Задача и алгоритмы кластеризации

Николай Анохин

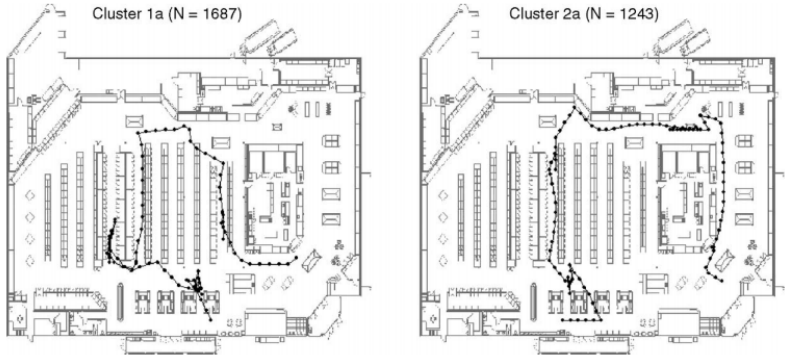
Обучение без учителя

В задачах **без учителя** значение целевой функции для объектов из обучающей выборки неизвестно. Решение таких задач подразумевает исследование “скрытой структуры” данных.

Задача **кластеризации** – задача без учителя, подразумевающая разбиение множества объектов на непересекающиеся подмножества (кластеры).

Мотивация

- Кластеризация позволяет больше узнать о данных (knowledge discovery!)



Типичные траектории покупателей супермаркета[1]

[1] An exploratory look at supermarket shopping paths // J.S. Larson et. al.

Мотивация

- Работать с кластерами удобнее, чем с отдельными объектами

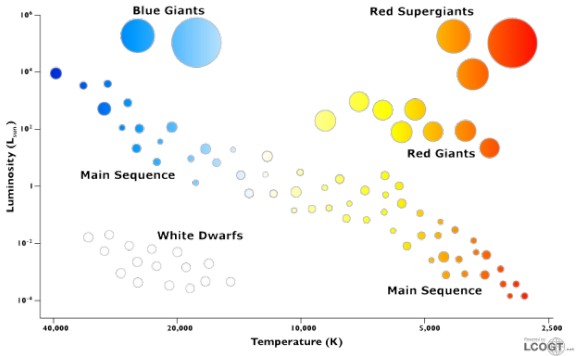


Диаграмма Герцшпрунга — Рассела[1]

[1] https://lcogt.net/spacebook/h-r-diagram/

Мотивация

- Кластеры можно использовать как признаки в других задачах

Задача кластеризации

Дано. Признаковые описания N объектов $\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{X}$, образующие тренировочный набор данных X

Найти. Модель из семейства параметрических функций

$$H = \{h(\mathbf{x}, \theta) : \mathcal{X} \times \Theta \to \mathcal{Y} \mid \mathcal{Y} = \{1, \ldots, K\}\},$$

ставящую в соответствие произвольному $\mathbf{x} \in \mathcal{X}$ один из K кластеров так, чтобы объекты внутри одного кластера были похожи, а объекты из разных кластеров различались