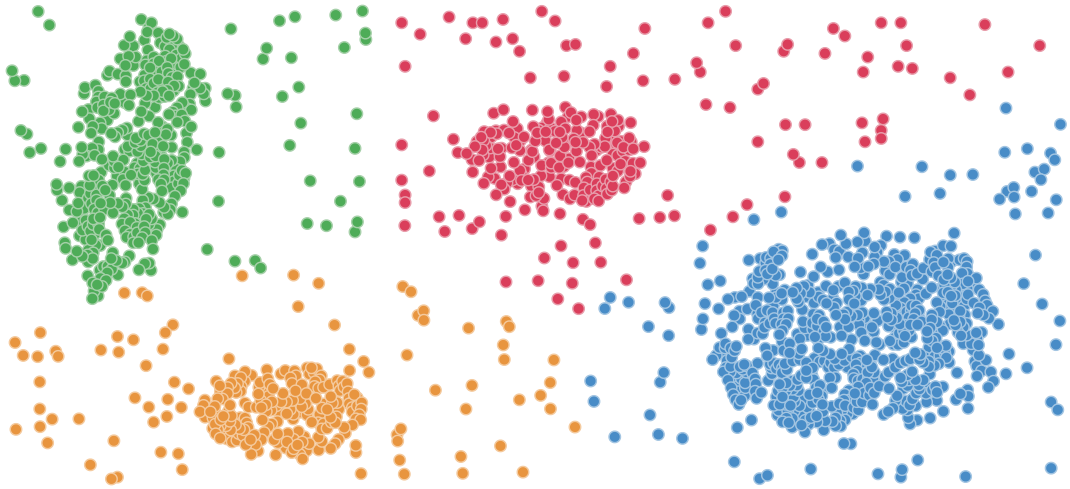
# Кластеризация (Clustering)

Синонимы: Сегментация, Segmentation
Разделы: Бизнес-задачи, Алгоритмы

Объединение объектов или наблюдений в непересекающиеся группы, называемые кластерами, на основе близости значений их признаков. В результате в каждом кластере будут находиться объекты, похожие по своим свойствам друг на друга и отличающиеся от объектов, которые содержатся в других кластерах. При этом чем больше подобие объектов внутри кластера и чем сильнее их отличие от объектов в других кластерах, тем лучше кластеризация.



Формальная постановка задачи кластеризации выглядит следующим образом:

Пусть заданы множества объектов $X = (x_1, x_2, \ldots, x_n)$ и номеров (имен, меток) кластеров $Y = (y_1, y_2, \ldots, y_k)$. Для $X$ определена некоторая функция расстояния между объектами $D(x, x')$, например, метрика $L_2$. Кроме этого, имеется конечная выборка обучающих примеров $X_m = (x_1, x_2, \ldots, x_m)$ из множества $X$, которую требуется разбить на $X_m$ непересекающиеся подмножества (кластеры) так, чтобы каждое из них состояло бы только из элементов, близких по метрике $D$. При этом каждому объекту $x_i$ из множества $X_m$ присваивается номер кластера $y_j$.

Тогда задача будет заключаться в поиске функции $f$, которая любому объекту $x$ из множества $X$ ставит в соответствие номер кластера $y$ из множества $Y$, которое само по себе бывает известно заранее. Однако в большинстве случаев приходится определять оптимальное число кластеров исходя из особенностей решаемой задачи.

Кластеризация позволяет добиться следующих целей:

- улучшает понимание данных за счет выявления структурных групп;
- разбиение набора данных на группы схожих объектов позволяет упростить дальнейшую обработку и принятие решений, применяя к каждому кластеру свой метод анализа;
- позволяет компактно представлять и хранить данные. Для этого вместо хранения всех данных можно оставить по одному типичному наблюдению из каждого кластера;
- поиск новизны — обнаружение нетипичных объектов, которые не попали ни в один кластер.

В Data Mining кластеризация используется для сегментации клиентов и рынков, медицинской диагностики, социальных и демографических исследований, определения кредитоспособности заемщиков и во многих других областях. Решение Loginom Customer Segmentation позволяет автоматизировать анализ поведения клиентов и их сегментацию на основе ключевых метрик.

Специальный обработчик кластеризация производит в Loginom кластеризацию объектов на основе алгоритмов k-means и g-means. А в статье «Алгоритмы кластеризации на службе Data Mining» описан целостный взгляд на последние достижения в области разработки эффективных подходов к кластеризации данных.