Алгоритмы машинного обучения без учителя

Здравствуйте, уважаемые слушатели! Тема нашей лекции –Алгоритмы машинного обучения без учителя.

План лекции:

- Сущность кластерного анализа
- Разновидности кластерного анализа

1. Вступительное слово

Прежде чем как углубиться в тему сегодняшней лекции, давайте вспомним что такое обучение без учителя. Как оговаривалось в предыдущих лекциях, это своего рода кластерный анализ, где изначально неизвестны классы. Цель кластерного анализа заключается в поиске существующих структур. Методы кластеризации необходимы для обнаружения структуры в данных, которую нелегко найти при визуальном обследовании или с помощью экспертов.

Существуют множество методов кластерного анализа. Но перед их изучением нужно знать некоторые предостережения общего характера:

- большинство методов кластерного анализа – эвристические, то есть, не имеют достаточного статистического обоснования;
- один и тот же метод кластерного анализа не всегда уместно применять для задач разных дисциплин, потому что каждая дисциплина предъявляет свои требования к отбору данных;
- одну и ту же задачу можно решить с помощью различных методов.

2. Сущность кластерного анализа

Кластерный анализ – анализ неклассифицированных многомерных объектов, в результате которого объекты внутри групп были бы похожи в некотором смысле друг на друга, а объекты из разных групп - непохожи.

При решении задачи кластерного анализа учитываются два фактора:

- во-первых, выбранное множество объектов в принципе допускают желательное разбиение на кластеры. Здесь возникает необходимость выбора свойств или характеристик объектов. Вообще предполагается, что проблема выбора характеристик решена до начала процесса кластеризации. Однако следует предупредить, что этим вносится некоторый произвол, что в отдельных случаях требует дополнительного рассмотрения;
- во-вторых, единицы измерения (масштаб) выбраны правильно. Как правило, данные нормализуют вычитанием среднего и делением на стандартное отклонение, так что дисперсия оказывается равной единице. В случае же, когда исходят из непосредственных (обычных) единиц измерения, возникает проблема интерпретации. Однако наиболее серьезная проблема возникает в связи с тем, что разбиение на кластеры зависит от выбора масштаба. Было бы желательно иметь такой метод кластеризации, который был бы инвариантен к изменению масштабов измерения.

Определение кластерного анализа в различных литературных источниках дается по-разному. Одной из причин этого является то обстоятельство, что кластерный анализ используется и совершенствуется в столь различных областях, как социология, психология, экология, геология, медицина и т.д. В кластерном анализе не существует однозначного количественного критерия, поскольку в различных прикладных задачах различными могут быть и цели анализа. Иногда необходимо выделить группы с высокой плотностью распределения и малой дисперсией, а иногда необходимо обнаружить связанные точные структуры.

3. Разновидности кластерного анализа

На сегодняшний момент общепринятой классификации методов кластерного анализа не существует. Давайте рассмотрим в чем отличаются задачи кластеризации.

- По методу: иерархические или неиерархические;
- По типу: жесткая или мягкая.

В иерархической кластеризации есть два метода выявления кластеров:

- агломеративные - в начале кластеризации каждый объект рассматривается как отдельный кластер, которые последовательно объединяются в большие. Данное объединение происходит до тех пор, пока все объекты будут составлять один кластер;
- дивизимные - в начале кластеризации все объекты принадлежат одному кластеру, которые последовательно разделяются на меньшие. Данное разделение происходит до тех пор, пока все объекты будут рассматривается как отдельный кластер.

Иерархическая кластеризация зависит от выбора расстояния или меры близости между кластерами.

В прошлых лекциях мы узнали, что расстояние между объектами можно вычислить с помощью некоторых функции расстояния. На этой занятии нам нужно вычислить расстояние между кластерами. Опишем некоторые из них:

Метод одиночной связи. Метод начинает свою работу с поиска двух наиболее близких объектов, образующих первичный кластер. Каждые последующий объект присоединяется к тому кластеру, к объекту которого он ближе всего. Расстояние между кластерами вычисляется с определением расстояния близких объектов из разных кластеров.

Метод полной связи. Метод начинается с поиска двух отдаленных объектов, каждый из которых будет считаться отдельным кластером. Последующие новые объекты будут присоединены к кластеру у которого самый отдаленный объект находится близко к новому объекту.

Метод средней связи. Метод вычисляет среднее арифметическое расстояние между всевозможными парами объектов между двух кластеров. Количество таких расстояний равно произведению количества объектов в кластерах. Последующие новые объекты будут присоединены к кластеру, с наименьшим средним расстоянием.

Иерархические методы используют для небольших наборов данных. Преимуществом является их наглядность, то есть, результат можно вывести с построением дендрограмм. Дендрограмма это древовидная схема, которая описывает близость отдельных точек и кластеров.

Неиерархические методы по сравнению с иерархическими можно применять для большого количества данных. Данный метод основан итеративном методе дробления. Деление происходит до выполнения правила остановки.

К популярным методам неиерархической кластеризации можно отнести метод k - средних (k-means). Принцип работы заключается в следующем, выбираем k количество случайных объектов, каждый из которых будет считаться центром кластера. Все оставшиеся объекты будут присоединены в тот класс, где расстояние от центра до объекта является самым минимальным. После каждого разделения новых объектов на классы, центроиды кластеров перевычисляются. В качестве центроида выбирается среднее арифметическое всех объектов. Данное перевычисление перевычисление происходит до тех пор пока все объекты не будут разделены.

Есть несколько вариации данного метода. Среди которых можно отметить MiniBatchKMeans и k -means++. В MiniBatchKMeans также задается число k кластеров, но в отличии от k -means он находит начальные точки по очереди. Данное вычисление позволяет выбрать центроиды таким образом, чтобы они находились далеко друг от друга. При этом основная идея заключается в вычислении среднее арифметическое не по всем объектам

кластера, а только в некоторой случайной выборке (batch). А размер этой выборки фиксированный, который задается нами.

k-means++ улучшенная версия k-means, так как в данном алгоритме центроиды не выбираются случайным образом, кроме первой. Остальные центроиды выбираются применяя принцип чем дальше от ближайшего центроида, тем больше вероятность что данный объект станет центроидом.

По типу отнесения объекта одному или нескольким кластерам мы делим их:

- Жесткая кластеризация: объект принадлежит только к одному из кластеров.
- Мягкая кластеризация: объект принадлежит к нескольким кластерам сразу (с некоторыми весами).

Так как мы работаем с библиотекой scikit-learn. Давайте рассмотрим выше описанные методы кластеризации в данном пакете:
Kmeans;
MiniBatchKMeans;
Hierarchy.