عمل الطالبة

خولة أحمد غيلان

0010_16_21

علوم حاسوب/ المستوى الرابع

التاريخ 29/ 1/ 2024

K-means

في خوارزمية K-means في مجال استخراج المعلومات(Data Mining) ، الـ K يمثل عدد العناصر أو العينات التي يتم تجميعها في مجموعات (clusters) مختلفة .تعتبر خوارزمية K-means من الخوارزميات الشهيرة في تجميع البيانات وتصنيفها.

عند تطبيق خوارزمية K-mean، يتم تحديد قيمة K بشكل مسبق قبل تشغيل الخوارزمية. يعتمد اختيار القيمة المناسبة لـ K على نوع وطبيعة البيانات المتاحة والهدف المرجو الوصول إليه.

هناك العديد من الطرق والمعايير التي يمكن استخدامها لتحديد قيمة K في خوارزميةK-means ، ومنها:

- المعرفة المسبقة: في بعض الحالات، يكون لدينا معرفة مسبقة عن عدد الفئات أو المجموعات المتوقعة في البيانات، في هذه الحالة، يمكن استخدام هذه المعرفة لتحديد قيمةK
- طريقة الكوع :(Elbow method) تعتمد هذه الطريقة على تجربة الخوارزمية لعدة قيم مختلفة لـ K وقياس مقياس الانحناء (inertia) أو مقياس الاختلاف (variance) لكل قيمة. يتم اختيار القيمة التي يكون فيها المقياس ثابتًا أو لا يتغير بشكل كبير بعد ذلك النقطة كأفضل قيمة لـK.
- المعاينة البصرية :(Visual inspection) يمكن تجربة الخوارزمية لعدة قيم مختلفة لـ K وفحص النتائج المتحققة بصريًا. يمكن ملاحظة الانسجام والتجزؤ في التجمعات واختيار القيمة التي تعطي أفضل تجزؤ وتجانس للتجمعات.
- استخدام معايير إحصائية أخرى: يمكن استخدام معايير إحصائية أخرى مثل معامل الارتباط (correlation coefficient) أو معامل الانكماش (shrinkage coefficient) لتحديد قيمة.K
- يجب ملاحظة أن اختيار قيمة K هو عملية متعددة الأبعاد وتعتمد على السياق والغرض من تطبيق خوارزمية K-means على البيانات المعينة. قد تتطلب هذه العملية تجربة ومقارنة عدة قيم للوصول إلى القيمة المناسبة لـK.

أمثلة على كيفية اختيار قيمة K باستخدام الطرق المذكورة:

- الاستخدام المسبق للمعرفة:
لنفترض أن لدينا مجموعة بيانات تحتوي على تفاصيل المنتجات في متجر. ونعلم أنه لدينا أربع فئات رئيسية من المنتجات: الملابس، الإلكترونيات، الأثاث، والأدوات المنزلية. في هذه الحالة، يمكننا استخدام المعرفة المسبقة لتحديد قيمة K كـ 4 لتكوين أربعة مجموعات تمثل هذه الفئات.

- طريقة الكوع:(Elbow method)
في هذا المثال، لنفترض أن لدينا مجموعة بيانات تحتوي على معلومات حول العملاء في متجر. نرغب في تجميع العملاء في مجموعات لتحليل السلوك الشرائي. نقوم بتشغيل خوارزمية K-means لقيم K تتراوح من 1 إلى 10 ونقيس مقياس الانحناء (inertia)لكل قيمة. الانحناء هو مجموع مربعات المسافة بين كل نقطة ومركز المجموعة الخاصة بها. نحصل على النتائج التالية:

K = 1: Inertia = 2560
K = 2: Inertia = 1800
K = 3: Inertia = 1200
K = 4: Inertia = 900
K = 5: Inertia = 800
K = 6: Inertia = 750
K = 7: Inertia = 720
K = 8: Inertia = 700
K = 9: Inertia = 680
K = 10: Inertia = 670

نلاحظ أن هناك تقلص كبير في الانحناء عند استخدام قيمة K تتراوح من 2 إلى 4، ولكن التحسن يصبح أقل واضحًا بعد ذلك. بالتالي، يمكننا اختيار قيمة K = 4 كأفضل قيمة تعكس الهيكل الرئيسي للبيانات.

- المعاينة البصرية:(Visual inspection)
في هذا المثال، لنفترض أن لدينا مجموعة بيانات ثنائية الأبعاد تحتوي على نقاط عشوائية. يمكننا تشغيل خوارزمية K-means لعدة قيم مختلفة لـ K وتصوير النتائج. بعد تجربة الخوارزمية لقيم K = 2 و K = 3 و K = 4 ، يمكننا ملاحظة النتائج وتحديد القيمة التي تعطي تجزؤًا واضحًا وفصلًا للنقاط في المجموعات.

- معامل الارتباط:(Correlation coefficient)
في هذا المثال، لنفترض أن لدينا مجموعة بيانات رقمية تحتوي على متغيرات متعددة .يمكننا حساب معامل الارتباط بين المتغيرات واستخدامه لتحديد قيمة K المناسبة. إذا كانت هناك قيمة K تؤدي إلى تجميع المتغيرات ذات الارتباط العالي في نفس المجموعة، فإن ذلك يشير إلى وجود تجزؤًا مناسبًا.

هذه هي بعض الطرق المشهورة لتحديد قيمة K في خوارزمية K-means. يمكن استخدام أي من هذه الطرق أو تجربة مجموعة متنوعة من القيم للحصول على نتائج مختلفة واختيار القيمة التي تناسب أفضل ظروف المشكلة المحددة.