



Сегментация - алгоритмы-3D
Ван Сеюй 323 групп

Используйте k-means для сегментации 3D-модели сетки

- K-means - это простой метод кластеризации, кластеризация относится к неконтролируемому обучению, но в выборке кластеризации нет заданного y , только функция x , например, предполагает, что звезды во Вселенной могут быть представлены как набор точек в трехмерное пространство (x, y, z). Цель кластеризации - найти потенциальную категорию y для каждого образца x и собрать вместе образцы x одной и той же категории y . Для вышеупомянутых звезд результатом кластеризации является скопление звезд. Точки в скоплении находятся относительно близко друг к другу, а расстояние между звездами в разных скоплениях относительно велико.

- (1) Случайным образом выбрать k центроидов из набора данных в качестве центра начального кластера;
- (2) Рассчитайте расстояние от всех точек в наборе данных до этих k точек и классифицируйте точки в ближайший кластер;
- (3) Отрегулируйте центр кластера, то есть переместите центр кластера к геометрическому центру (т.е. среднему);
- (4) Повторяйте шаги 2 и 3 до тех пор, пока центр кластера не перестанет перемещаться, после чего алгоритм сходится.

Преимущества

- 1) Его легко понять, и эффект кластеризации хороший. Хотя это локальный оптимум, локального оптимума часто бывает достаточно;
- 2) При работе с большими наборами данных алгоритм может обеспечить лучшую масштабируемость;
- 3) Когда кластер приблизительно гауссовский, эффект очень хороший;
- 4) Сложность алгоритма невысока.

Недостатки

- 1) Значение К необходимо установить вручную, а результаты, полученные с разными значениями К, будут разными;
- 2) Чувствительные к исходному центру кластера, разные методы отбора дадут разные результаты;
- 3) Чувствителен к выбросам;
- 4) Выборку можно отнести только к одной категории, что не подходит для задач множественной классификации;
- 5) Не подходит для слишком дискретной классификации, классификации несбалансированных образцов и классификации невыпуклой формы.

Настройка и улучшение алгоритмов

- Ввиду недостатков алгоритма К-средних у нас может быть много методов настройки: таких как предварительная обработка данных (удаление аномальных точек), разумный выбор значений K, многомерное отображение и так далее.
- Суть K-means - это алгоритм разделения данных, основанный на евклидовом расстоянии. Размеры с большим средним значением и дисперсией будут иметь решающее влияние на кластеризацию данных. Следовательно, данные, которые не были нормализованы и объединены, не могут быть напрямую задействованы в расчетах и сравнениях. Распространенными методами предварительной обработки данных являются: нормализация и стандартизация данных.
- Кроме того, выбросы или зашумленные данные будут иметь большее влияние на среднее значение, что приведет к смещению центра, поэтому нам также необходимо обнаруживать аномальные точки в данных.

- Выбор значения K имеет большое влияние на K-средних, что также является самым большим недостатком K-means. Распространенным методом выбора значения K является метод Gap statistic , ISODATA.
- Gap statistic:
 - $\text{Gap}(K) = E(\log D_K) - \log(D_K)$
- Где D_K - функция потерь, Здесь $E(\log D_K)$ относится к **математическому ожиданию** $\log D_K$ Это значение обычно создается с помощью моделирования методом Монте-Карло. Мы случайным образом генерируем столько случайных выборок, сколько исходное количество выборок в соответствии с равномерным распределением в области, где расположена выборка, и выполняем K-means для этой случайной выборки чтобы получить D_K Повторите это много раз, обычно 20 раз, мы можем получить 20 $\log D_K$. Посредством усреднения этих 20 значений получается приблизительное значение $E(\log D_K)$. Наконец, можно рассчитать Gap Statistic. К, соответствующий максимальному значению, полученному статистикой разрыва, является лучшим K.

- ISODATA:
- Полное название ISODATA - итеративный самоорганизующийся метод анализа данных. Это решает тот недостаток, что значение K необходимо заранее искусственно определять. При встрече с многомерными и массивными наборами данных людям часто бывает трудно точно оценить размер K. ISODATA внесла улучшения в эту проблему, и ее идея также очень интуитивно понятна: когда количество образцов, принадлежащих категории, слишком мало, категория удаляется, а когда количество образцов, принадлежащих к категории, слишком велико и степень разброс большой, категория удалена. Разделено на две подкатегории

алгоритм

- Скрытая переменная в K-средних - это категория, к которой принадлежит каждая категория. На итеративном этапе алгоритма K-средних центральная точка помечается заново каждый раз, когда центральная точка подтверждается, что соответствует шагу E. в алгоритме EM. Найдите ожидание при текущих параметрах. В соответствии с меткой снова найдите центральную точку, соответствующую шагу M в алгоритме EM, и найдите соответствующий параметр, когда функция правдоподобия максимизирована (когда функция потерь минимизирована)
-

- Сначала посмотрим на вид функции потерь

- $$J = \sum_{i=1}^C \sum_{j=1}^N r_{ij} \cdot v_{ij}(x_j, \mu_i)$$

- $$v(x_j, \mu_i) = \left| |x_j - \mu_i| \right|^2 \text{ if } x_j \in k, r_{nk} = 1 \text{ else } r_{nk} = 0$$

- Чтобы найти крайнее значение, мы заставляем функцию потерь принимать частную производную и равняться 0:

- $$\frac{dJ}{d\mu_k} = 2 \sum_{i=1}^N r(x_i - \mu_k) = 0$$

- k относится к k -й центральной точке, поэтому мы имеем:

- $$\mu_k = \frac{\sum_{i=1}^N (r_{ik} x_i)}{\sum_{i=1}^N r_{ik}}$$

- Видно, что новая центральная точка является центром масс всех классов. Недостатком EM-алгоритма является то, что он легко попадает в локальный минимум, поэтому K-means иногда находит локальное оптимальное решение.