SMART Закупка

1 минута назад

Спектральная кластеризация - как математика переопределяет процесс принятия решений

В современном мире больших данных и Интернета вещей для бизнеса характерно оказаться на вершине горы данных. Обладать им - это одно, но использовать его для принятия решений на основе данных - совсем другая игра с мячом. Чувства кишечника и узаконенная эвристика традиционно использовались для разработки протоколов и принятия решений, но мир искусственного интеллекта и больших разрозненных данных меняет это.

Каждый пытается понять и извлечь выгоду из своих данных. Те, кого нет, останутся позади. Эта проблема (и возможности) не ограничивается определенными отраслями. Например, большинство компаний изучают, как они могут использовать данные для принятия более эффективных маркетинговых решений, большинство ритейлеров используют данные для оптимизации своих цепочек поставок, а большинство производителей используют данные для контроля качества конечных продуктов.

Почти все бизнес-проблемы (с окружающими данными) можно разделить на две категории: контролируемое и неконтролируемое обучение. Возьмите, например, программное обеспечение для распознавания лиц. Один из способов распознавания лиц заключается в обучении программы на основе набора данных изображений и связанных тегов. Теги могут включать «Лицо», «Лицо, мужчина» или что-либо еще. Эти теги позволяют алгоритмам определять и узнавать, как выглядит лицо, и различать мужские и женские лица, или, если необходимо, более тонкие детали. Эту задачу можно переформулировать как проблему обучения без присмотра. Разница в том, что теги в примере контролируемого обучения больше не присутствуют. Скорее, алгоритм должен научиться самостоятельно распознавать лица. Технически говоря, алгоритм не сможет идентифицировать лица как лица, а скорее как наборы объектов / изображений, отличных от других объектов / изображений. Пользователь должен сказать компьютеру, что это те лица, которые он идентифицировал. Google представляет собой интересный пример того, как неконтролируемое обучение использовалось для идентификации кошек в видео на YouTube (см. здесь или для более технического лечения здесь).

Кластеризация (или сегментация) является широко распространенной формой обучения без присмотра в бизнесе. Это включает в себя группирование различных точек данных (клиентов,

продуктов, фильмов и т. Д.) В кластеры. В идеале каждый элемент, содержащийся в кластере, аналогичен любому другому элементу в этом кластере, но максимально отличается от элементов в других кластерах. Цель кластеризации - минимизировать разницу между элементами в кластере и максимизировать разницу между отдельными кластерами.

Почему способность кластеризации так важна?

Кластеризация дает предприятиям возможность достигать лучших результатов для инициатив и понимать клиентов и процессы на гораздо более глубоком уровне, чем тот, которого человек может достичь в одиночку. Если вы маркетолог, вы можете быть заинтересованы в разработке целевых маркетинговых стратегий. Прежде чем сделать это, вы должны знать, кому продавать. Это может быть достигнуто путем группировки клиентов на основе сходных атрибутов существующих клиентов. Это проблема, когда кластеры определяются атрибутами, используемыми для определения клиента: возраст, история платежей, история покупок и т. Д.

Предположим, что вы издательская фирма и хотите решить, как продавать новые книги или определять, как переоценивать или продавать старые книги. Книги могут быть сгруппированы с использованием схемы кластеризации, основанной на атрибутах книг. Они могут включать в себя длину, предмет, повторяющиеся группы слов и т. Д. Кластеризация даже всплывает в страховании, градостроительстве и определении использования земли. Они могут выражаться в определении групп держателей страховых полисов, стоимость претензий которых превышает среднюю, в определении групп домов, основанных на местоположении, типе и стоимости, или в идентификации участков земли вокруг использования.

Важно, чтобы исходная цель кластеризации была достигнута во всех этих примерах: минимизировать различия между элементами в кластере, максимизируя различия между кластерами. Сложность данных и используемые сегодня алгоритмы могут сделать это нетривиальной проблемой. Базовые алгоритмы часто не достигают желаемых результатов, поэтому требуется нечто большее. Ниже мы расскажем вам о некоторых распространенных методах, используемых для кластеризации, и сформулируем силу, которую предсказатель приносит на стол.

Традиционная кластеризация

Иерархическая кластеризация и кластеризация k-средних являются двумя наиболее основными и широко используемыми методами кластеризации.

Иерархическая кластеризация основана на организации точек данных в набор одинаковых кластеров, а затем рекурсивной группировки кластеров до тех пор, пока не останется один кластер. По сути, этот алгоритм назначает иерархию точкам данных. Преимущество этого метода заключается в том, что он позволяет пользователям выбирать желаемое количество кластеров и видеть взаимосвязь между каждым кластером.

Одним из основных недостатков этого метода является время, необходимое для запуска. Поскольку алгоритм должен проходить через каждую точку данных и сравнивать группы точек данных с другими группами точек данных, время выполнения резко увеличивается.

Ссылка на статью: [Спектральная кластеризация - как математика переопределяет процесс]

принятия решений