استاد: دکتر فاطمه منصوری

نیمسال دوم ۱۴۰۳-۱۴۰۲

# خوشه‌بندی

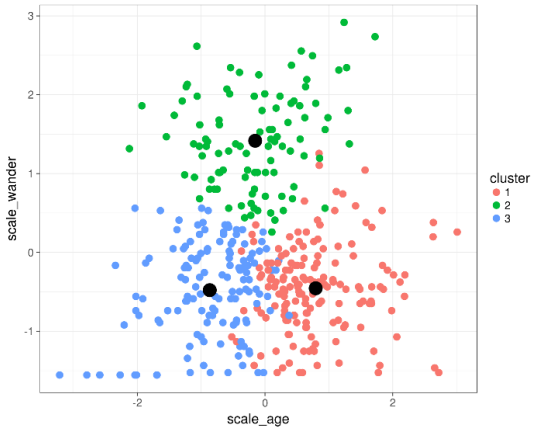
یادگیری عمیق

گردآورنده: ثنا سعیدمهر

## تعریف K-means

K-means یک الگوریتم خوشه‌بندی[1] محبوب است که در یادگیری ماشینی بدون نظارت[2] استفاده می‌شود. هدف آن تقسیم یک مجموعه داده به خوشه‌هایی به تعداد K است که در آن هر نقطه داده متعلق به خوشه با نزدیکترین میانگین است. "K" در K-means به تعداد خوشه‌هایی که می‌خواهید ایجاد کنید اشاره دارد و یک فراپارامتر[3] است که باید مشخص کنید.

(خوشه: گروهی از نقاط داده که مشابه یکدیگر هستند.)




شکل ۱ : نمایی از خوشه‌بندی داده‌ها به همراه مرکز آن‌ها

## نحوه عملکرد الگوریتم K-means

هدف K-means به حداقل رساندن مجموع فواصل مجذور بین نقاط داده و مرکز خوشه مربوطه آنها است.

[1] Clustering
[2] Unsupervised learning
[3] Hyperparameter

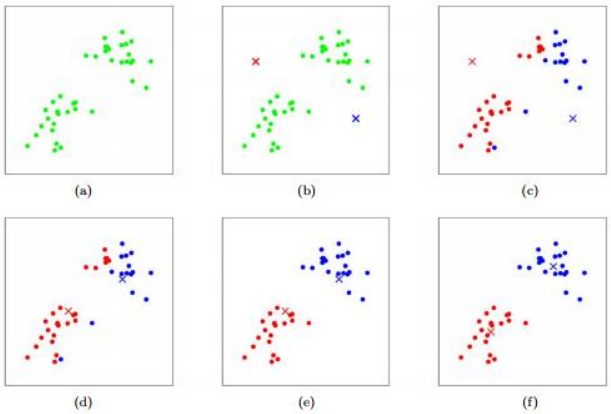
۱. مقداردهی اولیه: تعداد خوشه‌ها را انتخاب کنید. با مقداردهی اولیه تصادفی مرکزهای K خوشه شروع کنید. این مرکزها می توانند به صورت تصادفی از نقاط داده انتخاب شوند.

۲. اختصاص نقاط داده به خوشه‌ها: برای هر نقطه داده، فاصله هر مرکز را محاسبه کنید و نقطه را به خوشه‌ای با نزدیکترین مرکز اختصاص دهید. (معمولاً از فاصله اقلیدسی استفاده می‌شود، اما سایر معیارهای فاصله مانند فاصله منهتن یا شباهت کسینوس نیز می‌توانند استفاده شوند.)

۳. به‌روزرسانی مراکز خوشه‌ها: هنگامی که تمام نقاط داده به خوشه‌ها اختصاص داده شد، مراکز خوشه‌ها را با در نظر گرفتن میانگین تمام نقاط داده اختصاص داده شده به هر خوشه دوباره محاسبه کنید.

۴. تکرار: به طور مکرر فرآیند ۲ و ۳ را تا زمان همگرایی تکرار کنید. همگرایی زمانی اتفاق می‌افتد که تخصیص‌های خوشه دیگر تغییر قابل توجهی نداشته باشند، یا به حداکثر تعداد تکرار رسیده باشند.

۵. نهایی‌سازی: وقتی الگوریتم همگرا می‌شود، شما K خوشه را دارید که هر نقطه داده با خوشه خود مرتبط است.

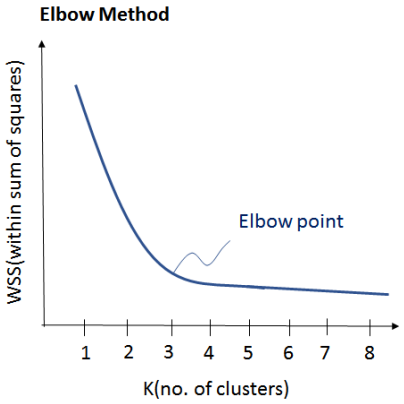



شکل ۲ : نحوه عملکرد الگوریتم K-means

### انتخاب تعداد خوشه‌ها در K-means

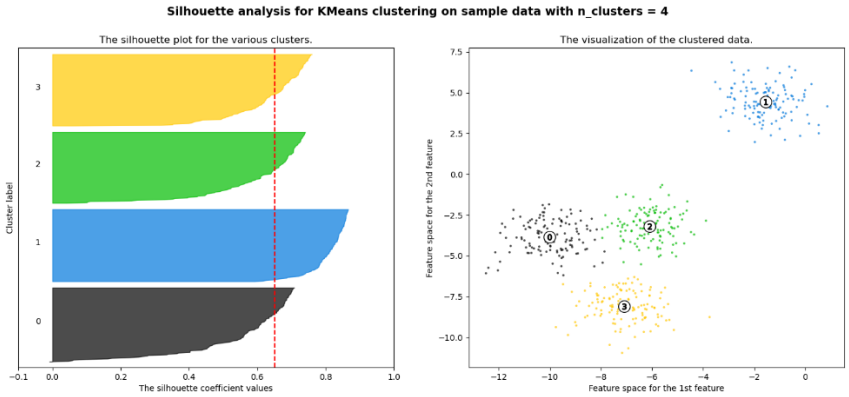
**۱. روش آرنج[4]:** مانند تلاش برای یافتن آرنج در نموداری از خوشه‌ها در مقابل مجموع مجذور فاصله نقاط تا مراکز خوشه اختصاص داده شده آن‌هاست. با افزایش تعداد خوشه‌ها، این مجموع کاهش می‌یابد زیرا نقاط به مراکز خوشه خود نزدیکتر هستند. با این حال، در برخی موارد، افزودن خوشه‌های بیشتر منجر به بهبود قابل توجهی نمی‌شود. این نقطه اغلب به عنوان "آرنج" نامیده می‌شود. تعداد خوشه‌ها را درست قبل از شروع بازده‌های کاهشی انتخاب می‌شود. (رایج‌ترین روش انتخابی)

[4] Elbow Method




شکل ۳ : انتخاب تعداد خوشه‌ها با استفاده از روش آرنج

**۲. امتیاز سیلوئت**[۵]: معیاری است که نشان می‌دهد یک نقطه داده در داخل خوشه (انسجام[۶]) در مقایسه با سایر خوشه‌ها چقدر شبیه است (جداسازی[۷]). محدوده مقدار سیلوئت بین +۱ و -۱ است. ضرایب سیلوئت نزدیک به +۱ نشان می‌دهد که نمونه از خوشه‌های همسایه دور است. مقدار ۰ نشان می‌دهد که نمونه روی مرز تصمیم بین دو خوشه همسایه یا بسیار نزدیک به آن است و مقادیر منفی نشان می‌دهد که آن نمونه‌ها ممکن است به خوشه اشتباهی اختصاص داده شده باشند. نمودار سیلوئت اندازه‌گیری نزدیکی هر نقطه در یک خوشه به نقاط در خوشه‌های همسایه را نشان می‌دهد و بنابراین راهی برای ارزیابی بصری پارامترهایی مانند تعداد خوشه‌ها ارائه می‌دهد.




شکل ۴ : انتخاب تعداد خوشه‌ها با استفاده از امتیاز سیلوئت

[5] Silhouette Score
[6] cohesion
[7] separation

**چند کاربرد K-means**

۱. تقسیم‌بندی مشتری: کسب و کارها از K-means برای تقسیم‌بندی مشتریان بر اساس رفتارهای خرید، ترجیحات یا جمعیت شناسی آن‌ها استفاده می‌کنند. این تقسیم‌بندی به تنظیم استراتژی‌های بازاریابی برای گروه‌های مختلف مشتریان کمک می‌کند.

۲. تقسیم‌بندی تصویر: در پردازش تصویر، از K-means می‌توان برای تقسیم‌بندی تصاویر به مناطق مجزا بر اساس شدت پیکسل یا شباهت رنگ استفاده کرد. این در کارهایی مانند تشخیص اشیا و فشرده‌سازی تصویر مفید است.

۳. تشخیص ناهنجاری: K-means می تواند نقاط پرت یا ناهنجاری در داده‌ها را با خوشه‌بندی نقاطی که به طور قابل توجهی از بقیه انحراف دارند، شناسایی کند. ناهنجاری ها اغلب به خوشه‌هایی با نقاط داده کمتر ختم می‌شوند.

۴. خوشه‌بندی اسناد: در پردازش زبان طبیعی، K-means برای خوشه‌بندی اسناد بر اساس محتوا یا شباهت آن‌ها استفاده می‌شود. این می‌تواند به سازماندهی و دسته‌بندی مجموعه داده‌های متنی بزرگ کمک کند.

۵. تجزیه و تحلیل بازار سهام: تحلیلگران مالی از K-means برای خوشه‌بندی سهام بر اساس حرکات تاریخی قیمت یا معیارهای مالی استفاده می‌کنند. این خوشه‌بندی می‌تواند به بینش‌هایی در مورد شباهت‌های بین سهام یا پرتفوی منجر شود.

۶. خوشه‌بندی ژنتیکی: در ژنتیک، K-means می‌تواند برای خوشه‌بندی ژن‌ها یا داده‌های ژنتیکی برای شناسایی الگوهای مربوط به عملکردها یا بیماری های بیولوژیکی خاص استفاده شود.

۷. سیستم‌های توصیه: تجارت الکترونیک و پلت‌فرم‌های آنلاین از K-means برای گروه‌بندی کاربران با اولویت‌های مشابه استفاده می‌کنند و توصیه‌های شخصی‌شده را بر اساس رفتار گذشته امکان‌پذیر می‌سازند.

۸. تجزیه و تحلیل شبکه اجتماعی: K-means را می‌توان برای خوشه‌بندی افراد در یک شبکه اجتماعی بر اساس ارتباطات یا تعاملات آن‌ها اعمال کرد. این می‌تواند جوامع یا کاربران با نفوذ در شبکه را نشان دهد.

۹. تشخیص کانون زلزله: دانشمندان زمین شناسی از K-means برای تجزیه و تحلیل داده‌های لرزه‌ای و تشخیص کانون‌های زلزله بر اساس شباهت امواج لرزه‌ای استفاده می‌کنند.

۱۰. بهینه‌سازی شبکه حسگر بی‌سیم: K-means به بهینه‌سازی شبکه‌های حسگر بی‌سیم با خوشه‌بندی گره‌های حسگر بر اساس مکان یا عملکردشان کمک می‌کند و کارایی ارتباط را بهبود می‌بخشد.

**مزایا K-means**

۱. سهولت اجرا: پیاده‌سازی K-means نسبتاً ساده و قابل درک است.

۲. کارایی: برای مجموعه داده‌های بزرگ کارآمد است.

۳. مقیاس پذیری: K-means می‌تواند مجموعه داده‌های بزرگ را به طور موثر مدیریت کند و آن را برای خوشه‌بندی برنامه‌های کاربردی با مقدار قابل توجهی داده مطابقت پیدا می‌کند.

۴. تفسیرپذیری: خوشه‌های تشکیل شده توسط K-means را می‌توان به راحتی تفسیر و تجسم کرد و آن را برای تجزیه و تحلیل داده‌های اکتشافی مفید می‌کند.

۵. تطبیق پذیری: K-means را می‌توان برای انواع مختلف داده‌ها و برنامه‌ها، از تقسیم بندی مشتری گرفته تا فشرده‌سازی تصویر، تطبیق داد.

## معایب K-means

۱. حساس به انتخاب اولیه مرکز: عملکرد K-means می‌تواند به شدت به قرارگیری اولیه مرکزهای خوشه وابسته باشد. انتخاب مرکز اولیه ضعیف می‌تواند به راه‌حل‌های غیربهینه منجر شود.

۲. فرض شکل خوشه: K-means فرض می‌کند که خوشه‌ها کروی و با اندازه مشابه هستند، که ممکن است برای همه مجموعه‌های داده صادق نباشد. این می‌تواند منجر به تخصیص خوشه‌های بهینه و عملکرد ضعیف در داده‌های غیر کروی شود.

۳. انتخاب تعداد خوشه‌ها: انتخاب تعداد مناسب خوشه‌ها (مقدار K) می‌تواند یک کار ذهنی و چالش برانگیز باشد.

۴. تأثیر نقاط پرت: نقاط پرت می‌توانند به طور قابل توجهی بر مرکزهای خوشه تأثیر بگذارند و منجر به تخصیص خوشه‌های کمتر معنی‌دار شوند.

۵. نتایج غیر قطعی: به دلیل مقداردهی اولیه تصادفی، K-means ممکن است نتایج متفاوتی را برای هر اجرا ایجاد کند. این تنوع گاهی اوقات می‌تواند بازتولید نتایج را دشوارتر کند.

## جایگزین K-means

خوشه بندی سلسله مراتبی[8]: سلسله مراتبی از خوشه‌ها را ایجاد می‌کند.

خوشه‌بندی فضایی مبتنی بر تراکم برنامه‌ها با نویز[9]: نقاط داده را بر اساس چگالی خوشه‌بندی می‌کند.

خوشه بندی طیفی[10]: از بردارهای ویژه یک ماتریس شباهت برای شناسایی خوشه‌ها استفاده می‌کند.

## نتیجه‌گیری

خوشه‌بندی K-means یک الگوریتم پرکاربرد برای تقسیم‌بندی داده‌ها به خوشه‌ها بر اساس شباهت است. درک اصول، نقاط قوت و ضعف می‌تواند به شما کمک کند تا آن را به طور موثر در مسائل مختلف دنیای واقعی به کار ببرید.

---

[8] Hierarchical Clustering
[9] Density-Based Spatial Clustering of Applications with Noise
[10] Spectral Clustering

## منابع


K-means Clustering

K-means Algorithm and Implementation

Determine the Optimal K for K-means

Clustering Algorithms Instead of K-means Clustering

Matrix methods in data mining Book