به نام خدا

عنوان:

تکلیف دهم درس یادگیری ماشین

استاد:

دکتر پدرام

دانشجو:

محمدعلی مجتهدسلیمانی

شماره دانشجویی:

4033904504

تاریخ:

29/09/1403

سوال 7

به طور کلی خوشه بندی سخت گیرانه و خوشه بندی فازی از چندین نظر بررسی کرد:

1. عضویت:

خوشه بندی سخت گیرانه: در این رویکرد، هر داده دقیقا به یک خوشه تعلق دارد. این مقدار مطلق و انحصاری است. داده یا جز خوشه هست یا نیست. هیچ درجه عضویتی وجود ندارد. یکی از مثال های این رویکرد k-means است.

خوشه بندی فازی: یک داده میتواند به چندین خوشه به صورت همزمان تعلق داشته باشد، با مقادیر مختلف درجه عضویت نسبت به هر خوشه. این درجه عضویت مقداری بین 0و1 خواهد بود. 0 به معنای عدم عضویت در آن خوشه، 1 به معنای عضویت کامل در آن خوشه و مقادیر بین 0 و1 نشانگر عضویت نسبی به آن خوشه است.

2. مرز خوشه‌ها:

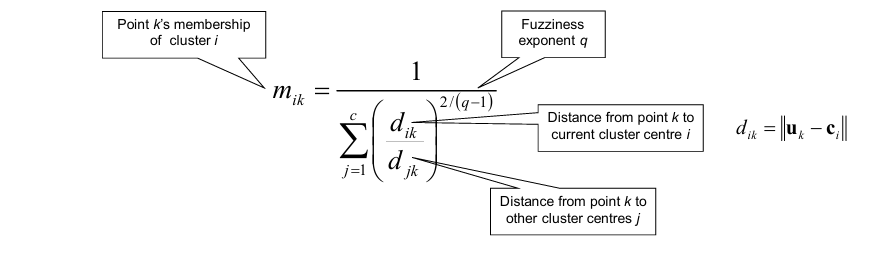
HCM: در این رویکرد مرزهای خوشه ها با هم هیچ همپوشانی ندارند و مجزا هستند.

FCM: در این رویکرد مرز خوشه ها نرم و دارای همپوشانی بین خوشه های مختلف است. یک داده میتواند در یک ناحیه ای قرار بگیرد که خوشه های مختلف در آنجا همپوشانی داشته باشد و درجه عضویت بیانگر حضور آن داده در خوشه های متعلق به آن است.

به سبب همپوشانی بین خوشه های مختلف، FCM مستعد عدم قطعیت خواهد بود زیرا عضویت های نسبی به هر داده نسبت به خوشه های مختلف میدهد. یک داده در نزدیک مرز بین 2 خوشه، درجه عضویت بالایی در هر 2 میتواند داشته باشد و باعث عدم قطعیت و ابهام در تخصیص یک خوشه به یک داده میشود. اما در HCM به سبب اینکه اجازه عضویت نسبی نمیدهد حتی اگر یک داده در نزدیکی خوشه دیگر باشد، جز آن خوشه حساب نمیشود.

همچنین در رویکرد فازی ما نمایه فازی هم داریم که در سوال بعد به صورت کامل توضیح داده شده است.

فرمول درجه عضویت خوشه بندی فازی:



نسبت فاصله: اگر نسبت کسر کوچک باشد به این معنی است که داده به خوشه i نزدیک تر از خوشه j خواهد بود. اگر نسبت کسر بزرگ باشد به این معنی است که داده به خوشه j نزدیک تر است نسبت به خوشه i.

سوال 8

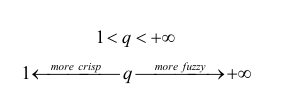
نمایه فازی یکی از پارامتر ها در خوشه بندی فازی هست که درجه همپوشانی بین خوشه های مختلف را مشخص میکند یا به طور بهتر درجه عضویت بین خوشه های مختلف را مشخص میکند که در رویکرد FCM مورد استفاده قرار میگیرد که داده ها نه لزوما فقط به یک خوشه بلکه میتوانند با یک درجه ای متعلق به تعدادی از خوشه ها باشند. این پارامتر تاثیر نسبت فاصله را کم یا زیاد میکند.

اگر q نزدیک به 1 باشد تخصیص عضویت بسیار سخت گیرانه خواهد بود مانند k-means معمولی و مرز بین خوشه ها بسیار تند خواهد بود.

اگر q کمی بزرگتر از 1 باشد، عضویت بیشتر فازی خواهد بود به این معنا که عضویت یک داده به چندین خوشه بسیار پر رنگ تر خواهد بود. مرز بین خوشه ها نرم تر و همپوشانی بیشتر خواهد شد.

اگر q به سمت بی نهایت میل کند، مقادیر عضویت برای همه خوشه ها تقریبا یکسان خواهد شد، به این معنی که فاصله عملا تاثیری ندارد و هر داده به همه خوشه ها به یک میزان مشابه عضویت دارد.

پس اگر پایین باشد، همپوشانی کمتر و hard خواهد بود. اگر زیاد باشد، همپوشانی بیشتر و مرز نرم تر خواهد بود و اگر خیلی بزرگ باشد عملا عضویت به یک میزان است.



سوال 9

الف)

ب)

مراکز خوشه ها در نهایت بیشتر احتمال دارد که در ناحیه ای که چگالی بیشتری دارد متمرکز میشوند، به چند دلیل:

الگوریتم k-mean و هر الگوریتمی که هدفشان کم کردن MSE است، چگالی تاثیر مستقیم بر آنها دارد. چونکه در نواحی که چگالی بالاتر است، داده ها نزدیک به یک دیگر هستند. قرار دادن مرکز یک خوشه در آن ناحیه باعث میشود که فواصل بین بسیاری از نقاط و مرکز کوچکتر شود در آن ناحیه، که این امر سبب میشود MSE به طور زیادی کاهش پیدا کند. اگر ناحیه ما چگالی پایین تری داشته باشد داده ها پراکنده هستند و مرکز یک خوشه فاصله بسیار زیادی تا نقاط دارد که باعث افزایش MSE میشود.

به همین دلیل k-mean به دنبال این است که MSE را کم کند به صورت طبیعی مراکز خوشه ها را در نواحی قرار میدهد که فاصله از بسیاری از نقاط به صورت همزمان کاهش پیدا کند، که به این معنی است که k-mean به سمت نواحی با چگالی بالاتر کشیده میشود.

پ)

مشکلی که در رابطه با k-mean عادی وجود دارد این است که خیلی به نقاط شروع حساس است و اگر نقاط شروع خوب توزیع نشده باشند ممکن است خوشه هایی که انتخاب میکند خیلی ضعیف و نامناسب باشند یا همگرایی کند صورت بگیرد. همچنین همانطور که اشاره شد k-mean نقاط ابتدایی مراکز را معمولا به صورت تصادفی از داده ها انتخاب میکند.

در نسخه تغییر داده شده k-mean که به k-mean++ معروف است، الگوریتم رویکرد بهتری را برای انتخاب اولیه مراکز در نظر میگیرد.

در این نسخه به سبب d^2، نقاطی که بسیار دورتر از مراکز فعلی ما هستند احتمال بیشتری برای انتخاب شدن به عنوان مراکز جدید را دارند. و توان 2 این تاثیر را بسیار تقویت میکند.

این نسخه به صورت کلی همگرایی سریعتری نسبت به نسخه معمولی k-mean دارد به سبب اینکه با انتخاب مراکز بهتری که خوب توزیع شده اند، خوشه های بهتری میسازد و iteration های کمتری برای رسیدن به همگرایی نیاز هست. همچنین در نتایج خوشه بندی هم بسیار خوب عمل میکنند و احتمال گیر کردن در local optima را کاهش میدهند.

اما سرعت این نسخه نسبت به نسخه اصلی k-mean کند تر است، به سبب محاسبه توزیع احتمالی در هر iteration.