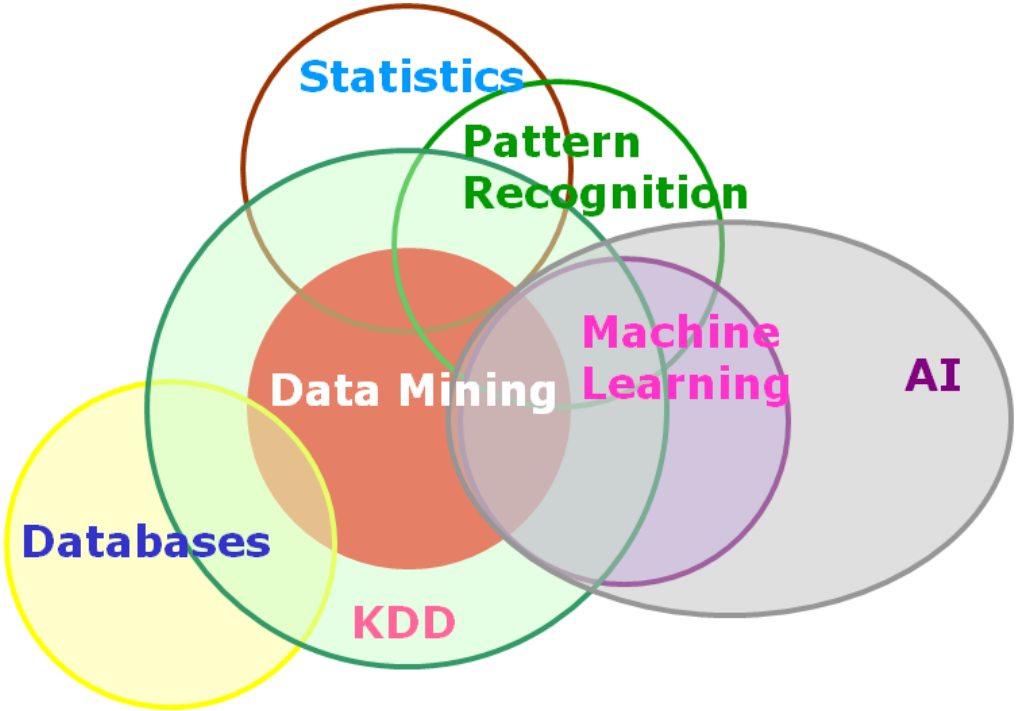
کشف دانش و داده‌کاوی

کشف دانش و داده‌کاوی

با رشد انفجاری داده‌ها و در دسترس بودن گسترده آن‌ها، نیاز مبرم به ابزاری قدرتمند برای کشف اطلاعات ارزشمند و تبدیل داده‌ها به دانش سازمان یافته، بیش از پیش احساس می‌شود.

داده‌کاوی به‌عنوان علمی نو و پویا، به مجموعه فرآیندهایی گفته می‌شود که در آن از حجم عظیم داده‌ها، دانش استخراج می‌شود.

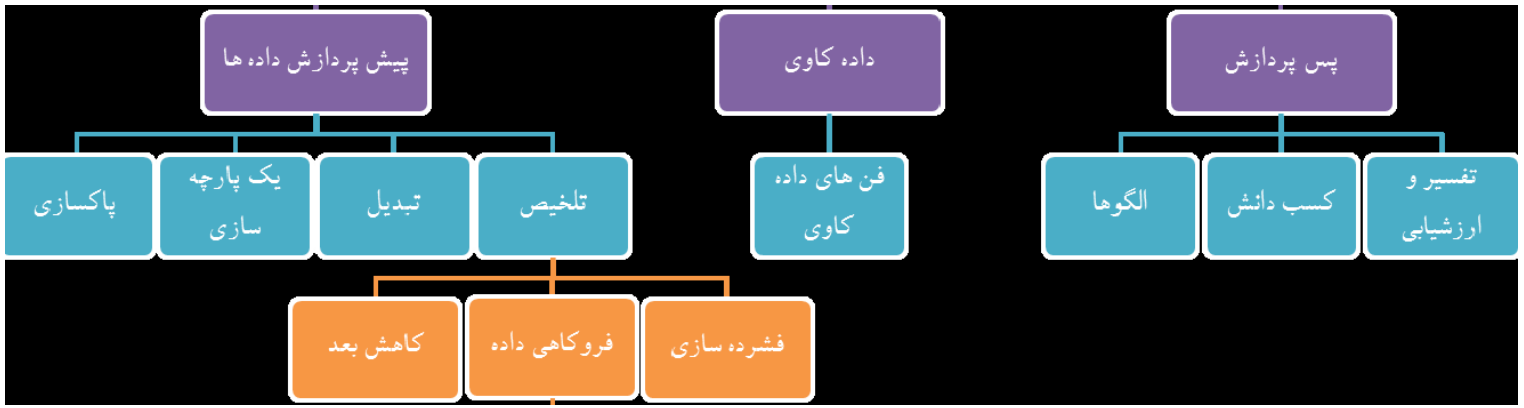
زمینه‌های علمی مرتبط با کشف دانش و داده‌کاوی



۲۱

[۱]KDD: Knowledge Discovery from Data.
[۲]AI: Artificial Intelligence.

مراحل کشف دانش



فن‌های داده‌کاوی

خوشه‌بندی (Clustering)

رده‌بندی (Classification)

قواعد پیوند (Association rules)

دیداری‌سازی (Visualization)

خوشه‌بندی

خوشه‌بندی

تعریف

خوشه‌بندی، یک روش توصیفی و غیرنظارتی است که در آن خوشه‌هایی شامل داده‌ها ایجاد می‌شوند به‌طوری که داده‌های یک خوشه بسیار مشابه بوده و داده‌های موجود در خوشه‌های متفاوت متمایز هستند.

معادل‌های کلمه خوشه‌بندی

Data Clustering- Clustering- Cluster Analysis- Segmentation Analysis- Taxonomy Analysis

کاربردهای خوشه‌بندی

کاربردهای خوشه‌بندی

۱) رتبه‌بندی شرکت‌ها براساس صورت‌های مالی، احتمال نکول و ...

۲) خوشه‌بندی شهرهای امریکا بر اساس جرم و جنایت

City	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto Theft
Atlanta	16.5	24.8	106	147	1112	905	494
Boston	4.2	13.3	122	90	982	669	954
Chicago	11.6	24.7	340	242	808	609	645
Dallas	18.1	34.2	184	293	1668	901	602
Denver	6.9	41.5	173	191	1534	1368	780
Detroit	13.0	35.7	477	220	1566	1183	788
Hartford	2.5	8.8	68	103	1017	724	468
Honolulu	3.6	12.7	42	28	1457	1102	637
Houston	16.8	26.6	289	186	1509	787	697
Kansas City	10.8	43.2	255	226	1494	955	765
Los Angeles	9.7	51.8	286	355	1902	1386	862
New Orleans	10.3	39.7	266	283	1056	1036	776
New York	9.4	19.4	522	267	1674	1392	848
Portland	5.0	23.0	157	144	1530	1281	488
Tucson	5.1	22.9	85	148	1206	756	483
Washington	12.5	27.6	524	217	1496	1003	793

روش‌های خوشه‌بندی

خوشه‌بندی شامل دو روش عمده است:

۱) سلسله مراتبی،

۲) غیرسلسله مراتبی.

روش غیرسلسله مراتبی شامل سه روش است:

۲٫۱) افرازی، (kمیانگین)

۲٫۲) مدل مبنا یا مبتنی بر مدل،

۲٫۳) مبتنی بر چگالی.

k میانگین

kمیانگین

kمیانگین، یکی از ساده‌ترین و معروف‌ترین الگوریتم‌های یادگیری غیرنظارتی است. این روش از الگوریتم‌های سلسله مراتبی کارآمدتر و سریع‌تر است. این روش برای خوشه‌بندی داده‌های کمی چندمتغیره استفاده می‌شود و هر داده فقط به یک خوشه تعلق می‌گیرد.

به طور کلی هدف این روش کمینه کردن معیار زیر است:

$$J = \sum_{j=۱}^{K} \sum_{i=۱}^{n} d(x_i^{(j)}, C_j)$$

الگوریتم kمیانگین

الگوریتم kمیانگین در حالت کلی به‌صورت زیر است:

۱) انتخاب تصادفی k نقطه به‌عنوان مرکز خوشه‌ها،

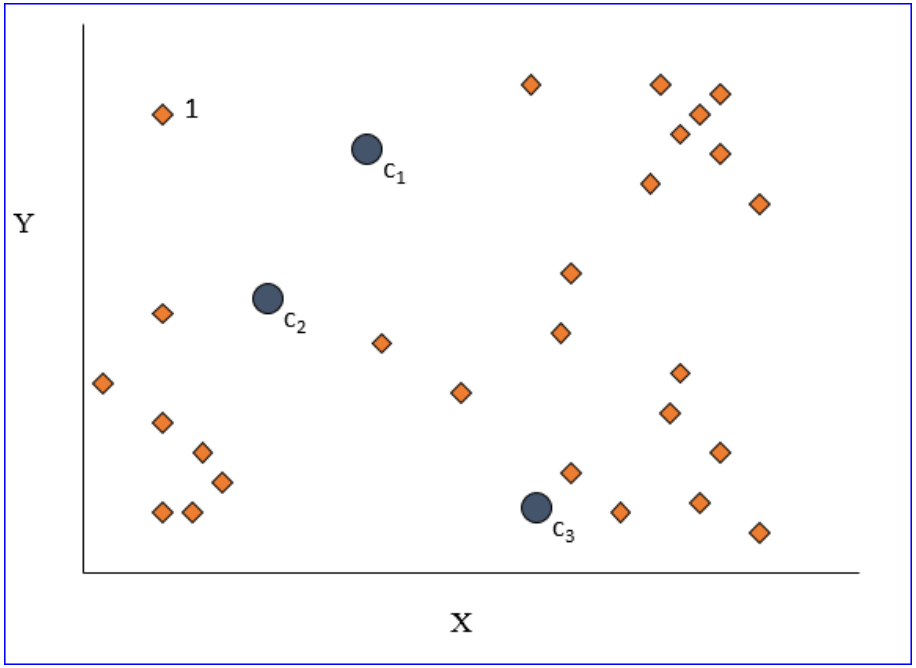
۲) نسبت دادن هر داده به یک خوشه که آن داده کمترین فاصله تا مرکز آن خوشه را دارا باشد،

۳) جایگزینی مرکز خوشه‌ها با میانگین داده‌های هر خوشه‌ها،

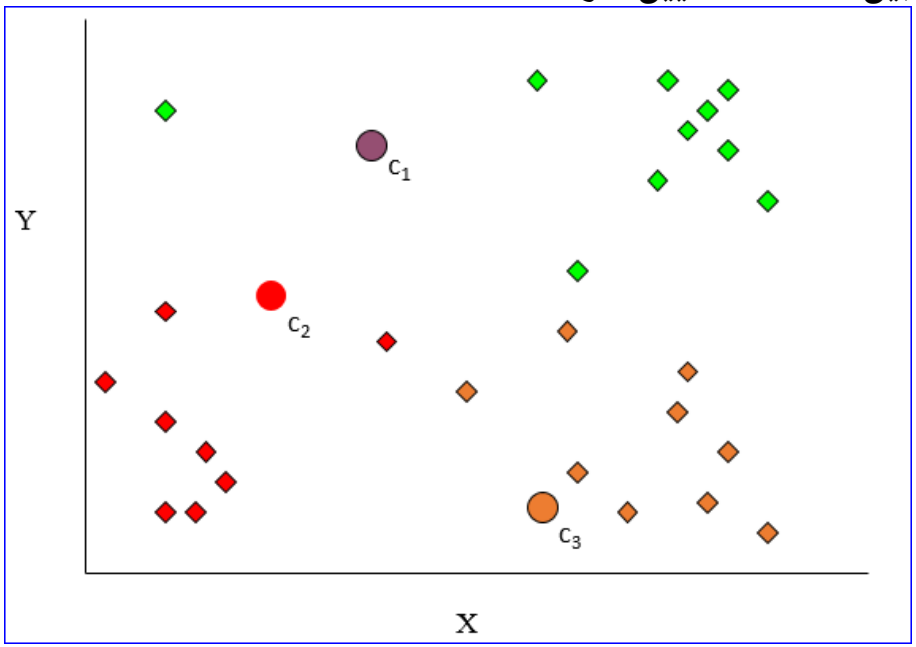
۴) تکرار مرحله ۲ و ۳ تا رسیدن به همگرایی. (تغییرات در خوشه‌ها ایجاد نشود)

الگوریتم kمیانگین

مرحله ۱: تعیین کردن k مرکز خوشه $(k = ۳)$

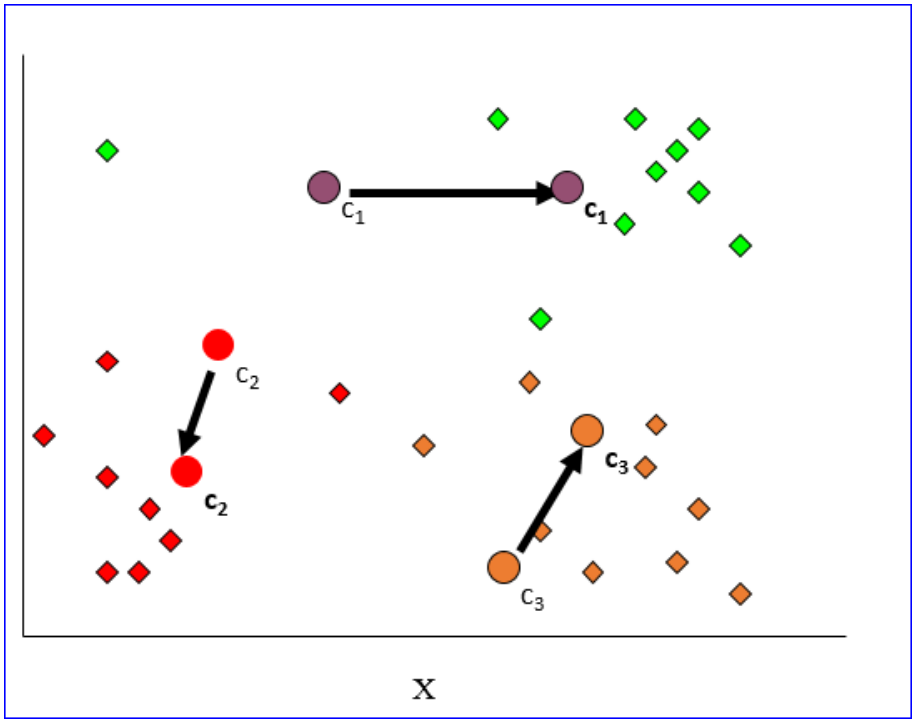


مرحله ۲: محاسبه فاصله‌های بین داده‌ها و تعیین خوشه داده‌ها

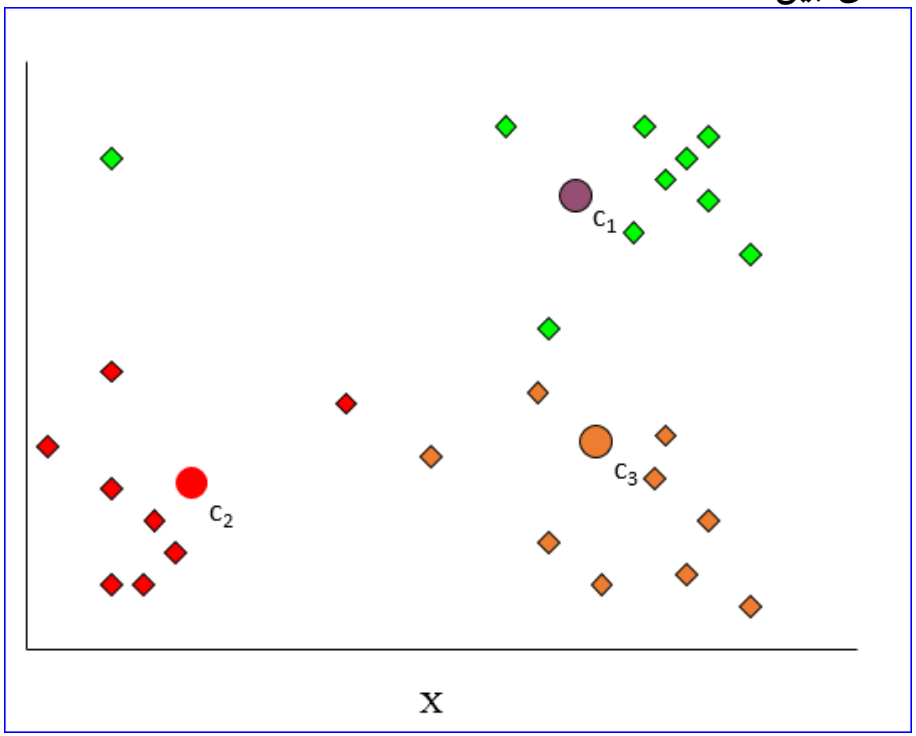


الگوریتم kمیانگین

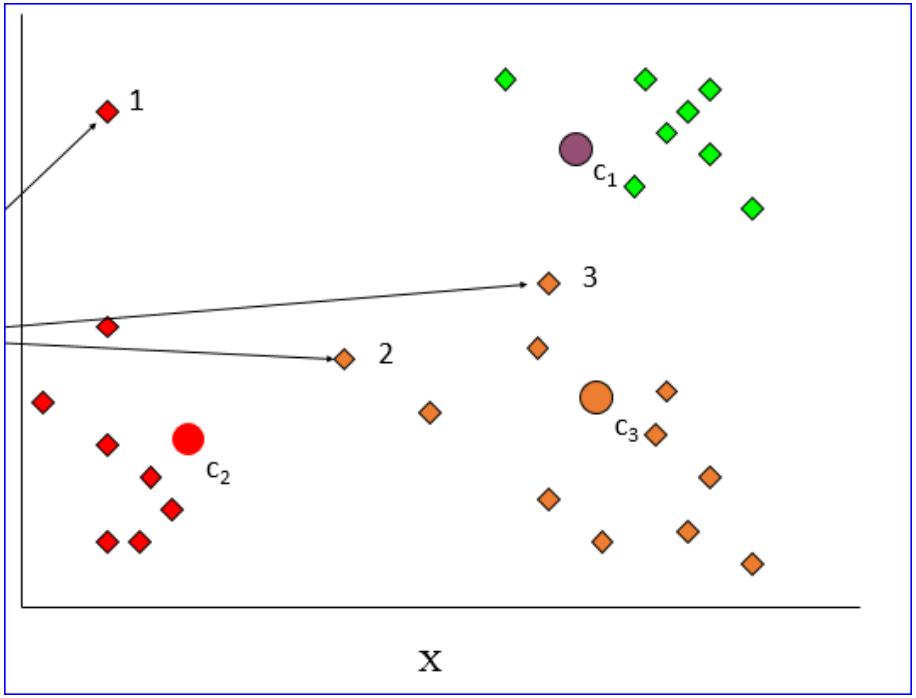
مرحله ۳: به‌روزرسانی مرکز خوشه‌ها



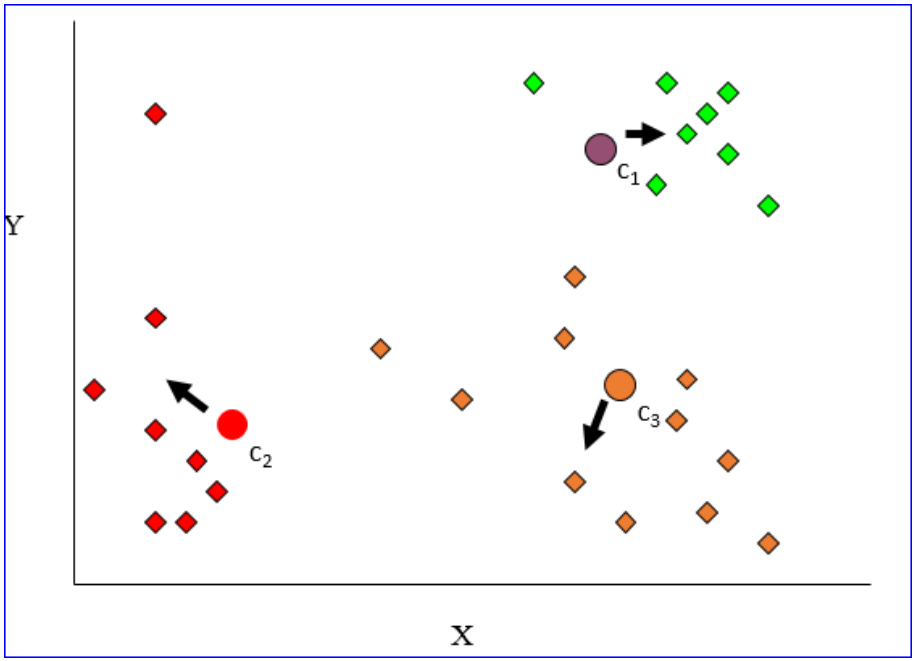
تکرار مرحله ۲: محاسبه فاصله‌های بین داده‌ها



تکرار مرحله ۲: تغییر خوشه ۳ داده

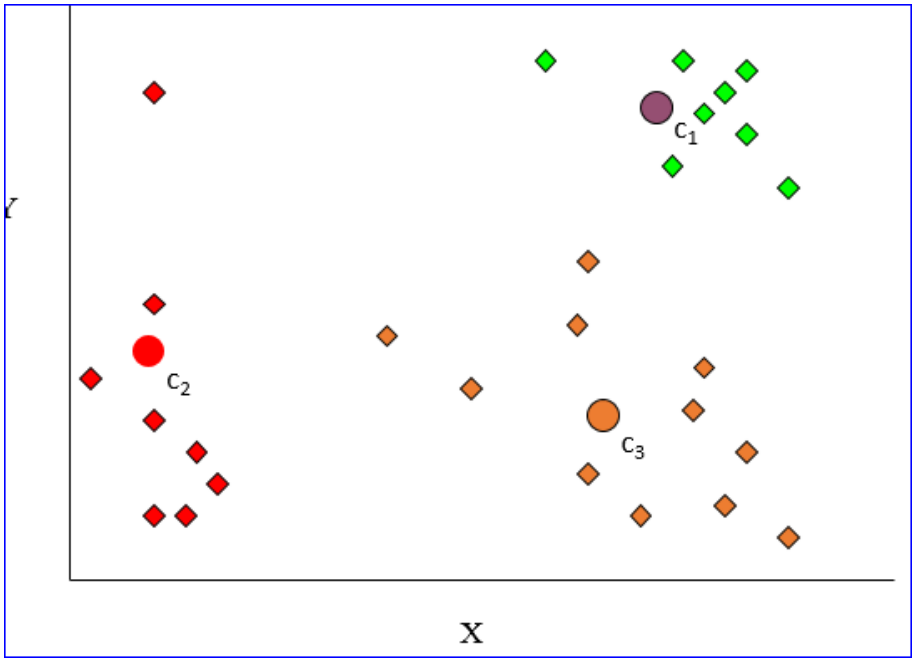


تکرار مرحله ۳: به‌روزرسانی مرکز خوشه‌ها



الگوریتم kمیانگین

همگرا شدن الگوریتم و خوشه‌های نهایی



الگوریتم kمیانگین

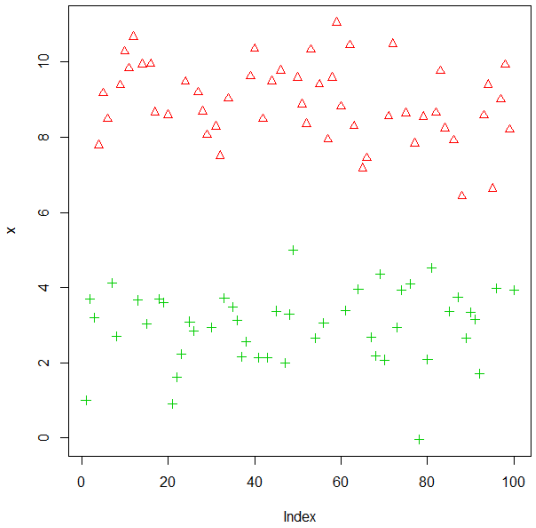
محدودیت‌ها

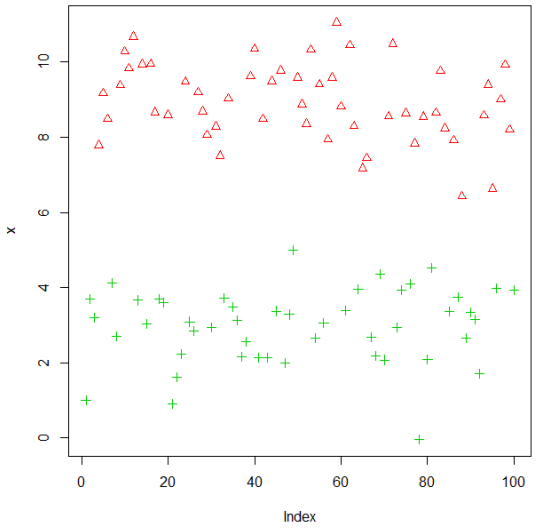
- این روش فقط با داده‌های عددی کار می‌کند،
- نداشتن معیاری برای تعیین تعداد و مرکز اولیه خوشه‌ها،
- اگر در خوشه‌ای هیچ داده‌ای وجود نداشته باشد راهی برای تغییر و بهبود آن وجود ندارد،
- همگرا شدن به مینیمم موضعی،
- تاثیر منفی ویژگی‌های بی‌ربط داده‌ها،
- مناسب نبودن برای داده‌های دم‌سنگین.

الگوریتم kمیانگین

شکل سمت چپ: داده‌های تولید شده از توزیع نرمال آمیخته با دو مولفه ($k = ۲$)

شکل سمت راست: خوشه‌بندی داده‌های تولید شده از توزیع نرمال آمیخته با متر اقلیدسی (خطای خوشه‌بندی: ۰ درصد)





الگوریتم kمیانگین

شکل سمت چپ: داده‌های تولید شده از توزیع لوی آمیخته با دو مولفه ($k = ۲$)

شکل سمت راست: خوشه‌بندی داده‌های تولید شده از توزیع لوی آمیخته با متر اقلیدسی (خطای خوشه‌بندی: ۵۰ درصد)