خوشه دوم باشد محاسبه می‌شود.

از نظر پیچیدگی زمانی برای هر سه حالت لازم است تا فاصله تمام جفت داده‌ها محاسبه شود تا بتوان به ترتیب بیشینه، کمینه و میانگین آن را محاسبه کرد. لذا پیاده‌سازی کلاسیک این سه روش تفاوتی از منظر پیچیدگی زمانی با یکدیگر نخواهند داشت.

از نظر حساسیت به نویز، طبیعتا Average Link باتوجه به حالت میانگین‌گیری که دارد کمترین حساسیت را نسبت به داده‌های نویز دارد. بین دو روش دیگر به طور قطعی نمی‌توان نظر داد ولی می‌توان گفت Single Link حساسیت بیشتری به داده‌های نویز دارد[1]. وقتی که داده‌های نویز وجود داشته باشد، این داده‌ها در میان خوشه‌های واقعی

[1] https://stats.stackexchange.com/q/304427/318893

قرار می‌گیرند و این امکان را ایجاد می‌کنند که برخی از خوشه‌های واقعی در مراحل اولیه در یک خوشه قرار بگیرند، اما طبیعتا Complete Link با این داده‌ها دچار مشکل نمی‌شود.

ب) در معیار Single Link دو خوشه ۱ و ۲ باهم خوشه می‌شوند و خوشه ۳ و ۴ باهم. چراکه در خوشه ۱ و ۲ داده‌هایی به هم خیلی نزدیک هستند. در معیار Complete Link دو خوشه ۱ و ۳ باهم خوشه می‌شوند و خوشه ۲ و ۴ باهم. چراکه یک چپ‌ترین داده خوشه ۱ از راست‌ترین داده خوشه ۲ فاصله زیادی دارد. در معیار Average Link هم خوشه ۱ و ۳ باهم خوشه می‌شوند و خوشه ۲ و ۴ باهم. چراکه به طور میانگین داده‌های ۱ به داده‌های ۳ نزدیک‌تر است تا ۲.

ج) برای مجموعه b معیار Single Link جواب می‌دهد چراکه موقع اجرای الگوریتم و در گام‌های اول که فاصله کم در نظر گرفته می‌شود تمام داده‌های یک خوشه به هم متصل می‌شوند چرا که در این معیار ملاک نزدیک‌ترین داده است و هر زیرخوشه از هر خوشه به زیرخوشه مجاور دیگری از آن خوشه دارای فاصله بسیار کمی است. اما طبیعی است که معیار Complete Link جواب ندهد. چراکه دو لبه‌ی هر خوشه از هم بسیار فاصله دارند و زمانی که قرار است تمام داده‌های یک خوشه به هم متصل شوند این مقدار فاصله برای دو زیرخوشه نهایی هر خوشه وجود خواهد داشت. در Average Link هم مشکل وجود خواهد داشت. چراکه داده‌های دو لبه‌ی هر خوشه میانگین فاصله را بالا می‌برند و این احتمال وجود دارد که لبه‌ی یک خوشه با داده‌های مرکزی خوشه دیگر زودتر تشکیل خوشه دهد.

برای مجموعه c اوضاع دو معیار Complete Link و Average Link متفاوت نخواهد. در این حالت روش Single Link هم به مشکل خواهد خورد. چراکه ممکن است لبه‌ی یک خوشه از طریق داده‌های جدید به مرکز یک خوشه دیگر متصل شود.