

КЛАСТЕРИЗАЦИЯ

Сергей Николенко

НИУ ВШЭ — Санкт-Петербург

19 мая 2017 г.

Random facts:

- 19 мая 1712 г. столица была перенесена из Москвы в Санкт-Петербург
- 19 мая 1536 г. была обезглавлена Анна Болейн, а 19 мая 1649 г. Англия была провозглашена республикой

КЛАСТЕРИЗАЦИЯ

- *Кластеризация* — типичная задача обучения без учителя: задача классификации объектов одной природы в несколько групп так, чтобы объекты в одной группе обладали одним и тем же свойством.
- Под свойством обычно понимается близость друг к другу относительно выбранной метрики.

- Есть набор тестовых примеров $X = \{x_1, \dots, x_n\}$ и функция расстояния между примерами ρ .
- Требуется разбить X на непересекающиеся подмножества (кластеры) так, чтобы каждое подмножество состояло из похожих объектов, а объекты разных подмножеств существенно различались.

- Есть точки $x_1, x_2, \dots, x_n$ в пространстве. Нужно кластеризовать.
- Считаем каждую точку кластером. Затем ближайшие точки объединяем, далее считаем единым кластером. Затем повторяем.
- Получается дерево.

HierarchyCluster($X = \{x_1, \dots, x_n\}$)

- Инициализируем $C = X, G = X$.
- Пока в C больше одного элемента:
 - Выбираем два элемента C c_1 и c_2 , расстояние между которыми минимально.
 - Добавляем в G вершину $c_1 c_2$, соединяем её с вершинами c_1 и c_2 .
 - $C := C \cup \{c_1 c_2\} \setminus \{c_1, c_2\}$.
- Выдаём G .

- В итоге получается дерево кластеров, из которого потом можно выбрать кластеризацию с требуемой степенью детализации (обрезать на том или ином максимальном расстоянии).
- Всё ли понятно?

- В итоге получается дерево кластеров, из которого потом можно выбрать кластеризацию с требуемой степенью детализации (обрезать на том или ином максимальном расстоянии).
- Всё ли понятно?
- Остаётся вопрос: как подсчитывать расстояние между кластерами?

- *Single-link* алгоритмы считают *минимум* из возможных расстояний между парами объектов, находящихся в кластере.
- *Complete-link* алгоритмы считают *максимум* из этих расстояний
- Какие особенности будут у single-link и complete-link алгоритмов? Чем они будут отличаться?

- Нарисуем полный граф с весами, равными расстоянию между объектами.
- Выберем некий предопределённый порог расстояния r и выбросим все рёбра длиннее r .
- Компоненты связности полученного графа — это наши кластеры.

- Минимальное остовное дерево — дерево, содержащее все вершины (связного) графа и имеющее минимальный суммарный вес своих рёбер.
- Алгоритм Краскала (Kruskal): выбираем на каждом шаге ребро с минимальным весом, если оно соединяет два дерева, добавляем, если нет, пропускаем.
- Алгоритм Борувки (Boruvka).

- Как использовать минимальное остовное дерево для кластеризации?

- Как использовать минимальное остовное дерево для кластеризации?
- Построить минимальное остовное дерево, а потом выкидывать из него рёбра максимального веса.
- Сколько рёбер выбросим, столько кластеров получим.

- Идея: кластер – это зона высокой плотности точек, отделённая от других кластеров зонами низкой плотности.
- Алгоритм: выделяем *core samples*, которые сэмплируются в зонах высокой плотности (т.е. есть по крайней мере n соседей, других точек на расстоянии $\leq \epsilon$).
- Затем последовательно объединяем *core samples*, которые оказываются соседями друг друга.
- Точки, которые не являются ничьими соседями, — это выбросы.

- Идея: строим дерево (CF-tree, от clustering feature), которое содержит краткие описания кластеров и поддерживает апдейты.
- $CF_i = \{N_i, LS_i, SS_i\}$: число точек в кластере CF_i ,
 $LS_i = \sum_{x \in CF_i} x_i$ (linear sum), $SS_i = \sum_{x \in CF_i} x_i^2$ (sum of squares).
- Этого достаточно для того, чтобы подсчитать разумные расстояния между кластерами.
- А также для того, чтобы слить два кластера: CF_i аддитивны.

- CF-дерево состоит из CF_i ; оно похоже на B-дерево, сбалансировано по высоте. Кластеры – листья дерева, над ними “суперкластеры”.
- Добавляем новый кластер, рекурсивно вставляя его в дерево; если от этого число элементов в листе становится слишком большим (параметр), лист разбивается на два.
- А когда дерево построено, можно запустить ещё одну кластеризацию (любым другим методом) на полученных “мини-кластерах”.

СПАСИБО!

Спасибо за внимание!