فاز 4: اعتبارسنجی داخلی

هدف: پیاده‌سازی روش‌های اعتبارسنجی داخلی برای تعیین خودکار تعداد خوشه‌ها در یک مجموعه داده معین.

شاید مهم‌ترین اشکال k-means این باشد که کاربر باید تعداد خوشه‌ها (K) را تامین کند. در بسیاری از کاربردها، این امر یا غیرممکن است یا غیرعملی.

در این مرحله، شما دو روش اعتبارسنجی داخلی را پیاده‌سازی می‌کنید که به k-means کمک می‌کند تا به‌طور خودکار تعداد خوشه‌ها را تعیین کند. یک روش اعتبارسنجی داخلی، تطابق بین یک پارتیشن به طور خودکار تولید شده از یک مجموعه داده و خود مجموعه داده را کمی می کند. اعتبار سنجی داخلی اغلب با استفاده از یک شاخص اعتبار داخلی انجام می شود، تابعی که یک پارتیشن، خود مجموعه داده و احتمالاً برخی پارامترهای اضافی را به عنوان ورودی می گیرد و یک مقدار عددی را نشان می دهد که کیفیت پارتیشن را به عنوان خروجی نشان می دهد.

در این مرحله، شاخص‌های اعتبار داخلی Calinski–Harabasz (CH) و Silhouette Width (SW) را پیاده‌سازی خواهید کرد. این شاخص ها در بسیاری از منابع توضیح داده شده است، به عنوان مثال به کتاب اخیر زکی و میرا جونیور مراجعه کنید: https://dataminingbook.info/book_html/

علاوه بر شاخص های CH و SW، شاخص Dunn (D) را نیز پیاده سازی کنید (برای توضیحات به کتاب فوق مراجعه کنید).

علاوه بر شاخص های CH، SW و D، شاخص Davies Bouldin (DB) را نیز پیاده سازی کنید (برای توضیحات، به کتاب فوق مراجعه کنید).

CH، SW، و D شاخص های بیشینه سازی هستند، در حالی که DB یک شاخص کمینه سازی است. نحوه استفاده از این شاخص ها بسیار ساده است. برای یک مجموعه داده معین، ابتدا Kmin و Kmax، حداقل و حداکثر تعداد خوشه‌هایی که ممکن است در مجموعه داده وجود داشته باشند را تعیین می‌کنیم. به عنوان مثال، برای مجموعه داده عنبیه، Kmin و Kmax به ترتیب می توانند 2 و 9 باشند. برای هر مجموعه داده، Kmin معمولاً 2 است، در حالی که، یک قانون کلی در مورد حداکثر مقدار ممکن Kmax نزدیکترین عدد صحیح به \sqrt{N}/2 است، که در آن N تعداد نقاط در مجموعه داده است. با فرض اینکه ما یک روش اولیه سازی تصادفی داریم، سپس k-means R بار برای K = 2 اجرا می کنیم و شاخص اعتبار داخلی (مثلاً CH) را روی پارتیشن تولید شده در بهترین اجرا محاسبه می کنیم (یعنی اجرایی که کوچکترین مقدار را تولید می کند. SSE). ما همین کار را برای K = 3، 4،...، 8، و در نهایت، K = 9 انجام می دهیم. از آنجایی که CH یک معیار بیشینه سازی است، سپس مقدار K را پیدا می کنیم که بیشترین مقدار CH را ایجاد می کند. این مقدار K تعداد "تخمینی" خوشه ها در مجموعه داده است. برای مثال عددی به مثال زیر و به مثال 17.8 در کتاب فوق الذکر مراجعه کنید. توجه داشته باشید که این شاخص ها همگی تخمین هایی را ارائه می دهند. بنابراین، ممکن است تعداد "درست" (3) خوشه‌ها را برای عنبیه (یا برای هر مجموعه داده دیگری) پیدا نکنید.

مثال عددی برای SW: مقادیر SW برای مجموعه داده عنبیه (فقط برای K = 2 و 3 خوشه - در آزمایش‌های شما، مقدار Kmax برای این مجموعه داده باید \sqrt{2/150}round{9} = { باشد) در زیر آورده شده. توجه داشته باشید که این فقط برای این است که به شما مثالی بزنم که چگونه مقادیر معقول SW باید به نظر برسد. ممکن است بسته به مقداردهی اولیه خود، اعداد دقیق یکسانی را دریافت نکنید. علاوه بر این، این اعداد بر روی مجموعه داده های عنبیه خام (غیر عادی) به دست می آیند. در آزمایشات خود، باید مجموعه داده ها را با استفاده از روش min-max نرمال کنید (فاز 3). 3 خوشه واقعی در عنبیه وجود دارد، اما دو تا از آن خوشه ها همپوشانی دارند، بنابراین K = 2 مقدار SW بسیار بهتری نسبت به K = 3 می دهد. بنابراین، اگر

بخواهیم فقط K = 2 و 3 خوشه را بر اساس شاخص SW امتحان کنیم. ، ما K = 2 را به عنوان بهترین تخمین برای تعداد خوشه ها در این مجموعه داده خاص انتخاب می کنیم. به خاطر داشته باشید که اگر مقادیر silhouette را خارج از محدوده [–1، 1] دریافت می‌کنید، حتماً مشکلی در برنامه شما وجود دارد.

test data/iris_bezdek.txt 2/.

تکرار 1: SSE = 539.413

تکرار 2: SSE = 366.075

تکرار 3: SSE = 237.899

تکرار 4: SSE = 175.767

تکرار 5: SSE = 154.339

تکرار 6: SSE = 152.513

تکرار 7: SSE = 152.348

SW(2) = 0.850351

test data/iris_bezdek.txt 3/.

تکرار 1: SSE = 230.163

تکرار 2: SSE = 81.9806

تکرار 3: SSE = 79.3943

تکرار 4: SSE = 78.9101

تکرار 5: SSE = 78.8514

SW(3) = 0.73566

نکته: ردیابی ماتریس پراکندگی درون خوشه ای، tr(Sw)، همانند SSE است که قبلاً با میانگین k محاسبه شده است. بنابراین، لازم نیست tr(Sw) را جداگانه محاسبه کنید. همچنین برای tr(Sb) نیازی به محاسبه کل ماتریس Sb نیست. ردیابی یک ماتریس مربع از مجموع عناصر مورب آن به دست می آید. بنابراین، تنها کاری که باید انجام دهید این است که عناصر مورب Sb را محاسبه کرده و سپس آنها را جمع کنید.

خروجی: جدول(های) مایکروسافت اکسل یا Apache OpenOffice Calc هر شاخص اعتبار داخلی برای مقادیر مختلف K برای هر مجموعه داده. برای هر مجموعه داده و شاخص اعتبار، تعداد تخمینی خوشه ها را به صورت پررنگ نشان دهید. بحث کنید که به نظر می رسد کدام شاخص تعداد خوشه ها را دقیق تر تخمین می زند. برای مقداردهی اولیه، 5372/4372 دانش‌آموز باید از روش «پارتیشن تصادفی» (فاز 3) (R = 100) استفاده کنند، در حالی که 6397 دانش‌آموز باید از روش «maximin» (فاز 3) (R = 100) استفاده کنند. همه مجموعه های داده باید با استفاده از روش "حداقل حداکثر" (فاز 3) نرمال سازی شوند.

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.