**Технический справочник по алгоритму линейной регрессии (Майкрософт)**

* 08.05.2018
* Время чтения: 10 мин
* Соавторы

**Область применения:** SQL Server Analysis Services Azure Analysis Services

Алгоритм линейной регрессии Microsoft — это частный случай алгоритма дерева принятия решений (Майкрософт), оптимизированный для моделирования пар непрерывных атрибутов.В данном разделе описывается реализация алгоритма и настройка его поведения. Приводятся также ссылки на дополнительную информацию о запросах к моделям.

**Реализация алгоритма линейной регрессии (Майкрософт)**

Алгоритм дерева принятия решений Майкрософт можно использовать во многих задачах, таких как линейная регрессия, классификация или анализ взаимосвязей. Для реализации этого алгоритма с целью изучения линейной регрессии параметры алгоритма настраиваются так, чтобы контролировать рост дерева и держать все данные модели в едином узле. Иными словами, несмотря на то, что алгоритм линейной регрессии основан на дереве принятия решений, это дерево содержит только один корень и не имеет ветвей: все данные содержатся в корневом узле.

Чтобы добиться такой конфигурации, параметру *MINIMUM\_LEAF\_CASES* алгоритма присваивается значение, большее или равное общему количеству вариантов, используемых алгоритмом для обучения модели интеллектуального анализа данных. Если параметр задан таким способом, алгоритм никогда не будет создавать разбиений, то есть проведет линейную регрессию.

Формула, представляющая линию регрессии, имеет общий вид y = ax + b и известна как уравнение регрессии. Переменная Y представляет выходную переменную, X — входную переменную, a и b — изменяемые коэффициенты. С помощью запросов к завершенной модели интеллектуального анализа данных можно получать коэффициенты, точки пересечения и другую информацию о формуле регрессии. Дополнительные сведения см. в разделе [Примеры запросов модели линейной регрессии](https://docs.microsoft.com/ru-ru/sql/analysis-services/data-mining/linear-regression-model-query-examples?view=sql-server-2017).

**Методы количественной оценки и выбор компонентов**

Выбор компонентов автоматически применяется всеми алгоритмами интеллектуального анализа данных служб Службы Analysis Services для улучшения качества анализа и снижения вычислительной нагрузки. Для выбора компонентов в моделях линейной регрессии применяется оценка интересности, поскольку модель поддерживает только непрерывные столбцы. В следующей таблице в информационных целях описана разница в выборе компонентов между алгоритмом линейной регрессии и алгоритмом дерева принятия решений.

| **Алгоритм** | **Метод анализа** | **Комментарии** |
| --- | --- | --- |
| Линейная регрессия | Оценка интересности | По умолчанию.Другие методы выбора компонентов, доступные алгоритму дерева принятия решений, применяются только к дискретным переменным, и, следовательно, неприменимы к моделям линейной регрессии. |
| Деревья решений | Оценка интересностиЭнтропия ШеннонаАлгоритм Байеса с априорной оценкой K2Эквивалент Дирихле метода Байеса с однородной априорной оценкой (выбор по умолчанию) | Если какие-либо столбцы содержат недвоичные непрерывные значения, то оценка интересности используется для всех столбцов в целях обеспечения согласованности. В противном случае используется предусмотренный по умолчанию или указанный метод. |

Выбором компонентов в модели дерева решений управляют следующие параметры алгоритма: MAXIMUM\_INPUT\_ATTRIBUTES и MAXIMUM\_OUTPUT.

**Настройка алгоритма линейной регрессии**

Алгоритм линейной регрессии Microsoft поддерживает несколько параметров, влияющих на поведение, производительность и точность итоговой модели интеллектуального анализа данных. Можно также изменять способ обработки данных, устанавливая на столбцах модели интеллектуального анализа данных или структуры интеллектуального анализа данных флаги модели.

**Задание параметров алгоритма**

В следующей таблице описываются параметры, которые можно использовать с алгоритмом линейной регрессии Microsoft :

| **Параметр** | **Описание** |
| --- | --- |
| *MAXIMUM\_INPUT\_ATTRIBUTES* | Определяет количество входных атрибутов, которые алгоритм может обработать перед вызовом выбора компонентов.Установите значение 0, чтобы отключить выбор компонентов.Значение по умолчанию — 255. |
| *MAXIMUM\_OUTPUT\_ATTRIBUTES* | Определяет количество выходных атрибутов, которые алгоритм может обработать перед вызовом выбора компонентов.Установите значение 0, чтобы отключить выбор компонентов.Значение по умолчанию — 255. |
| *FORCE\_REGRESSOR* | Приводит алгоритм к использованию указанных столбцов в качестве регрессоров, не обращая внимания на важность столбцов, вычисленную алгоритмом. |

**Флаги моделирования**

Алгоритм линейной регрессии Microsoft поддерживает приведенные ниже флаги модели.Чтобы задать порядок обработки в ходе анализа значений в каждом столбце, во время создания структуры или модели интеллектуального анализа данных определяются флаги модели. Дополнительные сведения см. в разделе [Флаги моделирования (интеллектуальный анализ данных)](https://docs.microsoft.com/ru-ru/sql/analysis-services/data-mining/modeling-flags-data-mining?view=sql-server-2017).

| **Флаг моделирования** | **Description** |
| --- | --- |
| NOT NULL | Указывает, что столбец не может принимать значение NULL. Если во время обучения модели службы Analysis Services обнаружат значение NULL, возникнет ошибка.Применяется к столбцам структуры интеллектуального анализа данных. |
| REGRESSOR | Указывает, что столбец содержит непрерывные численные значения, которые при анализе нужно обрабатывать как потенциальные независимые переменные.Применяется к столбцам модели интеллектуального анализа данных.Примечание. Если столбец указан как регрессор, это не значит, что он будет использован в качестве регрессора в окончательной модели. |

**Регрессоры в моделях линейной регрессии**

Модели линейной регрессии основаны на алгоритме дерева принятия решений Microsoft .Даже если не используется алгоритм линейной регрессии Microsoft , любая модель дерева принятия решений может содержать дерево или узлы, представляющие регрессию применительно к какому-то непрерывному атрибуту.

Необходимость указывать, что непрерывный столбец представляет регрессор, отсутствует.Алгоритм дерева принятия решений Microsoft секционирует набор данных на области со значимыми шаблонами, даже если для столбца не задан флаг REGRESSOR. Различие между указанными случаями состоит в том, что если пользователем задан флаг модели, то в алгоритме предпринимается попытка найти регрессионные уравнения в форме a\*C1 + b\*C2 + ... для подгонки шаблонов к узлам дерева. Вычисляется сумма остатков, и, если отклонение слишком велико, принудительно выполняется разбиение дерева.

Например, если осуществляется прогноз поведения клиента в процессе покупки с использованием дохода в качестве атрибута и на соответствующем столбце [Income] устанавливается флаг модели REGRESSOR, то в алгоритме вначале предпринимается попытка выполнить подгонку значений с применением стандартной формулы регрессии. Если отклонение слишком велико, то происходит отказ от применения формулы регрессии и разбиение дерева осуществляется по какому-то другому атрибуту. Затем в алгоритме дерева решений осуществляется попытка подгонки регрессора к доходу в каждой из ветвей, полученных после разбиения.

Можно применить параметр FORCED\_REGRESSOR для обеспечения того, чтобы в алгоритме использовался конкретный регрессор. Этот параметр может применяться с алгоритмом дерева принятия решений (Майкрософт) и алгоритмом линейной регрессии (Майкрософт).

**Требования**

Модель линейной регрессии должна содержать ключевой столбец, входные столбцы и хотя бы один прогнозируемый столбец.

**Входные и прогнозируемые столбцы**

Алгоритм линейной регрессии Microsoft поддерживает определенные входные столбцы данных и прогнозируемые столбцы, которые перечислены ниже в таблице. Дополнительные сведения о значении типов содержимого в применении к модели интеллектуального анализа данных см. в разделе [Типы содержимого (интеллектуальный анализ данных)](https://docs.microsoft.com/ru-ru/sql/analysis-services/data-mining/content-types-data-mining?view=sql-server-2017).

| **Столбец** | **Типы содержимого** |
| --- | --- |
| Входной атрибут | Continuous, Cyclical, Key, Table и Ordered |
| Прогнозируемый атрибут | Continuous, Cyclical и Ordered |

**Примечание**

Типы содержимого**Cyclical** и **Ordered** поддерживаются, но алгоритм обрабатывает их как дискретные величины и не производит их особой обработки.

**См. также**