فاز 2: پایه K-Means

هدف: پیاده سازی الگوریتم k-means معمولی (دسته ای).

ورودی: برنامه شما باید غیر تعاملی باشد (یعنی برنامه نباید با پرسیدن سوالات صریح از کاربر با او تعامل داشته باشد) و آرگومان های خط فرمان زیر را در نظر بگیرد:

<F> <K> <I> <T> <R>، که در آن

F : نام فایل داده

K : تعداد خوشه ها (عدد صحیح مثبت بزرگتر از یک)

I : حداکثر تعداد تکرار (عدد صحیح مثبت)

T : آستانه همگرایی (واقعی غیر منفی)

R : تعداد اجراها (عدد صحیح مثبت)

هشدار شماره 0: هیچ یک از این پارامترها (به عنوان مثال، با استفاده از عبارتی مانند "int R = 100;") را در برنامه خود کدگذاری نکنید. مقادیر این پارامترها باید توسط کاربر در زمان اجرا از طریق خط فرمان داده شود. برای مثال زیر را ببینید.

"run" اجرای k-means است تا زمانی که همگرا شود. خط اول F شامل تعداد نقاط (N) و ابعاد هر نقطه (D) است. هر یک از خطوط بعدی حاوی یک نقطه داده در قالب خالی جدا شده است. در برنامه خود، باید ویژگی‌های هر نقطه را با استفاده از یک نوع داده ممیز شناور با دقت مضاعف نشان دهید (به عنوان مثال، نوع داده «دوگانه» در C/C++/Java). به عبارت دیگر، نباید از یک نوع داده انتگرال (بایت، کوتاه، char، int، طولانی و غیره) برای نمایش یک ویژگی استفاده کنید. همچنین نباید از نوع داده رشته ای استفاده کنید (char * در C/C++، std::string در C++، رشته در جاوا).

مراکز خوشه اولیه باید به طور یکنواخت و به صورت تصادفی از نقاط داده انتخاب شوند. یک اجرا باید زمانی خاتمه یابد که تعداد تکرارها به I برسد یا بهبود نسبی SSE (Sum-of-Squared Error) بین دو تکرار متوالی به زیر T برسد، یعنی هر زمان (SSE(t – 1) – SSE(t)) / SSE(t – 1) < T، که در آن SSE(t) مقدار SSE را در پایان تکرار t (t = 1, 2, ..., I) و SSE(0) = ∞ (بی نهایت) نشان می دهد.

هشدار شماره 1: الگوریتم اصلی k-means تابع ریشه مربع (یعنی "sqrt" در C/C++/Java) را شامل نمی شود. اگر از sqrt استفاده می کنید، ممکن است الگوریتم همگرا نشود. تابع دیگری که باید از آن اجتناب کرد، تابع توان است (یعنی "pow" در C/C++/Java). الگوریتم پایه k معنی نیازی به توان نمایی توسط یک توان غیر صحیح ندارد که عملیات نسبتاً کندی است. Kmeans در عوض به مربع کردن نیاز دارد که می‌تواند با استفاده از عملگر ضرب ('*') به طور موثر انجام شود. به طور خلاصه، برای مربع یک عدد x، از pow(x, 2) استفاده نکنید. به جای آن از x * x استفاده کنید.

هشدار شماره 2: اگر از تابع ریشه مربع یا تقریبی از آن در برنامه خود استفاده کنید، به طور خودکار نمره صفر دریافت خواهید کرد! متاسفانه در گذشته این اتفاق افتاده است. الگوریتم باید R بار اجرا شود، هر اجرا با مجموعه متفاوتی از مراکز به طور تصادفی انتخاب شده شروع می شود. خروجی: مقدار SSE را در پایان هر تکرار نمایش دهید.

نمونه اجرای برنامه:

% ecoli.txt = F (نام فایل داده)

% K = 8 (تعداد خوشه ها)

% I = 100 (حداکثر تعداد تکرار در یک اجرا)

% T = 0.000001 (آستانه همگرایی)

% R = 3 (تعداد اجراها)

% test = نام فایل اجرایی

> تست 3 0.000001 100 8 ecoli.txt

اجرا 1

تکرار 1: SSE = 22.0312

تکرار 2: SSE = 18.3704

تکرار 3: SSE = 17.5163

تکرار 4: SSE = 17.3435

تکرار 5: SSE = 17.2725

تکرار 6: SSE = 17.2331

تکرار 7: SSE = 17.221

تکرار 8: SSE = 17.2185

تکرار 9: SSE = 17.2175

تکرار 10: SSE = 17.2108

تکرار 11: SSE = 17.1934

تکرار 12: SSE = 17.184

تکرار 13: SSE = 17.182

تکرار 14: SSE = 17.1766

تکرار 15: SSE = 17.1639

تکرار 16: SSE = 17.1639

اجرا 2

تکرار 1: SSE = 18.5142

تکرار 2: SSE = 16.7108

تکرار 3: SSE = 16.07

تکرار 4: SSE = 15.6179

تکرار 5: SSE = 15.3449

تکرار 6: SSE = 15.1791

تکرار 7: SSE = 15.1201

تکرار 8: SSE = 15.0975

تکرار 9: SSE = 15.0624

تکرار 10: SSE = 15.0292

تکرار 11: SSE = 15.0162

تکرار 12: SSE = 15.0162

اجرا 3

تکرار 1: SSE = 18.7852

تکرار 2: SSE = 16.4215

تکرار 3: SSE = 16.2141

تکرار 4: SSE = 16.1159

تکرار 5: SSE = 16.1081

تکرار 6: SSE = 16.1045

تکرار 7: SSE = 16.0995

تکرار 8: SSE = 16.0838

تکرار 9: SSE = 16.0619

تکرار 10: SSE = 16.0547

تکرار 11: SSE = 16.0327

تکرار 12: SSE = 16.0109

تکرار 13: SSE = 16.0109

بهترین اجرا: 2: SSE = 15.0162

تست: برنامه خود را بر روی مجموعه داده های ارائه شده در جدول زیر تست کنید (با استفاده از پارامترهای ارائه شده در همان جدول). توجه به این نکته مهم است که در یک اجرای معین، با پیشرفت تکرارها، مقادیر SSE هرگز افزایش نمی‌یابد، یعنی $SSE(t) \leq SSE(t-1)$ برای $t = 1, 2, \ldots, I$. اگر مشاهده کنید دقیقاً همان مقدار SSE در دو تکرار متوالی (مانند "نمونه اجرای برنامه")، این بدان معنی است که الگوریتم همگرا شده است و کاهش بیشتری در SSE ممکن نیست. توجه داشته باشید که خروجی برنامه شما ممکن است در هر اجرا متفاوت باشد به دلیل مقداردهی اولیه تصادفی در مرکز.

نکته: در اینجا یک راه ساده برای یافتن اینکه آیا اجرای k-means شما احتمالاً درست است یا خیر، وجود دارد. در مجموعه داده Iris Bezdek ($K = 3$)، اگر $R = 100$ اجرا از k-means را اجرا کنید، این امکانات عبارتند از:

1. حداقل یک اجرا مقدار SSE کمتر از 78.8514 تولید می کند: برنامه شما قطعا باگ است. 78.8514 بهینه جهانی برای این مجموعه داده است (برای K = 3 خوشه). دریافت مقدار SSE کمتر از 78.8514 غیرممکن است مگر اینکه برنامه شما باگ باشد.

2. هر اجرا مقدار SSE بیشتر از 78.8514 تولید می کند: برنامه شما به احتمال زیاد باگ است. این یک مجموعه داده کوچک و ساده است، بنابراین شما باید بتوانید مقدار SSE تقریباً برابر با 78.8514 را در حداقل یک اجرا بدست آورید.

3. هیچ یک از اجراها مقدار SSE کمتر از 78.8514 تولید نمی کند و حداقل یک اجرا مقدار SSE تقریباً برابر با 78.8514 تولید می کند: برنامه شما به احتمال زیاد درست است.

Data Set	***K***	***I***	***T***	***R***
Ecoli	8	100	0.001	100
Glass	6	100	0.001	100
Ionosphere	2	100	0.001	100
Iris Bezdek	3	100	0.001	100
Landsat	6	100	0.001	100
Letter Recognition	26	100	0.001	100
Segmentation	7	100	0.001	100
Vehicle	4	100	0.001	100
Wine	3	100	0.001	100
Yeast	10	100	0.001	100

هشدار شماره 3: در "نمونه اجرای برنامه" بالا، ما فقط R = 3 اجرا را به دلیل کمبود فضا نمایش می دهیم. همانطور که جدول بالا نشان می دهد، شما باید R = 100 اجرا بر روی هر مجموعه داده انجام دهید.

مجموعه داده های بزرگ: هدف از مجموعه داده های بزرگ (به عنوان مثال، تشخیص حروف و لندست) آزمایش صبر شما نیست، بلکه ایجاد انگیزه در شما برای نوشتن برنامه های کارآمدتر است. K-means در واقع یک الگوریتم بسیار سریع است، اما فقط در صورتی که با دقت اجرا شود (یک پیاده سازی کارآمد می تواند میلیون ها نقطه را در عرض چند ثانیه در یک رایانه مدرن جمع کند.) اگر الگوریتم را به شیوه ای نامرتب کدنویسی کنید (مثلاً با استفاده از تابع pow). به جای عملگر ضرب)، برنامه شما ممکن است به شدت کند شود. زبان برنامه نویسی: 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 جواب: کد منبع و فایل های خروجی خود را ارسال کنید (برای هر فایل ورودی، یک فایل خروجی با فرمت TXT)