Конспект лекции

А.Н. Шевляков

Машинное обучение

Линейная регрессия и ее обобщения

# Постановка задачи предсказания

Предсказание (prediction).есть множество объектов *М* с известными значениями признака *Y.* Найти значение признака *Y* для нового объекта *А∉М. Y* называется целевым признаком.

Предсказание значения количественного признака Y называется задачей регрессии.

Предсказание значения номинального (категориального) признака Y называется задачей классификации.

Например, предсказание IQ – это задача регрессии (Таблица 1), а предсказание пола – задача классификации (Таблица 2).

Таблица 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Рост** | **Вес** | **Пол** | **IQ** |
| Вася | 170 | 80 | 1 | 100 |
| Даша | 165 | 60 | 0 | 80 |
| Маша | 160 | 50 | 0 | 110 |
| Петя | 200 | 70 | 1 | 50 |

Таблица 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Рост** | **Вес** | **Пол** | **IQ** |
| А | 180 | 75 | 1 | ? |

Слово «регрессия» произошло из-за того, что исторически первая задача предсказания количественного признака обнаружила эффект «возврата (регресса) к среднему».

Задачи предсказания делятся на два типа: задача классификации и задача регрессии. Эти два типа задач требуют принципиально разных подходов к решению.

# План решения задачи регрессии

При решении задачи регрессии выборка разбивается на тренировочную и проверочную (тестовую) части.

Смысл модели линейной регрессии: необходимо минимизировать величину отклонений объектов от прямой (плоскости, гиперплоскости).

**Общий план решения задачи регрессии**

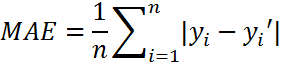
Множество объектов разбить на 2 множества: тренировочную выборку Train и тестовую (проверочную) выборку Test.

Модель предсказания строится по объектам Train, а качество модели проверяется по объектам Test.

Допустим по тренировочной выборке мы научились предсказывать целевой признак Y.

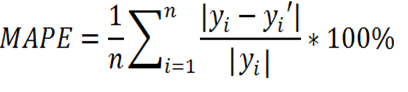
Как оценить качество предсказаний по тестовой выборке?

МАЕ (средняя абсолютная ошибка):



(*n* – объем тестовой выборки, *yi* – истинные *yi*’ – предсказанные значения).

МАPЕ (средняя абсолютная ошибка в процентах):



(*n* – объем тестовой выборки, *yi* – истинные *yi*’ – предсказанные значения).

Модель регрессии не обязана давать точный ответ на объектах тренировочной (!) выборки, которые использовались при построении модели.

Модель регрессии называется линейной, если значение предсказываемого признака Y вычисляется как сумма известных признаков *X1,X2,…Xm*, взятых с некоторыми коэффициентами.



Задача заключается в нахождении оптимальных весов (коэффициентов) *wi*. Как искать веса *wi*?

Принцип: для объектов тренировочной нужно минимизировать отклонение предсказываемых значений от истинных значений признака Y.

Важно: линейная регрессия неустойчива к выбросам – выбросы нужно заранее удалить.

# Построение модели линейной регрессии

Мы научимся строить линейную модель для решения задачи регрессии.

Для этого нужно знать, что такое частные производные.

Есть данные тренировочной выборки:

Так как нецелевой признак один Х, то модель предсказания признака Y будем искать в виде *y’=w1x+w0.*

Нужно найти оптимальные *w1,w0.*

Отклонение истинного от предсказанного значения равно *|y’-y|=|w1x+w0-y|,* эту величину нужно минимизировать для всех объектов тренировочной выборки. Иными словами, чем меньше выражение

L(w1,w0)=|*w1x1+w0-y1*|+|*w1x2+w0-y2*|+|*w1x3+w0-y3*|+|*w1x4+w0-y4*|,

тем лучше.

Поиск точки минимума этой функции осложняется тем, что модуль-функция – недифференцируемая. Поэтому на практике минимизируют несколько иную функцию: сумму квадратов отклонений.

Когда нецелевых признаков больше одного, все происходит аналогично, только параметров *wi* будет больше (и полученная зависимость *y’*=… будет уже определять не прямую, а гиперплоскость).

Таким образом, основная трудоемкость при построении линейной регрессии заключается в решении системы линейных уравнений на последнем шаге.

# Проблемы модели линейной регрессии

У линейной регрессии есть проблемы, связанные с наличием зависимости между признаками. Данные проблемы решаются с помощью процедуры регуляризации или лассо.

## Основные проблемы при построении линейной регрессии

Получаемая при построении линейной регрессии система линейных уравнений может:

* не иметь решений
* иметь более одного решения.

Эти проблемы возникают, когда между нецелевыми признаками существует линейная зависимость или высокая корреляция. Такую проблему еще называют «проблемой мультиколлинеарности».

Что нужно делать, чтобы найти хорошую модель регрессии?

- Отбор признаков. Нужно удалять нецелевые признаки, которые линейно зависят от других или имеют высокую корреляцию с другими признаками.

Коэффициент регрессии можно минимизировать («регуляризация» и «лассо»).

## Решение проблем моделей линейной регрессии

### Регуляризация

Нужно минимизировать

*R=L(w1,w2,…,wm,w0)+C(w02 +w12+w22+…+wm2),*

где *С –* заданная константа.

Рассмотрим регуляризацию при *С=1*, то есть будем минимизировать выражение

*R=L(w1,w2,…,wm,w0)+w12+w22+…+wm2+w02*

Уравнение регрессии ищется в виде

*y’=w1x1+w2x2+w0*

Составляем выражение для R и считаем его частные производные.

Какое значение выбрать для константы *С*?

* нужно взять как можно больше различных значений для *С*.
* для каждого значения построить модель регрессии и проверить ее качество на тестовой выборке.
* окончательно выбрать такое значение *С*, которое принадлежит модели с наилучшим качеством.

### Лассо

Нужно (как и в регуляризации) стремится сделать коэффициент регрессии небольшими, теперь мы будем минимизировать другое выражение:

*R=L(w1,w2,…,wm,w0)+C(|w0|+|w1|+|w2|+…+|wn|)*

где *С* – некоторая константа (это и есть лассо).

Оптимальные *wi* здесь уже нельзя найти простым способом, тут нужно использовать методы теории оптимизации.

Какое значение выбрать для константы *С*?

Совет такой же, как и для регуляризации:

* нужно взять как можно больше различных значений для *С*.
* для каждого значения построить модель регрессии и проверить ее качество на тестовой выборке.
* окончательно выбрать такое значение *С*, которое принадлежит модели с наилучшим качеством.

# Полиномиальная регрессия

Модель линейной регрессии легко допускает обобщения. Одним из таких обобщений является полиномиальная регрессия.

Можно ли искать зависимость не только с помощью линейных функций?

Хочется, чтобы:

- можно было бы строить нелинейные модели, а, например, полиномиальные (полиномиальная регрессия).

- алгоритм их построения не сильно бы отличался от алгоритма линейной регрессии.

Добавляй новые столбцы. В нашем примере можно добавить новый признак Х2 (являющийся квадратом признака Х).

И запустить стандартный алгоритм построения линейной регрессии.

То есть будем искать зависимость *y’=w1x1+w2x2+w0*,но фактически это уже не линейная,а квадратичная зависимость.

Таблица 3

|  |  |  |  |
| --- | --- | --- | --- |
| **Объект** | **X** | **Х2** | **Y** |
| А | -1 | 1 | 1 |
| B | 0 | 0 | 0 |
| C | 1 | 1 | 1 |
| D | 2 | 4 | 4 |

В общем случае, если хочется найти зависимость целевого признака в виде полинома *k*-ой степени, то нужно в таблицу добавить новые столбцы, содержащие все возможные произведения нецелевых переменных степени не больше *k*.

Таблица 4

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Объект** | **X** | **Z** | **XX** | **ZZ** | **XZ** | **Y** |
| А | 1 | 3 | 1 | 9 | 3 | 100 |
| B | 2 | 4 | 4 | 16 | 8 | 200 |

В таблице показано расширение таблицы при двух нецелевых признаках для квадратичной регрессии.