

Линейная регрессия

Описание

Линейная регрессия представляет собой модель зависимости между входными и выходными переменными с линейной функцией связи.

Линейная регрессия является одним из наиболее часто используемых алгоритмов в машинном обучении. Этот алгоритм зачастую дает хороший результат даже на небольших наборах данных.

Широкое применение линейной регрессии обусловлено тем, что большое количество реальных процессов в науке, экономике и бизнесе можно описать линейными моделями. Так, с помощью линейной регрессии можно оценивать объем ожидаемых продаж в зависимости от установленной цены.

Компонент может использоваться для решения различных задач Data Mining, например, таких, как прогнозирование и численное предсказание.

Для получения результирующих наборов требуется предварительное обучение узла.

Порты

Вход

-  — Входной источник данных (таблица данных) — обязательный порт.
-  — Управляющие переменные (переменные) — необязательный порт, переменными можно задать значения параметров мастера настройки.

Выходы

-  — Выход регрессии — таблица, состоящая из полей: регрессия поля выходных данных, поля исходного набора данных.
-  — Коэффициенты регрессионной модели — таблица данных.
-  — Сводка — переменные.

Мастер настройки узла

Мастер настройки включает в себя следующие группы параметров:

- Настройка входных столбцов.
- Настройки нормализации.
- Разбиение на множества.
- Настройка линейной регрессии.
- Детальные настройки.

Настройка входных столбцов

На первом этапе необходимо задать назначение полей входного набора данных:

-  **Входное** — поле содержит значения одного из входных параметров.
-  **Выходное** — поле содержит значение выходного параметра.
-  **Не задано** — поле не участвует в обработке. Устанавливается по умолчанию для прочих полей.

Важно:

- Входные данные никогда не должны содержать пропусков, выходные данные не должны содержать пропусков во время обучения.
- Назначение *Выходное* можно задать только для поля с *Вещественным типом данных и Непрерывным видом данных*. Вид данных *Входных* полей не регламентируется.

Настройка нормализации

На этом этапе входные данные приводятся к определенным диапазонам в соответствии с выбранным алгоритмом нормализации. Использование нормализации увеличивает качество и скорость обучения *Линейной регрессии*.

Разбиение на множества

Страница *Разбиение на множества* мастера настройки узла позволяет разделить множество на обучающее и тестовое:

- Обучающее — структурированный набор данных, применяемый для обучения аналитических моделей. Каждая запись обучающего множества представляет собой обучающий пример, содержащий заданное входное воздействие и соответствующий ему правильный выходной (целевой) результат.
- Тестовое — подмножество обучающей выборки, содержащее тестовые примеры, т.е. примеры, использующиеся не для обучения модели, а для проверки его результатов.

Доступные параметры:

- Размер обучающего и тестового множества в процентах или строках. Может быть задан с помощью переменных.
- Метод разбиения на обучающее и тестовое множество. Существует три метода разбиения:
 - Случайный — случайно разбивает множество записей на обучающее и тестовое множество.
 - Последовательный — группы строк множеств (обучающее, неиспользуемое, тестовое) выбираются последовательно, т.е. сначала выбираются те записи, которые входят в первое множество, затем — во второе и т.д. Порядок множеств можно менять (кнопки *Сдвинуть вверх*, *Сдвинуть вниз*).
 - По столбцу — разбиение на обучающее и тестовое множества задаётся при помощи параметра. Параметром выступает столбец с логическим типом данных, где значение «ИСТИНА» указывает на то, что запись относится к тестовому набору, а значение «ЛОЖЬ» — на то, что запись принадлежит обучающему набору (т.е. можно разбить множество на обучающее и тестовое в узле Разбиение на множества и подать данные из порта *Общий выходной набор* на вход узла *Линейная регрессия*, выбрав в качестве параметра разбиения по столбцу колонку «Тестовое множество»). При выборе данного метода таблица выбора соотношения обучающего и тестового множеств становится неактивной.
- Метод валидации, который может принимать следующие значения:
 - Без валидации.
 - K-fold кросс-валидация — позволяет выбрать *Метод сэмплинга* и количество *Колод кросс-валидации*.
 - Монте-Карло — позволяет выбрать *Количество итераций ресемплинга* и задать размер обучающего и валидационного множества.

Важно: Кросс-валидация используется при обучении модели для автоматического подбора коэффициентов регуляризации (автоматического выбора модели) в режиме автоматической настройки - максимальная точность, повышенная точность, средняя скорость. Кросс-валидация выполняется на обучающем множестве (на обучающие и тестовые колоды делится обучающее множество).

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 порядок случайного разбиения на тестовое и обучающее множество и 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.

Настройка линейной регрессии

Набор параметров для настройки линейной регрессии можно сгруппировать в следующие блоки:

Настройка метода

- Автоматическая настройка:
 - Значение логического типа. По умолчанию включено.
 - Влияет на использование следующих блоков параметров: если включена, то можно настраивать блок *Приоритет автоматической настройки*, если выключена, то можно выбирать алгоритм отбора факторов и защиту от переобучения и настраивать приоритеты.
- Приоритет автоматической настройки:
 - Влияет на выбор конкретного метода и его настроек по шкале Точность — Скорость.
 - Целочисленный тип. Может принимать следующие значения:
 - Максимальная точность.
 - Повышенная точность.
 - Средняя скорость.
 - Повышенная скорость.
 - Максимальная скорость.

Настройка параметров

Используется, если не выбран флаг *Автоматическая настройка*.

- **Отбор факторов и защита от переобучения** — значение перечисления:
 - **Принудительное включение (Enter)** — включение в регрессионную модель всех заданных признаков независимо от того, оказывают ли они значимое влияние или нет.
 - **Пошаговое включение (Forward)** — метод, который базируется на принципе: начать с отсутствия признаков и постепенно найти самые «лучшие», которые будут добавлены в подмножество.
 - **Пошаговое исключение (Backward)** — метод основан на следующем: начать со всех доступных признаков и последовательными итерациями исключить самые «худшие».
 - **Пошаговое включение/исключение (Stepwise)** — модификация метода *Forward* с тем отличием, что на каждом шаге после включения новой переменной в модель осуществляется проверка на значимость остальных переменных, которые уже были введены в нее ранее.
 - **Ridge** (Гребневая регрессия) — процедура регуляризации L2, предназначена для защиты от переобучения (снижения степени переобучения обучаемой модели), что обеспечивает более качественное прогнозирование. Подразумевает введение «штрафов» для уменьшения значений коэффициентов регрессии. Величина «штрафов» вычисляется как сумма квадратов коэффициентов переменных умноженная на коэффициент регуляризации.
 - **LASSO** (регрессия Лассо) — также как и *Ridge*, применяется для регуляризации (защиты от переобучения) обучаемой модели. Подразумевает введение «штрафов» (вычисляются как сумма модулей коэффициентов переменных умноженные на коэффициент регуляризации) для уменьшения значений коэффициентов регрессии. Регуляризация Lasso (L1) позволяет снизить размерность и упростить регрессионную модель, за счёт зануления коэффициентов некоторых признаков.

- **Elastic-Net** (регрессия «Эластичная сеть») — модель регрессии с двумя регуляризаторами $L1$, $L2$. Частными случаями являются модели *LASSO* $L1 = 0$ и *Ridge* регрессии $L2 = 0$. Данный тип регуляризации учитывает эффективность обоих методов регуляризации и применяется в тех случаях, когда важно сохранить корреляционную связь переменных и не допустить зануления коэффициентов регрессионной модели (как в случае с *LASSO*).
- Приоритет точность/скорость.
 - Целочисленный тип. Может принимать следующие значения:
 - Максимальная точность.
 - Повышенная точность.
 - Средняя скорость.
 - Повышенная скорость.
 - Максимальная скорость.
- Приоритет точные/недостовверные данные.
 - Целочисленный тип. Может принимать следующие значения:
 - Точные данные.
 - Повышенная точность.
 - Средняя точность.
 - Пониженная точность.
 - Недостовверные данные.
- Приоритет меньше/больше факторов.
 - Целочисленный тип. Может принимать следующие значения:
 - Минимум факторов.
 - Меньше факторов.
 - Среднее число факторов.
 - Больше факторов.
 - Максимум факторов.

Перечисленные опции доступны для различных методов:

Метод	Приоритет точность/скорость	Приоритет точные/ недостовверные данные	Приоритет меньше/ больше факторов
Enter	•		
Forward	•		•
Backward	•		•
Stepwise	•		•
Ridge	•	•	
LASSO	•		•
Elastic-Net	•	•	•

- **Денормализировать коэффициенты модели** — денормализация необходима для интерпретации результатов. Т.к. модель может работать только с нормализованными данными, то для ее работы необходимо сначала нормализовать данные, которые поступили в модель, а затем провести денормализацию для того, чтобы данные приняли вид, который был до нормализации. Является значением логического типа, по умолчанию включено.
- **Использовать детальные настройки** — позволяет более развернуто настроить линейную регрессию (появляется дополнительная страница мастера — блок детальных настроек). Является значением логического типа, по умолчанию выключено. Доступен для редактирования, если снят флаг *Автоматическая настройка*.
- **Показывать коэффициенты опорных категорий** — возвращает нулевые коэффициенты регрессии для опорных категорий. Является значением логического типа, по умолчанию включено.

Примечание: все доступные параметры настройки линейной регрессии можно задавать с помощью переменных.

Детальные настройки

Используются, если включен блок настроек параметров и в нем установлен флаг *Использовать детальные настройки*.

Детальные настройки объединяются в следующие блоки параметров:

Настройки метода

Доступные параметры:

- **Точность решения** — критерий остановки итераций. Настройка, которая позволяет определить точность нахождения минимума функции ошибки. Значение вещественного типа от 0 до 1. Представляет собой редактор с шагом изменения значения 0,000001.
- **Включить в модель константу** — добавляет в модель константу (свободный член). Если удалить этот переключатель, линия регрессии будет проходить через начало координат. Параметр доступен для редактирования, если не установлен флаг *Денормализировать коэффициенты модели*.

Настройки расчета статистики

- Уровень доверия, %

Настройки регуляризации

Доступные параметры:

- **Установка коэффициента L1-регуляризации** — настройка данного параметра возможна только для алгоритмов *LASSO*, *Elastic-Net*.

- **Установка коэффициента L2-регуляризации** — настройка данного параметра возможна только для алгоритмов *Ridge*, *Elastic-Net*.

Для каждого из параметров можно задать либо автоматическую установку значения, либо ввести необходимое значение вручную.

Настройки отбора факторов

Доступные параметры:

- **Критерий отбора факторов** — позволяет выбрать один из следующих информационных критериев:
 - Критерий Фишера (F-тест).
 - Коэффициент детерминации.
 - Скорректированный коэффициент детерминации.
 - Информационный критерий Акаике.
 - Информационный критерий Акаике скорректированный.
 - Информационный критерий Байеса.
 - Информационный критерий Ханнана-Куинна.
- Порог значимости при добавлении фактора.
- Порог значимости при исключении фактора.

Примечание: все доступные параметры детальных настроек можно задавать с помощью переменных.

Статьи в разделе:

- Сводка
- Коэффициенты регрессии