فاز 3: عادی سازی و اولیه سازی

هدف: داده های خود را عادی کنید و یک روش اولیه سازی جایگزین برای k-means پیاده سازی کنید.

بخش اول – عادی سازی: در برنامه های کاربردی خوشه بندی، عادی سازی صفت یک مرحله پیش پردازش متداول است که برای جلوگیری از تسلط ویژگی های با دامنه های بزرگ بر محاسبات فاصله و جلوگیری از ناپایداری های عددی ضروری است. در این مرحله، روش "minmax normalization" را برای عادی سازی ویژگی های خود قبل از خوشه بندی پیاده سازی خواهید کرد. شما می توانید به راحتی شرح این روش عادی سازی ساده را در هر کتاب داده کاوی یا در WWW بیابید.

هشدار شماره 1: ماتریس داده های ورودی خود را در بین ردیف ها عادی نکنید. باید در سرتاسر ستون ها نرمال شود. به عبارت دیگر، هر یک از ویژگی های مجموعه داده ها باید مستقل از سایر ویژگی ها نرمال سازی شوند.

علاوه بر "نرمال سازی حداقل حداکثر"، روش "نرمال سازی امتیاز z" را نیز اجرا کنید.

نکته 1: هنگام اجرای هر یک از روش های عادی سازی، باید از تقسیم بر صفر اجتناب کنید.

قسمت دوم - مقداردهی اولیه: در فاز 2، الگوریتم استاندارد دسته ای k-means را پیاده سازی کردید. متأسفانه، k-means به مقداردهی اولیه بسیار حساس است. به عبارت دیگر، مراکز اولیه مختلف می توانند نتایج بسیار متفاوتی را به همراه داشته باشند. روش های اولیه متعددی برای رفع این مشکل پیشنهاد شده است. در این مرحله شما یک روش مقداردهی اولیه را برای kmeans پیاده سازی کرده و آن را با روش "انتخاب تصادفی" که در فاز 2 پیاده سازی کرده اید مقایسه می کنید.

Initial SSE: این مقدار SSE است که پس از مرحله اولیه سازی، قبل از مرحله خوشه بندی محاسبه می شود. این به خودی خود معیاری از اثربخشی یک روش اولیه سازی را به ما می دهد.

1. SSE نهایی: این مقدار SSE است که پس از مرحله خوشه بندی محاسبه می شود. هنگامی که خروجی آن توسط kmeans پالایش می شود، اندازه گیری اثربخشی روش مقداردهی اولیه را به ما می دهد. توجه داشته باشید که این تابع هدف الگوریتم k-means است.

2. تعداد تکرارها: این تعداد تکرارهایی است که k-means تا رسیدن به همگرایی زمانی که توسط یک روش اولیه سازی مقداردهی اولیه می شود، طول می کشد. این یک معیار کارایی مستقل از زبان برنامه نویسی، سبک پیاده سازی، کامپایلر و معماری CPU است.

در فاز 2، روش "انتخاب تصادفی" را اجرا کردید. در این مرحله روش پارتیشن تصادفی را پیاده سازی خواهید کرد. در این روش، با شروع از خوشه های خالی، ابتدا هر نقطه را به خوشه ای که به طور یکنواخت و به صورت تصادفی انتخاب شده است، اختصاص دهید. سپس مرکز این خوشه های اولیه را به عنوان مراکز اولیه در نظر می گیرید.

علاوه بر این، برای "پارتیشن تصادفی"، "روش maximin" را پیاده سازی کنید. این روش اولین مرکز را به طور دلخواه از بین نقاط داده انتخاب می کند و مراکز باقی مانده (K - 1) به صورت متوالی به شرح زیر انتخاب می شوند. در تکرار i (i = 2، 3، ...، K) ، مرکز i به عنوان نقطه داده ای با بیشترین مجذور فاصله اقلیدسی تا نزدیکترین مراکز انتخاب شده قبلی (i - 1) انتخاب می شود. به عبارت دیگر، مرکز 2 به عنوان دورترین نقطه داده از مرکز 1 انتخاب می شود. مرکز 3 به عنوان نقطه داده ای با بیشترین فاصله به نزدیکتر از مراکز 1 و 2 انتخاب می شود، یعنی نقطه داده x با بیشترین دقیقه. مقدار (d(x, c1), d(x, c2) که در آن c1 و c2 به ترتیب مرکز اول و دوم هستند، 'd' مجذور فاصله اقلیدسی و تابع "min(a, b)" است. کوچکتر «a» و «b» را برمی گرداند. مرکز i به عنوان نقطه داده x با بیشترین مقدار (d(x، (d(x، c1)، (d(x، c2)، ...، (ci-1) انتخاب می شود. با استفاده از حافظه اضافی، maximin را می توان در زمان O(NDK) پیاده سازی کرد، جایی که N، D و K به

ترتیب تعداد نقاط، ویژگی ها و خوشه ها را نشان می دهند. با این حال، یک پیاده سازی ساده به زمان O(N2DK) نیاز دارد. اگر به دنبال اجرای ساده‌ای هستید، در تکمیل اجراهای خود با مشکل مواجه خواهید شد.

نکته شماره 2: «خودسرانه» لزوماً به معنای «تصادفی» نیست. تصمیم تصادفی خودسرانه است، اما یک تصمیم خودسرانه لزوما تصادفی نیست.

نکته 3: Maximin با هر دو فاصله اقلیدسی (معمولی) و مربع اقلیدسی کار می کند. در هر دو مورد، نتیجه یکسان است. بنابراین، دلیل خوبی برای استفاده از فاصله اقلیدسی معمولی گران قیمت محاسباتی وجود ندارد.

اگر روش مقداردهی اولیه شما تصادفی است، آن را R بار اجرا کنید و بهترین نتیجه را برای هر اندازه‌گیری عملکرد جداگانه (بهترین SSEهای اولیه، بهترین SSEهای نهایی و بهترین # تکرار) برای هر مجموعه داده جدول‌بندی کنید. توجه داشته باشید که بهترین SSE اولیه، بهترین SSE نهایی و بهترین # تکرار ممکن است در اجراهای مختلف تولید شوند. به عبارت دیگر، اجرائی که بهترین SSE اولیه را می دهد، لزوماً اجرای بهترین SSE نهایی نخواهد بود. اگر روش مقداردهی اولیه شما قطعی (غیر تصادفی) باشد، از سوی دیگر، یک جدول ساده از مجموعه داده ها و اندازه گیری های مربوطه کافی خواهد بود. راه های زیادی برای جدول بندی این نوع داده ها وجود دارد. سعی کنید یک راه بصری پیدا کنید

خروجی: جدول(های) مایکروسافت اکسل یا Apache OpenOffice Calc اندازه گیری عملکرد برای هر ترکیبی از روش های عادی سازی و مقداردهی اولیه. به عبارت دیگر، {10 مجموعه داده} x {1 یا 2 روش عادی سازی} x {2 یا 3 روش اولیه} x {3 معیار عملکرد} = 60 یا 180 مقدار. البته می توانید این جدول را به جداول کوچکتر تقسیم کنید. به عنوان مثال، می توانید 6 یا 18 مقدار را در یک جدول قرار دهید که نشان دهنده یک مجموعه داده خاص است.

زبان: C، C++، یا جاوا. شما فقط می توانید از امکانات داخلی این زبان ها استفاده کنید. به عبارت دیگر، شما نمی توانید از کتابخانه های شخص ثالث استفاده کنید.

ارسال: کد منبع و فایل(های) خروجی خود را ارسال کنید