

 ЕМ Кластеризация

Описание

В основе ЕМ кластеризации лежит масштабируемый алгоритм ЕМ, который опирается на предположение, что исследуемое множество данных может быть смоделировано с помощью линейной комбинации многомерных нормальных распределений. Целью при этом является оценка параметров распределения, которые максимизируют логарифмическую функцию правдоподобия, используемую в качестве меры качества модели. Иными словами, предполагается, что данные в каждом кластере подчиняются определенному закону распределения, а именно, нормальному распределению.

Таким образом, любое наблюдение (объект) принадлежит ко всем кластерам, но с разной вероятностью. Объект должен быть отнесен к тому кластеру, для которого данная вероятность выше.

Порты

Вход

-  Входной источник данных (таблица данных).

Требования к принимаемым данным

Поле будет запрещено к использованию, если:

- оно является дискретным и содержит всего одно уникальное значение;
- оно непрерывное и с нулевой дисперсией;
- оно содержит пропущенные значения.

Выход

-  Разбиение на кластеры (таблица данных).

Таблица, состоящая из полей:

- **Номер кластера** — каждому объекту присвоен номер того кластера, в который он входит.
- **Вероятность принадлежности** — для каждого объекта проставляется вероятность принадлежности к данному кластеру.
- Поля исходного набора данных (значения не изменяются).

-  Центры кластеров (таблица данных).

Центр кластера — среднее значение переменных объектов, входящих в кластер. Результат — таблица, количество записей которой соответствует числу кластеров, т.е. данные сгруппированы по кластерам. Состоит из полей:

- **Номер кластера** — перечислены номера сформированных кластеров.
- Поля исходного набора данных, в ячейках которых рассчитано среднее значение параметров.

Мастер настройки

Мастер настройки включает в себя следующие группы параметров:

- Настройка входных столбцов.
- Настройка нормализации.
- ЕМ Кластеризация.

Настройка входных столбцов

- Выбор полей для кластеризации:
 - Для полей, участвующих в кластеризации, выставить назначение *Используемое*.
 - Для прочих полей оставить *Не задано*.

ЕМ Кластеризация

- Автоопределение числа кластеров. При установлении флага становится доступен для настройки блок параметров *Автоматическое определение числа кластеров*.
 - Минимальное число кластеров. По умолчанию — 1.
 - Максимальное число кластеров. По умолчанию — 10.
 - Порог разделения кластеров (в интервале от 0,1 до 5). Чем больше порог разделения, тем больше кластеров будет сгенерировано при кластеризации. По умолчанию — 1.
- Заданное число кластеров. Блок параметров, доступный для настройки, при неактивном флаге *Автоопределение числа кластеров*.
 - Число кластеров. По умолчанию — 3.
- Параметры ЕМ-кластеризации.
 - Переменные независимы. Учет зависимости между переменными. Если флаг активен, зависимости нет.
 - Медианная модификация. Включение флага означает, что на втором М-шаге алгоритма моментные оценки максимального правдоподобия заменяются более устойчивыми оценками медианного типа. Это может повысить устойчивость алгоритма по отношению к начальным данным.
- **Random seed** — начальное число (целое, положительное), которое используется для инициализации генератора псевдослучайных чисел. Последовательность чисел генератора полностью определяется начальным числом. Если генератор повторно инициализируется с тем

же начальным числом, он выдаст ту же последовательность чисел.

Параметр влияет на воспроизводимость результата обучения. Можно повторить результат обучения узла, если подать те же данные и выставить тот же random seed.

Для параметра доступны следующие команды:

- Всегда случайно — начальное число всегда будет случайным.
- Генерировать — сгенерируется новое начальное число.
- Копировать — в буфер обмена будет скопировано указанное значение.