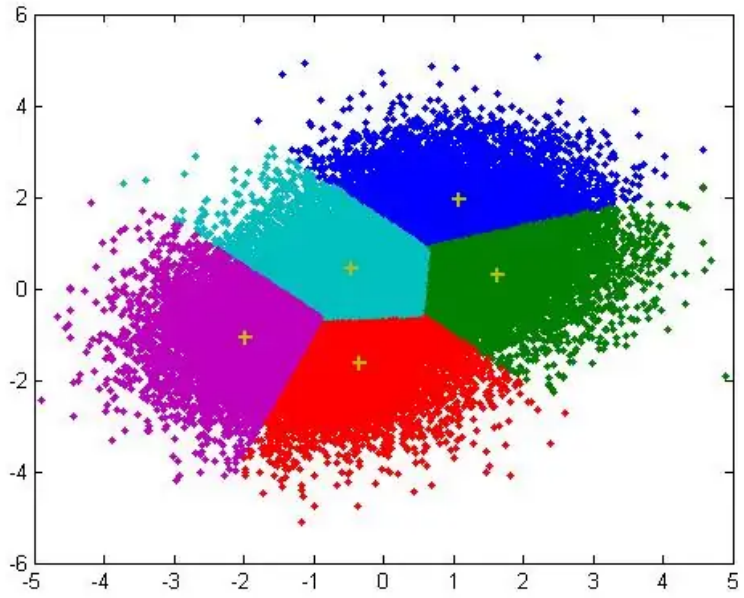
۱ خوشه‌بندی k-means

روش‌ها و الگوریتم‌های متعددی برای تبدیل اشیاء به گروه‌های همشکل یا مشابه وجود دارد. الگوریتم k-means یا میانگین-k یکی از ساده‌ترین و محبوب‌ترین الگوریتم‌هایی است که در «داده‌کاوی» (Data Mining) بخصوص در حوزه «یادگیری نظارت نشده» (Unsupervised Learning) به کار می‌رود.

معمولا در حالت چند متغیره، باید از ویژگی‌های مختلف اشیا به منظور طبقه‌بندی و خوشه کردن آن‌ها استفاده کرد. به این ترتیب با داده‌های چند بعدی سروکار داریم که معمولا به هر بعد از آن، ویژگی یا خصوصیت گفته می‌شود. با توجه به این موضوع، استفاده از توابع فاصله مختلف در این جا مطرح می‌شود. ممکن است بعضی از ویژگی‌های اشیا کمی و بعضی دیگر کیفی باشند. به هر حال آنچه اهمیت دارد روشی برای اندازه‌گیری میزان شباهت یا عدم شباهت بین اشیاء است که باید در روش‌های خوشه‌بندی لحاظ شود.

الگوریتم خوشه‌بندی میانگین-k از گروه روش‌های خوشه‌بندی تفکیکی (Partitioning Clustering) می‌شود و درجه پیچیدگی محاسباتی آن برابر با $O(n^{dk+1})$ است، به شرطی که n تعداد اشیاء، d بعد ویژگی‌ها و k تعداد خوشه‌ها باشد. همچنین پیچیدگی زمانی برای این الگوریتم برابر با $O(nkdi)$ است، که البته منظور از i تعداد تکرارهای الگوریتم برای رسیدن به جواب بهینه است.

در خوشه‌بندی میانگین-k از بهینه‌سازی یک تابع هدف (Object Function) استفاده می‌شود. پاسخ‌های حاصل از خوشه‌بندی در این روش، ممکن است به کمک کمینه‌سازی (Minimization) یا بیشینه‌سازی (Maximization) تابع هدف صورت گیرد. به این معنی که اگر ملاک «میزان فاصله» (Distance Measure) بین اشیاء باشد، تابع هدف براساس کمینه‌سازی خواهد بود پاسخ عملیات خوشه‌بندی، پیدا کردن خوشه‌هایی است که فاصله بین اشیاء هر خوشه کمینه باشد. در مقابل، اگر از تابع مشابهت (Dissimilarity Function) برای اندازه‌گیری مشابهت اشیاء استفاده شود، تابع هدف را طوری انتخاب می‌کنند که پاسخ خوشه‌بندی مقدار آن را در هر خوشه بیشینه کند.



معمولا زمانی که هدف کمینه‌سازی باشد، تابع هدف را «تابع هزینه» (Cost Function) نیز می‌نامند.

روش خوشه بندی میانگین-k ، توسط «مک‌کوئین» (McQueen) جامعه شناس و ریاضیدان در سال ۱۹۶۵ ابداع و توسط

دیگر دانشمندان توسعه و بهینه شد. برای مثال در سال ۱۹۵۷ نسخه دیگری از این الگوریتم به عنوان الگوریتم استاندارد خوشه‌بندی میانگین-k ، توسط «لوید» (Lloyd) در آزمایشگاه‌های بل (Bell Labs) برای کدگذاری پالس‌ها ایجاد شد که بعدها در سال ۱۹۸۲ منتشر گردید. این نسخه از الگوریتم خوشه‌بندی، امروزه در بیشتر نرم‌افزارهای رایانه‌ای که عمل خوشه‌بندی میانگین-k را انجام می‌دهند به صورت استاندارد اجرا می‌شود. در سال ۱۹۵۶ «فورجی» (W.Forgy) به طور مستقل همین روش را ارائه کرد و به همین علت گاهی این الگوریتم را با نام لوید-فورجی می‌شناسند. همچنین روش هارتیگان- ونگ (Hartigan-Wong) که در سال ۱۹۷۹ معرفی شد یکی از روش‌هایی است که در تحقیقات و بررسی‌های داده‌کاوی مورد استفاده قرار می‌گیرد. تفاوت در این الگوریتم‌ها در مرحله آغازین و شرط همگرایی الگوریتم‌ها است ولی در بقیه مراحل و محاسبات مانند یکدیگر عمل می‌کنند. به همین علت همگی را الگوریتم‌های خوشه‌بندی میانگین-k می‌نامند.

۱.۱ روش خوشه‌بندی میانگین-k

فرض کنید مشاهدات $(x_1, x_2, \ldots, x_n)$ که دارای d بعد هستند را باید به k بخش یا خوشه تقسیم کنیم. این بخش‌ها یا خوشه‌ها را با مجموعه‌ای به نام $S = \{S_1, S_2, , S_k\}$ می‌شناسیم. اعضای خوشه‌ها باید به شکلی از مشاهدات انتخاب شوند که تابع «مجموع مربعات درون خوشه‌ها» (within-cluster sum of squares-WCSS) که در حالت یک بعدی شبیه واریانس است، کمینه شود.

بنابراین، تابع هدف در این الگوریتم به صورت زیر نوشته می‌شود.

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

در اینجا منظور از μ_i میانگین خوشه S_i و $|S_i|$ تعداد اعضای خوشه iام است. البته می‌توان نشان داد که کمینه کردن این مقدار به معنی بیشینه‌سازی میانگین مربعات فاصله بین نقاط در خوشه‌های مختلف (between-Cluster sum of Squares-BCSS) است زیرا طبق قانون واریانس کل، با کم شدن مقدار WCSS ، مقدار BCSS افزایش می‌یابد، زیرا واریانس کل ثابت است.

در ادامه به بررسی روش خوشه بندی میانگین-k به روش لوید-فورجی (استاندارد) و هارتیگان-ونگ می‌پردازیم.

۲.۱ خوشه‌بندی میانگین-k با الگوریتم لوید (Lloyd's Algorithm)

به عنوان یک الگوریتم استاندارد برای خوشه‌بندی میانگین-k از الگوریتم لوید بخصوص در زمینه علوم کامپیوتر، استفاده می‌شود. ابتدا به علائمی که در این رابطه به کار می‌رود، اشاره می‌کنیم.

$m_j^{(i)}$: میانگین مقدارهای مربوط به خوشه jام در تکرار iام از الگوریتم را با این نماد نشان می‌دهیم.

$S_j^{(i)}$: مجموعه اعضای خوشه j ام در تکرار iام الگوریتم.

الگوریتم لوید را با توجه به نمادهای بالا می‌توان به دو بخش تفکیک کرد. ۱- بخش مقدار دهی (Assignment Step)، ۲- بخش به روز رسانی (Update Step). حال به بررسی مراحل اجرای این الگوریتم می‌پردازیم. در اینجا فرض بر این است که نقاط مرکزی اولیه یعنی $m_1^{(1)}, m_2^{(1)}, \cdots, m_k^{(1)}$ داده شده‌اند.

۱. بخش مقدار دهی: هر مشاهده یا شی را به نزدیکترین خوشه نسبت می‌دهیم. به این معنی که فاصله اقلیدسی هر مشاهده از مراکز، اندازه گرفته شده سپس آن مشاهده عضو خوشه‌ای خواهد شد که کمترین فاصله اقلیدسی را با مرکز آن خوشه دارد. این قانون را به زبان ریاضی به صورت $S_i^{(t)} = \left\{x_p : \left\|x_p - m_i^{(t)}\right\|^2 \leq \left\|x_p - m_j^{(t)}\right\|^2 \ \forall j, 1 \leq j \leq k\right\}$ می‌نویسیم.

۲. بخش به روز رسانی: میانگین خوشه‌های جدید محاسبه می‌شود. در این حالت داریم: $m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$

توجه داشته باشید که منظور از $|S_i^{(t)}|$ تعداد اعضای خوشه iام است. الگوریتم زمانی متوقف می‌شود که مقدار برچسب عضویت مشاهدات تغییری نکند. البته در چنین حالتی هیچ تضمینی برای رسیدن به جواب بهینه (با کمترین مقدار برای تابع هزینه) وجود ندارد. کاملا مشخص است که در رابطه بالا، فاصله اقلیدسی بین هر نقطه و مرکز خوشه ملاک قرار گرفته است. از این جهت از میانگین و فاصله اقلیدسی استفاده شده که مجموع فاصله اقلیدسی نقاط از میانگینشان کمترین مقدار ممکن نسبت به هر نقطه دیگر است.

نکته: ممکن است فاصله اقلیدسی یک مشاهده از دو مرکز یا بیشتر، برابر باشد ولی در این حالت آن شئ فقط به یکی از این خوشه‌ها تعلق خواهد گرفت.

نکته: به نقاط مرکزی هر خوشه مرکز (Centroid) گفته می‌شود. ممکن است این نقطه یکی از مشاهدات یا غیر از آن‌ها باشد. مشخص است که در الگوریتم لوید، k مشاهده به عنوان مرکز خوشه‌ها (Centroids) در مرحله اول انتخاب شده‌اند ولی در مراحل بعدی، مقدار میانگین هر خوشه نقش مرکز را بازی می‌کند.

۳.۱ خوشه‌بندی میانگین-k با الگوریتم هارتیگان-ونگ (Hartigan-Wong)

یکی از روش‌های پیشرفته و البته با هزینه محاسباتی زیاد در خوشه‌بندی میانگین-k ، الگوریتم هارتیگان-ونگ است. برای آشنایی با این الگوریتم بهتر است ابتدا در مورد نمادهایی که در ادامه خواهید دید توضیحی ارائه شود.

$\phi(S_j)$: از این نماد برای نمایش «تابع هزینه» برای خوشه S_j استفاده می‌کنیم. این تابع در خوشه‌بندی میانگین-k برابر است با:

$$\phi(S_i) = \sum_{x \in S_j} (x - \mu_j)^۲$$

S_j : از آنجایی که هدف از این الگوریتم، تفکیک اشیاء به k گروه مختلف است، گروه‌ها یا خوشه‌ها در مجموعه‌ای با نام S قرار دارند و داریم، $S = \{S_۱, S_۲, \cdots, S_k\}$.

μ_j : برای نمایش میانگین خوشه jام از این نماد استفاده می‌شود. بنابراین خواهیم داشت:

$$\mu_j = \frac{\sum_{x \in S_j} x}{n_j}$$

n_j : این نماد تعداد اعضای خوشه jام را نشان می‌دهد. بطوری که $j = \{۱, ۲, \cdots, k\}$ است. البته مشخص است که در اینجا تعداد خوشه‌ها را با k نشان داده‌ایم.

۴.۱ مراحل اجرای الگوریتم

در خوشه‌بندی میانگین-k با الگوریتم هارتیگان می‌توان مراحل اجرا را به سه بخش تقسیم کرد: ۱- بخش مقدار دهی اولیه (Assignment Step) ، ۲- بخش به روز رسانی (Update Step) ، ۳- بخش نهایی (Termination) . در ادامه به بررسی این بخش‌ها پرداخته می‌شود.

۱. بخش مقدار دهی اولیه: در الگوریتم هارتیگان-ونگ، ابتدا مشاهدات و یا اشیاء به طور تصادفی به k گروه یا خوشه تقسیم می‌شوند. به این کار مجموعه S با اعضایی به صورت $\{S_j\}_{j \in \{i, \cdots, k\}}$ مشخص می‌شود.

۲. بخش به روز رسانی: فرض کنید که مقدارهای n و m از اعداد ۱ تا k انتخاب شده باشد. مشاهده یا شیئ از خوشه n ام را در نظر بگیرید که تابع $\Delta(m, n, x) = \phi(S_n) + \phi(S_m) - \phi(S_n \setminus \{x\}) - \phi(S_m \cup \{x\})$ را کمینه سازد، در چنین حالتی مقدار x از خوشه nام به خوشه mام منتقل می‌شود. به این ترتیب شی مورد نظر در S_m قرار گرفته و خواهیم داشت $x \in S_m$.

۳. بخش نهایی: زمانی که به ازای همه n, m, x مقدار $\Delta(m, n, x)$ کوچکتر از صفر باشد، الگوریتم خاتمه می‌یابد.

نکته: منظور از نماد $\phi(S_n \setminus \{x\})$ محاسبه تابع هزینه در زمانی است که مشاهده x از مجموعه S_n خارج شده باشد. همچنین نماد $\phi(S_m \cup \{x\})$ به معنی محاسبه تابع هزینه در زمانی است که مشاهده x به خوشه S_m اضافه شده باشد.

در تصویر زیر مراحل اجرای الگوریتم هارتیگان به خوبی نمایش داده شده است. هر تصویر بیانگر یک مرحله از اجرای الگوریتم است. نقاط رنگی نمایش داده شده، همان مشاهدات هستند. هر رنگ نیز بیانگر یک خوشه است.
در تصویر اول مشخص است که در بخش اول از الگوریتم به طور تصادفی خوشه‌بندی صورت پذیرفته. ولی در مراحل بعدی خوشه‌ها اصلاح شده و در انتها به نظر می‌رسد که بهترین تفکیک برای مشاهدات رسیده‌ایم.
در تصویر آخر نیز مشخص است که مراکز خوشه‌ها، محاسبه و ثابت شده و دیگر بهینه‌سازی صورت نخواهد گرفت. به این ترتیب پاسخ‌های الگوریتم با طی تکرار ۵ مرحله به همگرایی می‌رسد.

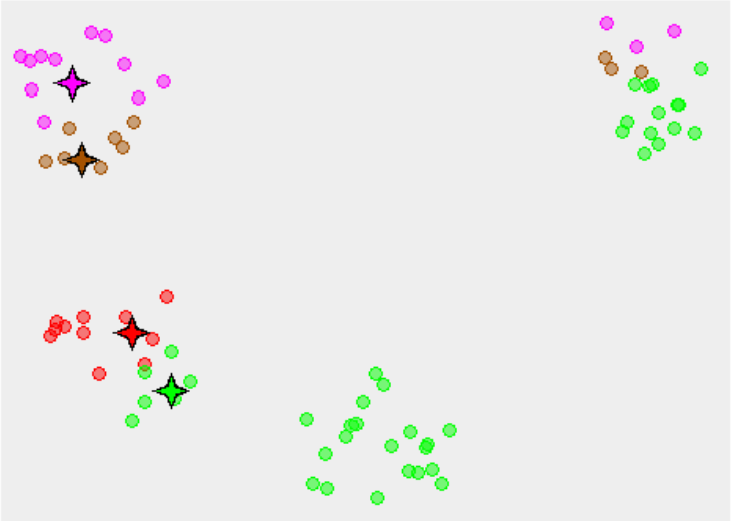


Figure ۱: الگوریتم هارتیگان بخش مقدار دهی اولیه

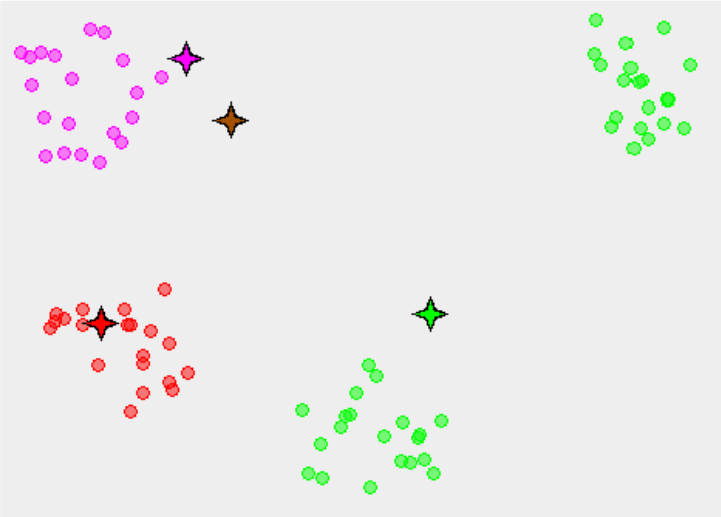


Figure ۲: الگوریتم هارتیگان تکرار ۱

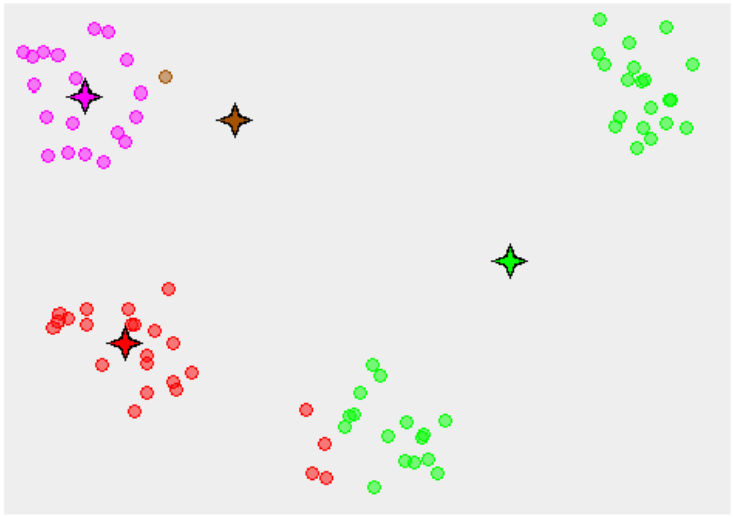


Figure ۳: الگوریتم هارتیگان تکرار ۲

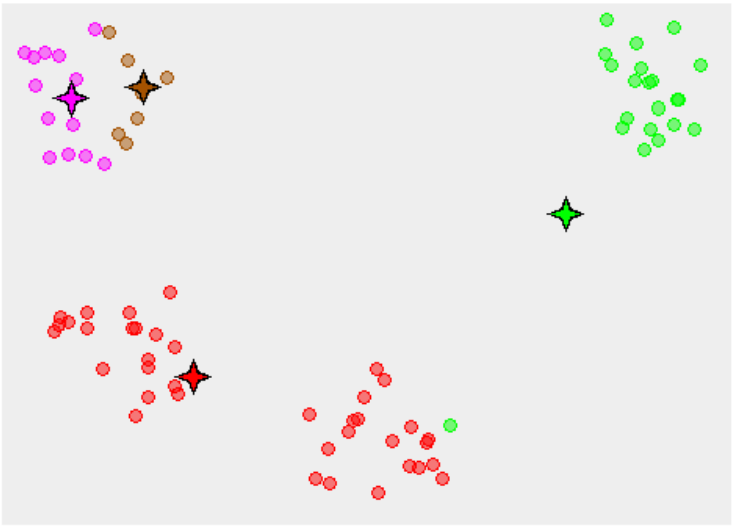


Figure ۴: الگوریتم هارتیگان تکرار ۳

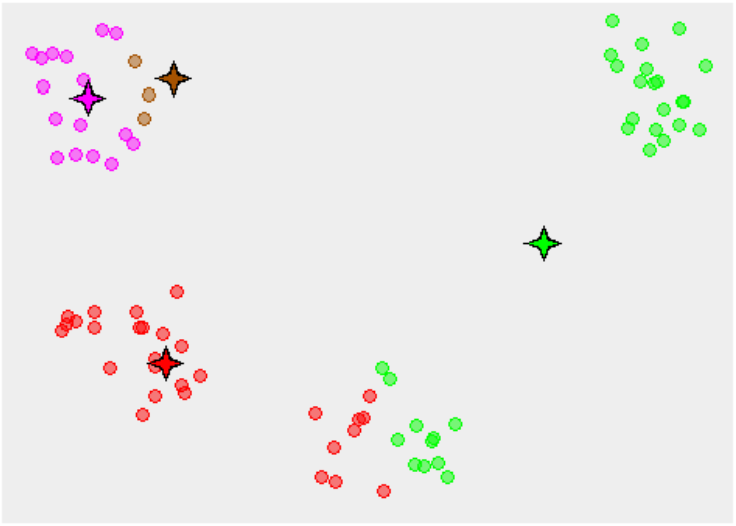


Figure ۵: الگوریتم هارتیگان تکرار ۴

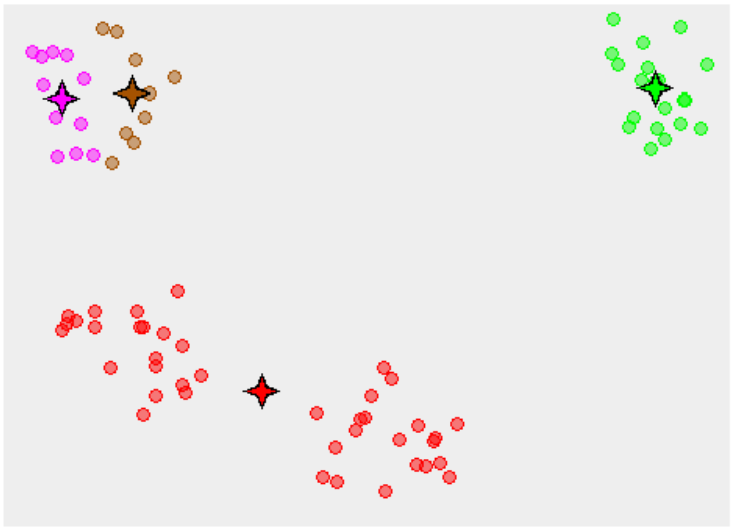


Figure ۶: الگوریتم هارتیگان تکرار ۵