یادگیری بدون نظارت

خوشه‌بندی

المپیاد هوش مصنوعی - یادگیری ماشین

۱ تعریف خوشه‌بندی

۲ کابردهای خوشه‌بندی

۳ دسته‌بندی الگوریتم‌های خوشه بندی

۴ روش افرازی

۵ روش سلسله مراتبی

۱ تعریف خوشه‌بندی
۲ کابردهای خوشه‌بندی
۳ دسته‌بندی الگوریتم‌های خوشه بندی
۴ روش افرازی
۵ روش سلسله مراتبی

خوشه‌بندی

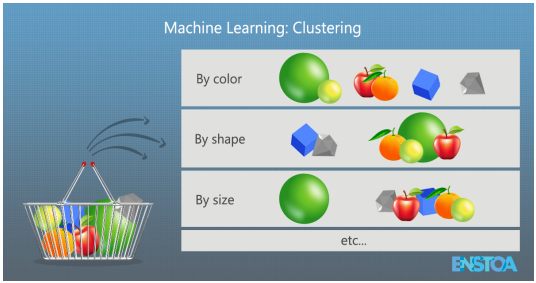
خوشه (Cluster) به گروهی از موجودیت‌ها گفته می‌شود که بیشترین شباهت را با یکدیگر دارند و از طرف دیگر بیشترین تفاوت را با موجودیت‌های قرار گرفته در خوشه‌های دیگر داشته باشند. شباهت‌های موجودیت‌ها بر اساس ویژگی‌هایی که دارند و میزان نزدیکی ارتباط‌شان با این ویژگی‌ها در مقایسه با سایر موجودیت‌ها تعیین می‌شود.

به عنوان مثال، فرض کنید دو نقطه در یک نمودار دو بعدی داریم. با استفاده از فاصله اقلیدسی، می‌توانیم میزان نزدیکی این دو نقطه را اندازه‌گیری کنیم. به همین ترتیب، با استفاده از معیارهای شباهت مختلف، می‌توانیم میزان نزدیکی یا شباهت نقاط داده را پیدا کنیم.

تمام نقاط داده مشابه، خوشه‌ها یا گروه‌ها را تشکیل می‌دهند. ایجاد این خوشه‌ها به روشی معنادار، خوشه‌بندی نامیده می‌شود.

خوشه‌بندی - ادامه

تصور کنید تعدادی شیء در یک سبد دارید. هر شکل مجموعه‌ای از ویژگی‌های مشخص (اندازه، شکل، رنگ و غیره) دارد. حال فرض کنید که از شما خواسته شده تا هریک از اشیاء را در سبد گروه‌بندی کنید. یک سوال طبیعی که باید بپرسید این است: "بر اساس چه معیاری این اشیاء را گروه‌بندی کنم؟" شاید بر اساس اندازه، شکل یا رنگ، شاید هم نه.

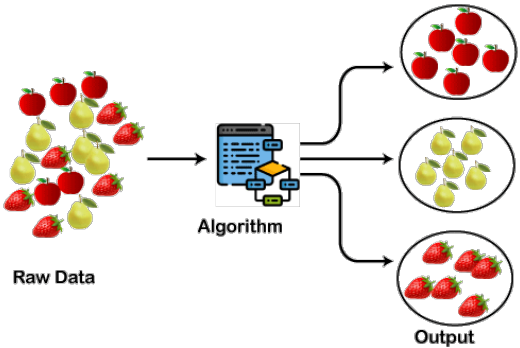


خوشه‌بندی - ادامه

پاسخی که انتخاب می‌کنید ممکن است به عوامل زیادی بستگی داشته باشد، از جمله تعداد اشیاء، میزان شباهت (یا تفاوت) آنها، یا حتی سنگینی برخی از اشیاء. اصلاً واضح نیست که چه معیاری برای گروه‌بندی باید انتخاب کنید. اغلب، بهترین پاسخ این است که اشیاء را بر اساس تمام ویژگی‌هایشان گروه‌بندی کنید. با بزرگتر شدن سبد و افزایش تعداد اشیاء، کار گروه‌بندی آنها به‌طور فزاینده‌ای پیچیده می‌شود - به همان اندازه‌ای که شیء در سبد وجود دارد، راه برای گروه‌بندی آنها وجود خواهد داشت!
به زبان ساده، الگوریتم‌های خوشه‌بندی ابزارهای مفیدی در مدیریت سبدهای دست و پاگیر هستند.

خوشه‌بندی - ادامه

نمودار زیر نحوه عملکرد الگوریتم خوشه‌بندی را توضیح می‌دهد. همان‌طور که می‌بینیم، میوه‌های مختلف بر اساس ویژگی‌های مشابه، به چند گروه تقسیم شده‌اند.

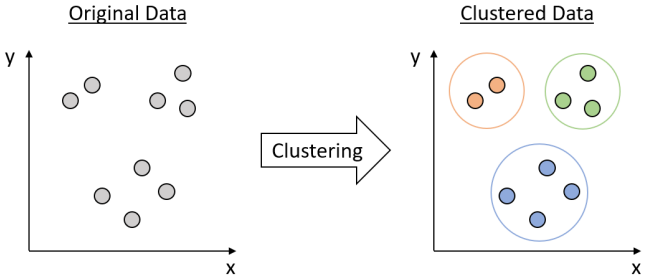


خوشه‌بندی - ادامه

در دنیای یادگیری ماشین، خوشه‌بندی فرآیندی است که در آن انبوهی از نقاط داده بدون برچسب را بر اساس ویژگی‌هایشان به خوشه‌ها تقسیم می‌کنیم.

- این تکنیک با پیدا کردن الگوهای مشابه مانند شکل، اندازه، رنگ یا رفتار در داده‌های بدون برچسب، آن‌ها را بر اساس وجود یا عدم وجود این الگوها دسته‌بندی می‌کند.
- از آنجایی که خوشه‌بندی یک روش یادگیری بدون نظارت است، هیچ اطلاعات از پیش تعیین‌شده‌ای در اختیار الگوریتم قرار نمی‌گیرد و الگوریتم مستقیماً با داده‌های بدون برچسب کار می‌کند.
- پس از انجام خوشه‌بندی، به هر خوشه یا گروه یک شناسه اختصاص داده می‌شود. سیستم‌های یادگیری ماشین می‌توانند از این شناسه‌ها برای آسان‌تر کردن پردازش مجموعه‌های داده‌های بزرگ و پیچیده استفاده کنند.

خوشه‌بندی - ادامه



هدف خوشه‌بندی

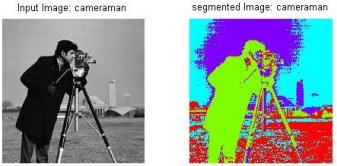
- مرحله پیش‌پردازش برای نمایه‌سازی، فشرده‌سازی یا کاهش داده‌ها
- نمایش داده‌های با ابعاد بالا در فضای با ابعاد پایین (به‌عنوان مثال، برای مقاصد بصری)
- کشف دانش از داده‌ها: به‌عنوان ابزاری برای درک ساختار پنهان در داده‌ها یا گروه‌بندی آن‌ها
- کسب بینش در مورد ساختار داده‌ها (قبل از طراحی طبقه‌بند)
- فراهم کردن اطلاعات در مورد ساختار داخلی داده‌ها
- برای گروه‌بندی یا تقسیم داده‌ها زمانی که برچسبی در دسترس نیست

۱ تعریف خوشه‌بندی
۲ کابردهای خوشه‌بندی
۳ دسته‌بندی الگوریتم‌های خوشه بندی
۴ روش افرازی
۵ روش سلسله مراتبی

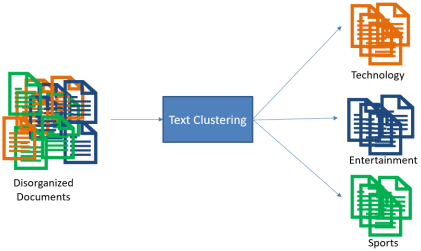
کاربردهای خوشه‌بندی

- دریافت اطلاعات (جستجو و مرور)
  - خوشه‌بندی مستندات متنی یا تصاویر بر اساس محتوای آن‌ها
  - خوشه‌بندی گروه‌های کاربران بر اساس الگوهای دسترسی آن‌ها به صفحات وب
- خوشه‌بندی کاربران شبکه‌های اجتماعی بر اساس علایق (شناسایی جامعه).
- بیوانفورماتیک
  - خوشه‌بندی پروتئین‌های مشابه با هم (شبیه‌سازی از نظر ساختار شیمیایی و/یا عملکرد و غیره) یا ژن‌های مشابه بر اساس داده‌های میکرواری
- تقسیم‌بندی بازار
  - خوشه‌بندی مشتریان بر اساس تاریخچه خرید و ویژگی‌های آن‌ها
- تقسیم‌بندی تصویر
- و بسیاری از کاربردهای دیگر

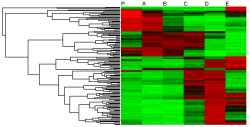
مثال‌هایی از کاربردهای خوشه‌بندی



(ب) تقسیم‌بندی تصویر



(آ) خوشه‌بندی مستندات متنی



(ج) خوشه‌بندی ژن‌های مشابه

۱ تعریف خوشه‌بندی

۲ کابردهای خوشه‌بندی

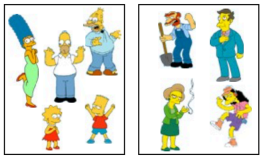
۳ دسته‌بندی الگوریتم‌های خوشه بندی

۴ روش افرازی

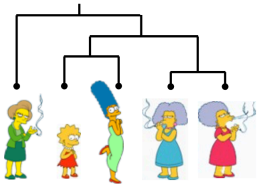
۵ روش سلسله مراتبی

الگوریتم‌های خوشه‌بندی

الگوریتم‌های خوشه‌بندی را می‌توان به دو دسته افرازی و سلسله‌مراتبی دسته‌بندی کرد.



(د) روش افرازی



(ه) روش سلسله‌مراتبی

۱ تعریف خوشه‌بندی
۲ کابردهای خوشه‌بندی
۳ دسته‌بندی الگوریتم‌های خوشه بندی
۴ روش افرازی
۵ روش سلسله مراتبی

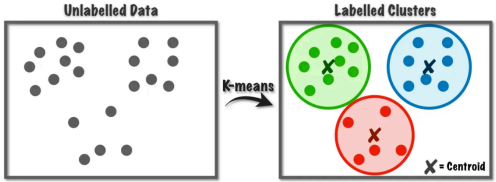
روش افرازی

در این روش، هدف این است که یک مجموعه داده را به K خوشه مختلف تقسیم کنیم، بطوریکه:

- هر داده به طور دقیق در یکی از K خوشه‌ای که با هم هیچ اشتراکی ندارند، قرار می‌گیرد.
- تمام داده‌های موجود در مجموعه داده باید در یکی از این K خوشه قرار بگیرند. به عبارت دیگر، اجتماع تمام خوشه‌ها برابر با کل مجموعه داده است.
- هیچ دو خوشه‌ای نباید عضو مشترک داشته باشند. این به این معنی است که هر داده فقط می‌تواند به یک خوشه تعلق داشته باشد. این نوع افراز بندی به عنوان خوشه‌بندی سخت نیز شناخته می‌شود، جایی که هر داده فقط به یک خوشه تعلق دارد.
- از آنجایی که خروجی این روش فقط یک مجموعه از خوشه‌ها است، کاربر باید از قبل **تعداد خوشه‌های مورد نظر** (K) را مشخص کند.

الگوریتم k-means

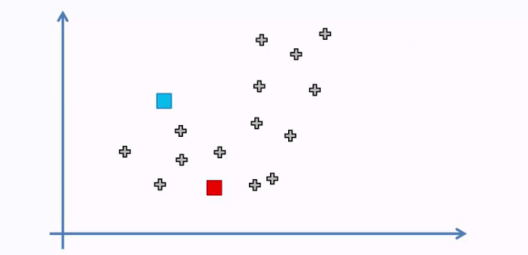
K-means یکی از محبوب‌ترین و پرکاربردترین الگوریتم‌ها در دسته‌ی الگوریتم‌های افرازی است.



الگوریتم K-means برای پردازش داده‌های آموزشی، کار خود را با یک گروه اولیه از مراکز خوشه آغاز می‌کند. این مراکز به عنوان نقاط شروع برای هر خوشه مورد استفاده قرار می‌گیرند و سپس الگوریتم محاسبات تکراری انجام می‌دهد تا موقعیت این مراکز را بهینه کند. در ادامه هریک از گام‌ها به ترتیب توضیح داده شده است.

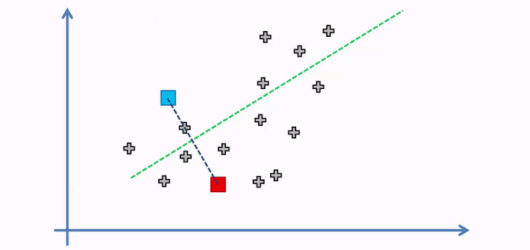
گام اول

ابتدا به صورت تصادفی تعدادی نقطه را به عنوان مرکز اولیه برای هر خوشه انتخاب می‌کنیم (تعداد این نقاط همان K است که از قبل مشخص کرده‌ایم). مثل اینکه چند میخ را به صورت تصادفی روی صفحه قرار دهیم.



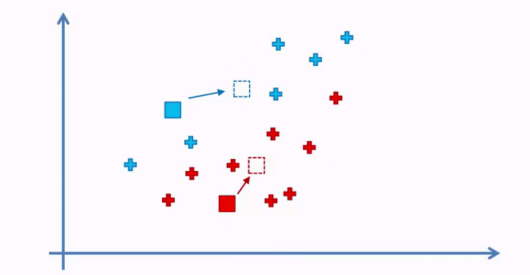
گام دوم

هر نقطه از داده‌ها را به نزدیک‌ترین مرکزی که انتخاب کرده‌ایم نسبت می‌دهیم. مثل اینکه با یک نخ، هر نقطه را به نزدیک‌ترین میخ وصل کنیم. به این ترتیب، نقاط در اطراف هر مرکز، یک خوشه را تشکیل می‌دهند.



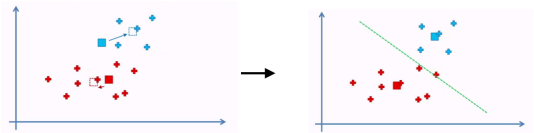
گام سوم

حالا برای هر خوشه، مرکز جدیدی را محاسبه می‌کنیم. این مرکز جدید، میانگین تمام نقاطی است که در آن خوشه قرار دارند. مثل اینکه جای میخ‌ها را به وسط مجموعه‌ای از نقاطی که به آن وصل شده‌اند، تغییر دهیم.



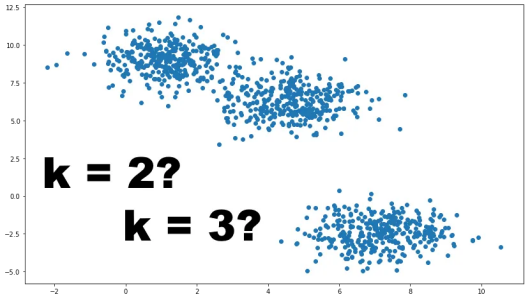
گام چهارم

گام‌های دوم و سوم را دوباره تکرار می‌کنیم. یعنی دوباره هر نقطه را به نزدیک‌ترین مرکز جدید نسبت می‌دهیم و سپس مراکز جدید را محاسبه می‌کنیم. این کار را آن‌قدر ادامه می‌دهیم تا دیگر جای مراکز تغییر نکند یا تغییرات بسیار کم شود.



روش انتخاب بهینه مقدار K

تصور کنید می‌خواهید یک دسته از افراد را در چند گروه دسته‌بندی کنید. هرچه تعداد گروه‌ها بیشتر باشد، افراد داخل هر گروه شبیه‌تر خواهند بود (خطای کمتری داریم). اما اگر تعداد گروه‌ها خیلی زیاد شود، ممکن است گروه‌های خیلی کوچکی تشکیل شوند که دیگر معنی خاصی ندارند. بنابراین در گام اول الگوریتم k-means از کجا مطمئن باشیم که چند خوشه باید تولید شود؟



نقطه آرنج

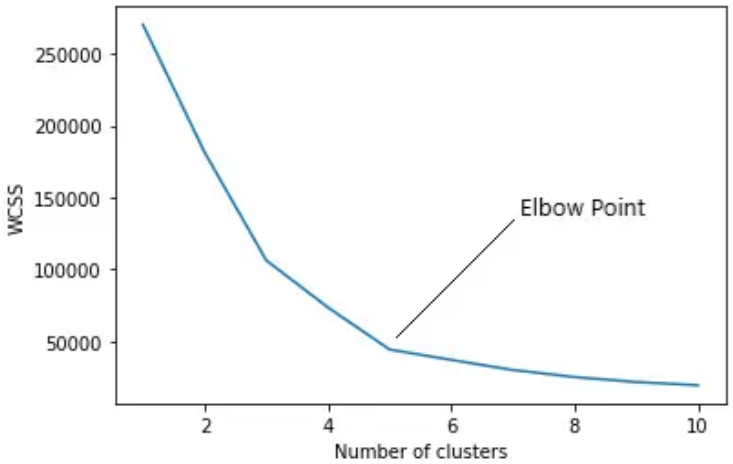
نقطه آرنج (Elbow Point) یک روش بصری است که برای پیدا کردن تعداد بهینه خوشه‌ها (مقدار مناسب K) در الگوریتم K-means استفاده می‌شود. به این صورت کار می‌کند:

- شما الگوریتم K-means را چندین بار اجرا می‌کنید، هر بار با تعداد خوشه‌های مختلف (مثلاً از ۱ تا ۱۰).
- برای هر تعداد خوشه‌ای که امتحان می‌کنید، یک معیار به نام مجموع مربعات خطا را محاسبه می‌کنید. این معیار نشان می‌دهد که چقدر نقاط داده در داخل هر خوشه به مرکز آن خوشه نزدیک هستند. هرچه این مقدار کمتر باشد، خوشه‌بندی بهتر انجام شده است.
- حالا شما یک نمودار رسم می‌کنید که در محور افقی آن تعداد خوشه‌ها (K) و در محور عمودی آن مقدار SSE قرار دارد.

نقطه آرنج

- پیدا کردن آرنج: در این نمودار، معمولاً یک منحنی می‌بینید که با افزایش تعداد خوشه‌ها، مقدار SSE کاهش پیدا می‌کند. در ابتدا، این کاهش خیلی سریع است، اما بعد از یک نقطه، شیب منحنی کم می‌شود و کاهش SSE کندتر می‌گردد. به نقطه‌ای در این نمودار که این تغییر شیب ناگهانی رخ می‌دهد، "نقطه آرنج" می‌گویند، چون شکل نمودار در آن نقطه شبیه آرنج خم‌شده به نظر می‌رسد.
- انتخاب K بهینه: مقدار K در نقطه آرنج، معمولاً به عنوان تعداد بهینه خوشه‌ها در نظر گرفته می‌شود. دلیلش این است که قبل از این نقطه، افزایش تعداد خوشه‌ها باعث کاهش چشمگیر خطا می‌شود، اما بعد از این نقطه، افزودن خوشه‌های بیشتر تاثیر چندانی در کاهش خطا ندارد و ممکن است باعث ایجاد خوشه‌هایی شود که خیلی معنی‌دار نیستند.

نقطه آرنج


250000
200000
150000
100000
50000
WCSS
Elbow Point
2
4
6
8
10
Number of clusters

۱ تعریف خوشه‌بندی

۲ کابردهای خوشه‌بندی

۳ دسته‌بندی الگوریتم‌های خوشه بندی

۴ روش افرازی

۵ روش سلسله مراتبی

روش سلسله مراتبی

خوشه‌بندی سلسله مراتبی یک روش دیگر برای گروه‌بندی داده‌ها است که برخلاف روش افرازی، نیازی به تعیین تعداد خوشه‌ها از قبل ندارد. در عوض، این روش یک ساختار درختی از خوشه‌ها ایجاد می‌کند که نشان می‌دهد خوشه‌ها چگونه با یکدیگر ارتباط دارند.

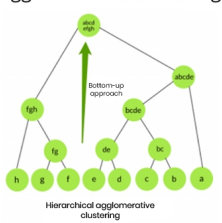
تصور کنید می‌خواهید یک دسته از اشیاء را بر اساس شباهت‌هایشان دسته‌بندی کنید، مثل دسته‌بندی جانوران. شما ممکن است ابتدا آن‌ها را به گروه‌های بزرگتر مثل پستانداران، پرندگان، خزندگان و غیره تقسیم کنید.

سپس هر کدام از این گروه‌ها را به گروه‌های کوچکتر تقسیم کنید، مثلاً پستانداران را به سگ‌سانان، گربه‌سانان و غیره. این یک نوع ساختار سلسله مراتبی است.

روش سلسله مراتبی

روش ادغام‌شونده

روش ادغام شونده (Agglomerative Clustering) از پایین به بالا کار می‌کند. یعنی ابتدا هر داده به عنوان یک خوشه جداگانه در نظر گرفته می‌شود. سپس، در هر مرحله، نزدیک‌ترین خوشه‌ها با هم ادغام می‌شوند تا اینکه در نهایت همه داده‌ها در یک خوشه بزرگ قرار بگیرند. مثل این است که شما از تک‌تک برگ‌های یک درخت شروع کنید و آن‌ها را به شاخه‌های کوچک، سپس به شاخه‌های بزرگ‌تر و در نهایت به تنه درخت وصل کنید.



روش تقسیم شونده

روش تقسیم‌شونده (Divisive Clustering) از بالا به پایین کار می‌کند. یعنی ابتدا همه داده‌ها در یک خوشه بزرگ قرار دارند. سپس، در هر مرحله، این خوشه به خوشه‌های کوچکتر تقسیم می‌شود تا اینکه هر داده به یک خوشه جداگانه تبدیل شود. مثل این است که شما از کل یک درخت شروع کنید و آن را به شاخه‌های اصلی، سپس به شاخه‌های کوچکتر و در نهایت به برگ‌ها تقسیم کنید.