Задача предсказания и линейная регрессия

Лекция 4

Постановка задачи предсказания

Формальная постановка

Предсказание (prediction). Есть множество объектов M с известными значениями признака Y. Найти значение признака Y для нового объекта $A \notin M$. Y называется целевым признаком.

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

	Рост	Вес	Пол	IQ
A	180	75	1	?

Задачи предсказания бывают

1) Предсказание значения количественного признака Y называется задачей регрессии.
2) Предсказание значения номинального (категориального) признака Y называется задачей классификации.

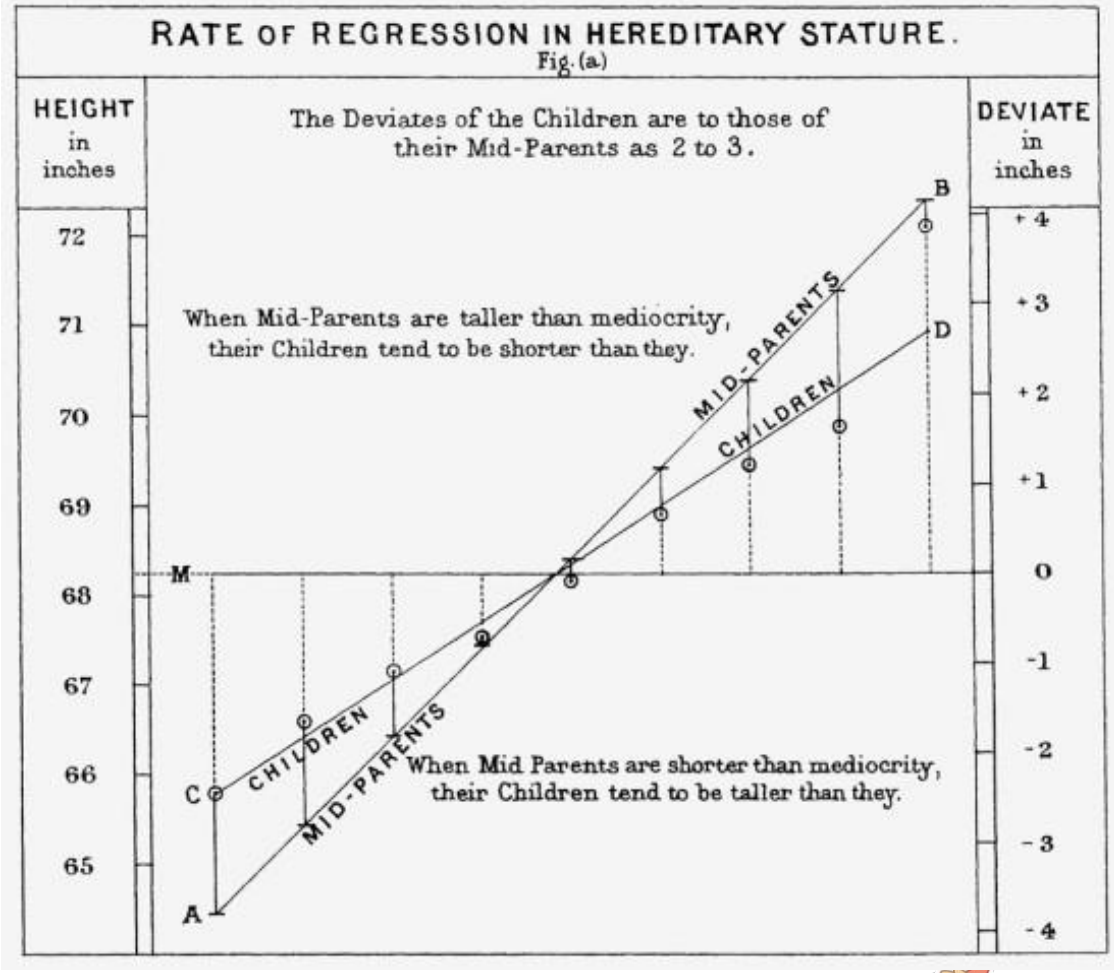
Например, предсказание IQ – это задача регрессии, а предсказание пола – задача классификации.

Откуда слово «регрессия»?

Это тот случай, когда видовое имя используется при описании всего рода (примеры такого типа: «чойс», «ксерокс», «памперс»…)

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

Когда исследовали зависимость роста сына от роста его отца, то было замечено «рост сына приближался (регрессировал) к среднему росту мужчин»



План решения задачи регрессии (для классификации тоже сгодится)

Общий план решения задачи регрессии

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

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

Train	Рост	Вес	Пол	IQ
Вася	170	80	1	100
Даша	165	60	0	80

Test	Рост	Вес	Пол	IQ
Маша	160	50	0	110
Петя	200	70	1	50

Общий план решения задачи регрессии

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

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

Train	Рост	Вес	Пол	IQ
Вася	170	80	1	100
Даша	165	60	0	80

Test	Рост	Вес	Пол	IQ
Маша	160	50	0	110
Петя	200	70	1	50

Оценивание качества модели по тестовой выборке

Допустим по тренировочной выборке мы научились предсказывать целевой признак Y. Как оценить качество предсказаний по тестовой выборке?
(В таблице в скобках указаны предсказанные значения.)

Test	Рост	Вес	Пол	IQ
Маша	160	50	0	110 (100)
Петя	200	70	1	50 (70)

Показатели качества регрессии

MAE (средняя абсолютная ошибка):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_i'|$$

(n – объем тестовой выборки, y_i – истинные y_i' – предсказанные значения). В нашем примере: MAE=1/2(|110-100|+|50-70|)=15.

Test	Рост	Вес	Пол	IQ
Маша	160	50	0	110 (100)
Петя	200	70	1	50 (70)

Показатели качества регрессии

MAPE (средняя абсолютная ошибка в процентах):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - y_i'|}{y_i} * 100\%$$

(n – объем тестовой выборки, y_i – истинные y_i' – предсказанные значения). В нашем примере: MAPE=1/2(|110-100|/110+|50-70|/50)*100%=25%.

Test	Рост	Вес	Пол	IQ
Маша	160	50	0	110 (100)
Петя	200	70	1	50 (70)

Замечание о точности регрессии

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

Замечание о точности регрессии

То есть если модель предсказания была построена по данным

Train	Рост	Вес	Пол	IQ(Y)
Вася	170	80	1	100
Даша	165	60	0	80

а вы решили приколоться и подать ей на вход объект

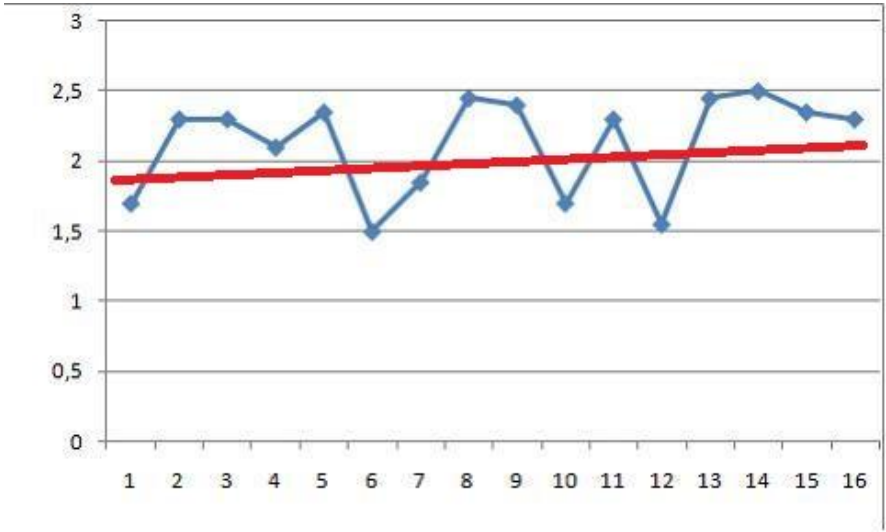
Joke	Рост	Вес	Пол	IQ(Y)
Даша	165	60	0	??

то модель не обязана выдавать вам 80 в качестве ответа.

Как тут предсказать Y по значению X

Узлы – это объекты тренировочной выборки. Предсказывать значения Y в новых точках можно с помощью а) ломаной линии б) красной прямой.

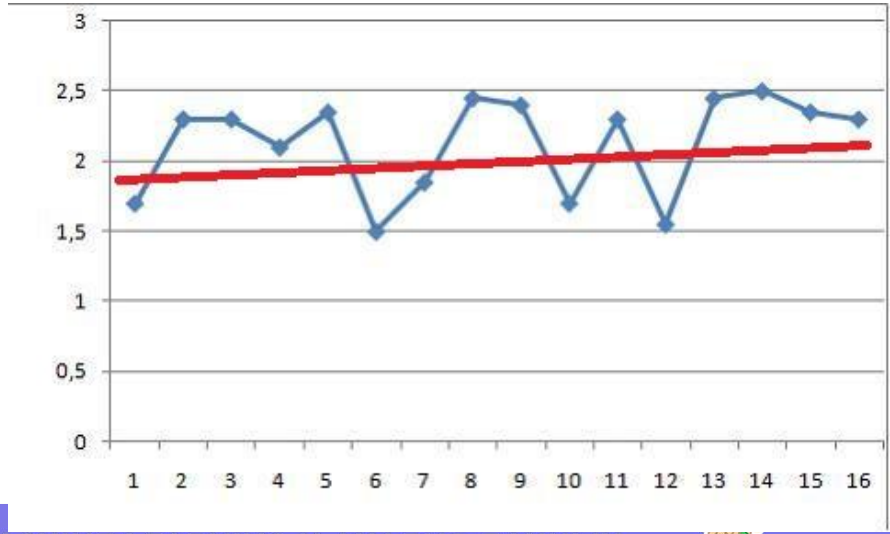
Какие «минусы» имеет предсказание с помощью ломаной линии?



Как тут предсказать Y по значению X

«Минусы» ломаной линии:

1. Модель предсказания имеет сложность сравнимую с объемом данных.
2. Модель нельзя никак проинтерпретировать.
3. Нет уверенности, что на тестовой выборке будут небольшие ошибки.



Красная прямая – это линия регрессии (см. след. слайды)

Формальное определение линейной модели

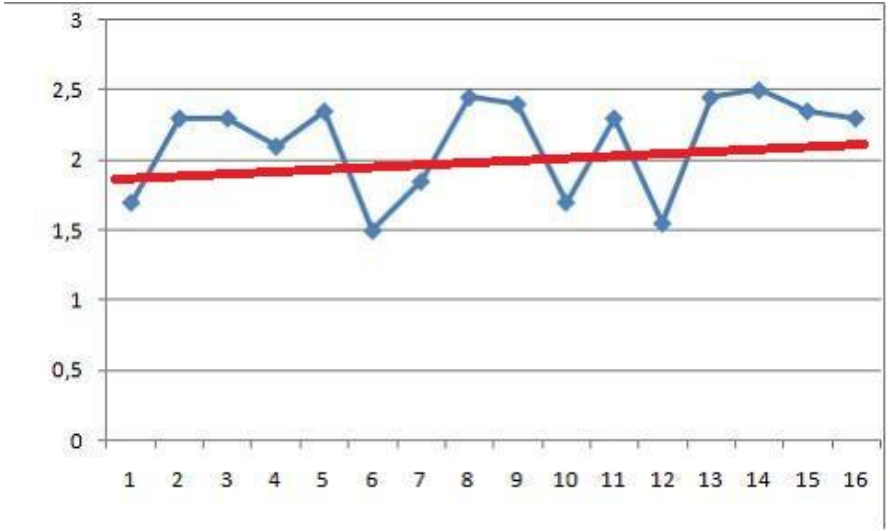
Модель регрессии называется линейной, если значение предсказываемого признака Y вычисляется как сумма известных признаков $X_1, X_2, \ldots X_m$, взятых с некоторыми коэффициентами.

$$y' = w_1 x_1 + w_2 x_2 + \cdots + w_m x_m + w_0$$

Задача заключается в нахождении оптимальных весов (коэффициентов) w_i.

Как искать веса w_i?

Принцип: для объектов тренировочной нужно минимизировать отклонение предсказываемых значений от истинных значений признака Y. (Важно: лин. регрессия неустойчива к выбросам – выбросы нужно заранее удалить)



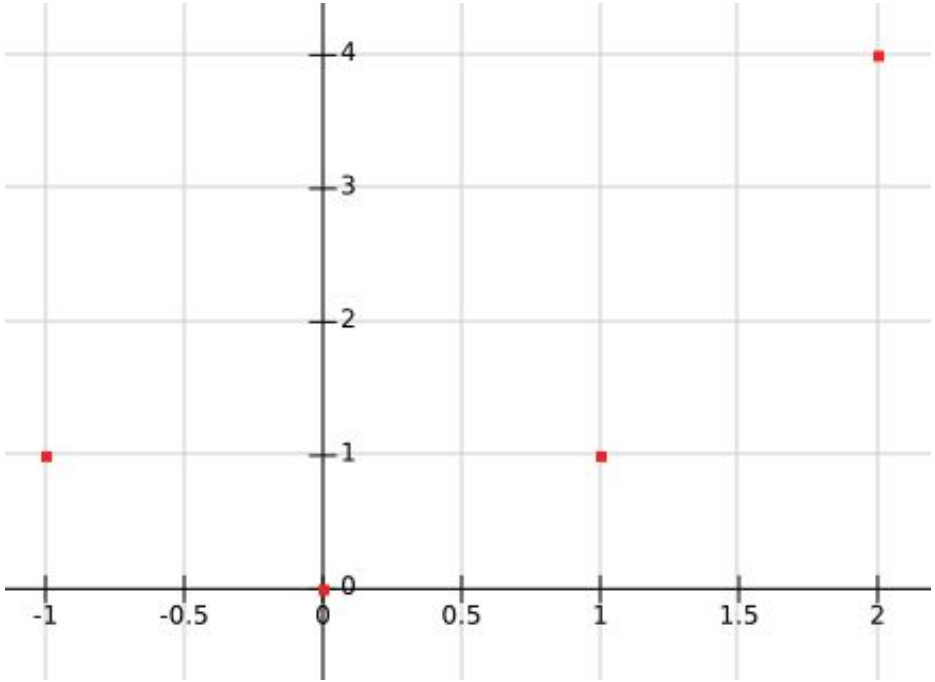
Поучительный пример

Есть данные тренировочной выборки:
Так как нецелевой признак один X, то модель предсказания признака Y будем искать в виде
$y'=w_1x+w_0$. *Нужно найти оптимальные* w_1, w_0.

Объект	X	Y
A	-1	1
B	0	0
C	1	1
D	2	4

Простой, но поучительный пример

Можно нанести точки на плоскость. Можете представить, как нужно нарисовать прямую, чтобы сумма отклонений точек от этой прямой была минимальной?



Что нужно минимизировать?

Отклонение истинного от предсказанного значения равно $|y'-y|=|w_1x+w_0-y|$, эту величину нужно минимизировать для всех объектов тренировочной выборки. Иными словами, чем меньше выражение $L(w_1,w_0)=|w_1x_1+w_0-y_1|+|w_1x_2+w_0-y_2|+|w_1x_3+w_0-y_3|+|w_1x_4+w_0-y_4|$
тем лучше.

Поиск точки минимума этой функции осложняется тем, что модуль – функция НЕ…(какая?)

Что нужно минимизировать?

… - не дифференцируемая (у нее нет производной). Поэтому на практике минимизируют несколько иную функцию: сумму квадратов отклонений.

Т.е. в нашем примере нужно найти минимальное значение функции:

$$L(w_1,w_0)=(w_1(-1)+w_0-1)^2+(w_1 0+w_0-0)^2+$$
$$+(w_1 1+w_0-1)^2+(w_1 2+w_0-4)^2=$$
$$=6w_1^2+4w_0^2+4w_1w_0-16w_1-12w_0+18$$

Объект	X	Y
A	-1	1
B	0	0
C	1	1
D	2	4

А как найти точку минимума?

$L(w_1,w_0)=6w_1^2+4w_0^2+4w_1w_0-16w_1-12w_0+18$

Вычисляем частные производные:

$$\frac{\partial L}{\partial w_1} = 12w_1 + 4w_0 - 16$$

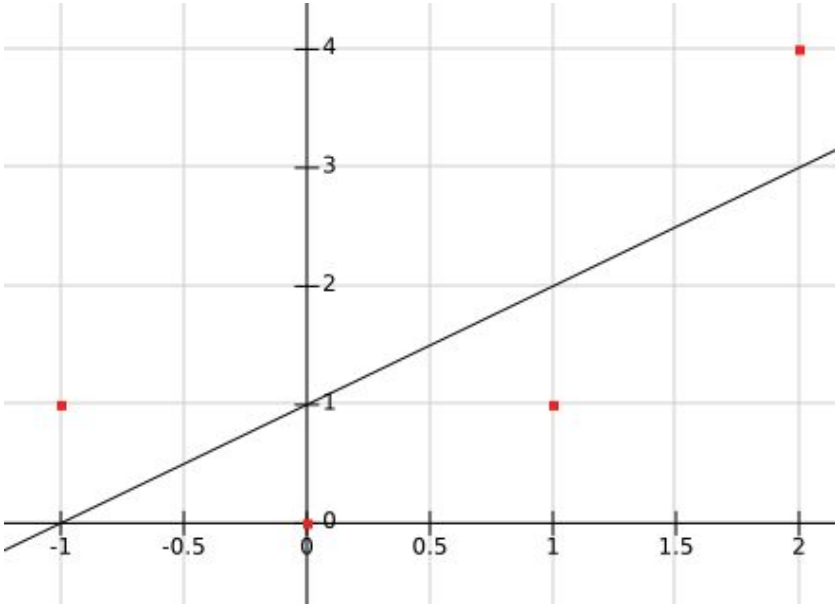
$$\frac{\partial L}{\partial w_0} = 4w_1 + 8w_0 - 12$$

Приравниваем их к 0, решаем систему уравнений

$$\begin{cases} 12w_1 + 4w_0 - 16 = 0, \\ 4w_1 + 8w_0 - 12 = 0 \end{cases}$$

Получаем модель предсказания

Решая систему уравнений, получаем $w_1=1$, $w_0=1$.
То есть модель предсказания признака Y будет
$y'=x+1$



В общем случае

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

Есть готовые формулы, позволяющие сразу вычислить вид частных производных по всем w_i. Таким образом, основная трудоемкость при построении линейной регрессии заключается в решении системы линейных уравнений на последнем шаге.

Основные проблемы линейной регрессии связаны с системой линейных уравнений

Получаемая при построении линейной регрессии система линейных уравнений может:

1) не иметь решений;
2) иметь более одного решения;
3) случаи (1,2) возникают, когда определитель системы $|A|=0$, но при $|A|\approx 0$ тоже плохо.

Эти проблемы возникают, когда

… между нецелевыми признаками существует линейная зависимость или сильная корреляция (это еще называют «проблемой мультиколлинеарности»)!!!

Рассмотрим пример, когда между нецелевыми признаками X1, X2 есть линейная зависимость (они вообще совпадают))). В этом случае уравнение регрессии ищется в виде

$$y' = w_1 x_1 + w_2 x_2 + w_0$$

Объект	X1	X2	Y
A	-1	-1	1
B	0	0	0
C	1	1	1
D	2	2	4

Эти проблемы возникают, когда

Уравнение регрессии ищется в виде
$y' = w_1 x_1 + w_2 x_2 + w_0$
Аналогично составляется выражение для функции $L(w_1, w_2, w_0)$ и считаются ее частные производные:

$$\frac{\partial L}{\partial w_1} = 12w_1 + 12w_2 + 4w_0 - 16$$

$$\frac{\partial L}{\partial w_2} = 12w_1 + 12w_2 + 4w_0 - 16$$

$$\frac{\partial L}{\partial w_0} = 4w_1 + 4w_2 + 8w_0 - 12$$

Объект	X1	X2	Y
A	-1	-1	1
B	0	0	0
C	1	1	1
D	2	2	4

Эти проблемы возникают, когда

Нужно решать систему уравнений

$$\begin{cases} 12w_1 + 12w_2 + 4w_0 = 16 \\ 12w_1 + 12w_2 + 4w_0 = 16 \\ 4w_1 + 4w_2 + 8w_0 = 12 \end{cases}$$

здесь фактически 2 (а не 3) уравнения. Можно показать, что множество решений этой системы бесконечно. Любая точка вида *(a,1-a,1)* $a \in R$ ей удовлетворяет.

А что в этом плохого? Ну, возьмём любое из решений, например *(0.5,0.5,1)* этой системы, получим модель $y' = 0.5x_1 + 0.5x_2 + 1$…

Какие тут проблемы

Любая модель регрессии, чьи параметры удовлетворяют системе имеют одинаковое значение выражения L. То есть все эти модели одинаково хороши (или плохи).

Проблема первая: зависимость целевого признака Y никак нельзя проинтерпретировать (и заказчик никогда вас не поймет). Например, для последнего пример можно взять модель

$y' = x_1 + 1$

а можно и

$y' = x_2 + 1$

Какие тут проблемы

Одна модель

$y' = x_1 + 1$

утверждает, что признак Y не зависит от признака X2.

А модель

$y' = x_2 + 1$

утверждает, что признак Y не зависит от признака X1.

Кто прав?

Объект	X1	X2	Y
A	-1	-1	1
B	0	0	0
C	1	1	1
D	2	2	4

Какие тут проблемы

Проблема вторая: возможна Б-О-О-Ольшая ошибка на объектах, не попавших в тренировочную выборку. Это произойдет, когда в качестве коэффициентов будут выбраны большие числа.

Например, если в прошлом примере взять $y'=10^6x_1+(1-10^6)x_2+1$, то

Объект	X1	X2	y	y’
B (из трен. выборки)	0	0	0	1
B’ (близкий объект)	0.01	0	0	1000

Какие тут проблемы

Проблема третья: если между нецелевыми признаками существует сильная корреляция, то при построении регрессии придется решать систему с почти нулевым определителем. При численном решении таких систем возникают умножения (деления) на большие (малые) числа – удовольствие и качество результата так себе.

Что нужно делать, чтобы найти хорошую модель регрессии?

1. Отбор признаков. Нужно удалять нецелевые признаки, которые линейно зависят от других или имеют высокую корреляцию с другими признаками.
2. Можно вводить новые ограничения на коэффициенты регрессии. Например, можно ограничивать их величину по модулю («лассо», см. ниже).
3. Коэфф. регрессии можно явно не ограничивать а пытаться их минимизировать («регуляризация», см.ниже)

Регуляризация

Основная идея

Нужно стремится сделать коэфф. регрессии небольшими, чтобы не было такой фигни как здесь (модель $y'=10^6x_1+(1-10^6)x_2+1$).

Объект	X1	X2	y	y’
B (из трен. выборки)	0	0	0	1
B’ (близкий объект)	0.01	0	0	1000

То есть нужно вместо выражения $L(w_1,w_2,\dots,w_m,w_0)$ минимизировать…

$R=L(w_1,w_2,\dots,w_m,w_0)+w_1^2+w_2^2+\dots+w_m^2+w_0^2$

Основная идея

нужно минимизировать

$$R=L(w_1,w_2,\ldots,w_m,w_0)+C(w_1^2+w_2^2+\ldots+w_m^2+w_0^2),$$

где C – заданная константа.

На след. слайдах рассмотрим регуляризацию при C=1, то есть будем минимизировать выражение

$$R=L(w_1,w_2,\ldots,w_m,w_0)+w_1^2+w_2^2+\ldots+w_m^2+w_0^2$$

Для нашего примера регуляризация дает…

Уравнение регрессии ищется в виде
$y' = w_1 x_1 + w_2 x_2 + w_0$
Составляем выражение для R и считаем его частные производные:

$$\frac{\partial R}{\partial w_1} = (12w_1 + 12w_2 + 4w_0 - 16) + 2w_1$$
$$\frac{\partial R}{\partial w_2} = (12w_1 + 12w_2 + 4w_0 - 16) + 2w_2$$
$$\frac{\partial R}{\partial w_0} = (4w_1 + 4w_2 + 8w_0 - 12) + 2w_0$$

Объект	X1	X2	Y
A	-1	-1	1
B	0	0	0
C	1	1	1
D	2	2	4

Для нашего примера регуляризация дает…

… и решаем систему уравнений (теперь-то она имеет единственное решение)

$$\begin{cases} 12w_1 + 12w_2 + 4w_0 + 2w_1 = 16 \\ 12w_1 + 12w_2 + 4w_0 + 2w_2 = 16 \\ 4w_1 + 4w_2 + 8w_0 + 2w_0 = 12 \end{cases}$$

получаем (приближенно) $w_1=1,36$ $w_2=-0.25$ $w_0=0.75$

Основной вопрос регуляризации

Какое значение выбрать для константы *C*?

Совет:

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

П.С. Способы выбора оптимальных параметров (например, кросс-валидация) будут рассмотрены на след. лекциях.

Лассо

Основная идея

Нужно (как и в регуляризации) стремится сделать коэфф. регрессии небольшими, добавляя новое условие

$$\begin{cases} L(w_1, w_2, \dots, w_n, w_0) \to min \\ \sum_{i=1}^{m} |w_i| < C \end{cases}$$

где C – некоторая константа (это и есть лассо).

Оптимальные w_i здесь уже нельзя найти простым способом, тут нужно использовать крутые методы теории оптимизации.

Основной вопрос лассо (как и регуляризации)

Какое значение выбрать для константы *C*?

Совет (такой же как и для регуляризации):

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

П.С. Способы выбора оптимальных параметров (например, кросс-валидация) будут рассмотрены на след. лекциях.

Полиномиальная регрессия

Хочется, чтобы...

1) можно было бы строить нелинейные модели, а, например, полиномиальные (полиномиальная регрессия)
2) алгоритм их построения не сильно бы отличался от алгоритма линейной регрессии.

Хочется, чтобы...

1) можно было бы строить нелинейные модели, а, например, полиномиальные (полиномиальная регрессия)
2) алгоритм их построения не сильно бы отличался от алгоритма линейной регрессии.

Добавляй новые столбцы

В нашем примере можно добавить новый признак X2 (являющийся квадратом признака X).

И запустить стандартный алгоритм построения линейной регрессии. То есть будем искать зависимость

$y' = w_1 x_1 + w_2 x_2 + w_0$,

по фактически это уже не линейная, а квадратичная зависимость.

Объект	X	X2	Y
A	-1	1	1
B	0	0	0
C	1	1	1
D	2	4	4

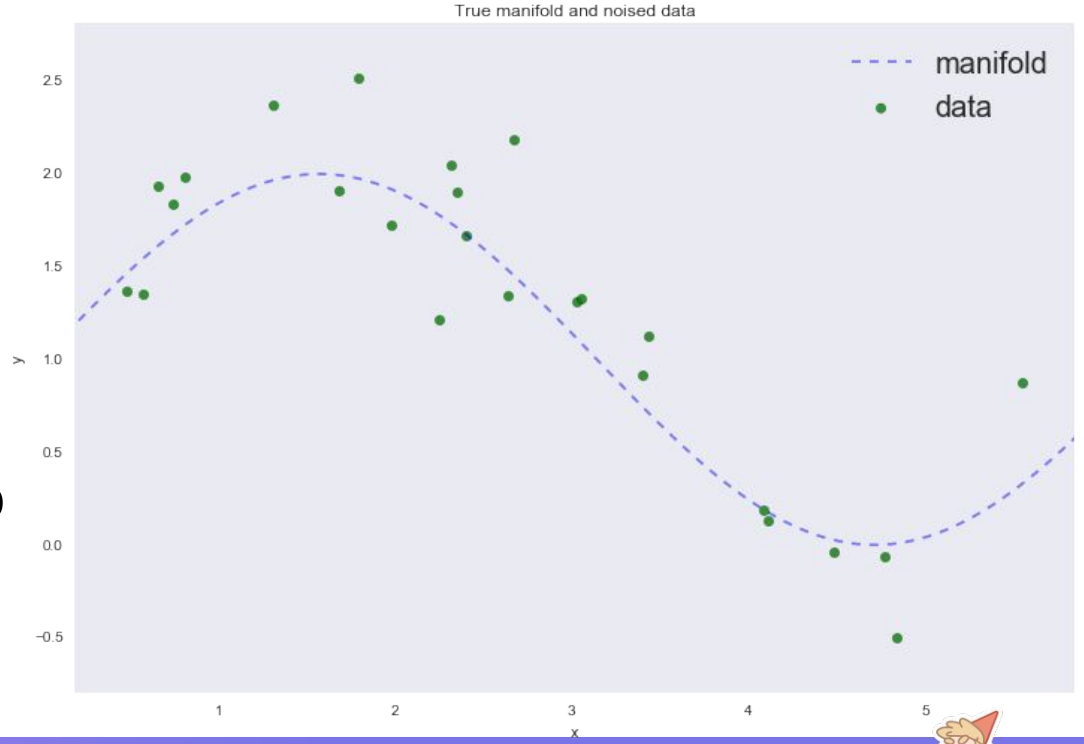
Добавляй новые столбцы

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

Объект	X	Z	XX	ZZ	XZ	Y
A	1	3	1	9	3	100
B	2	4	4	16	8	200

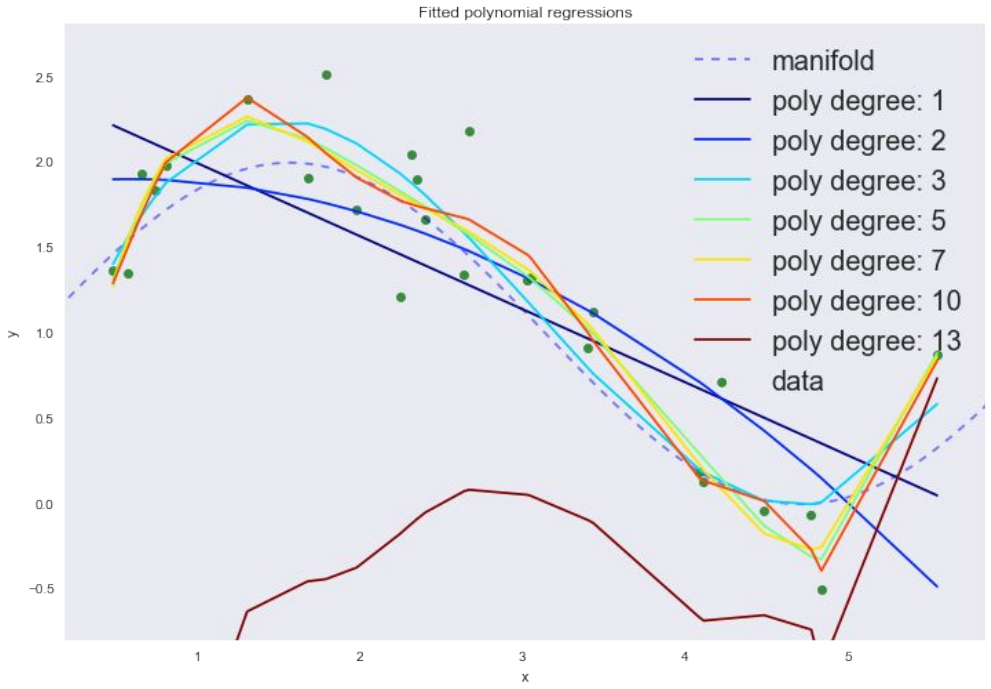
Примеры полиномиальной регрессии

Здесь значение признака Y было получено как sin(X)+(небольшое случайное число). Как можно предсказывать значение Y с помощью полиномов от X?



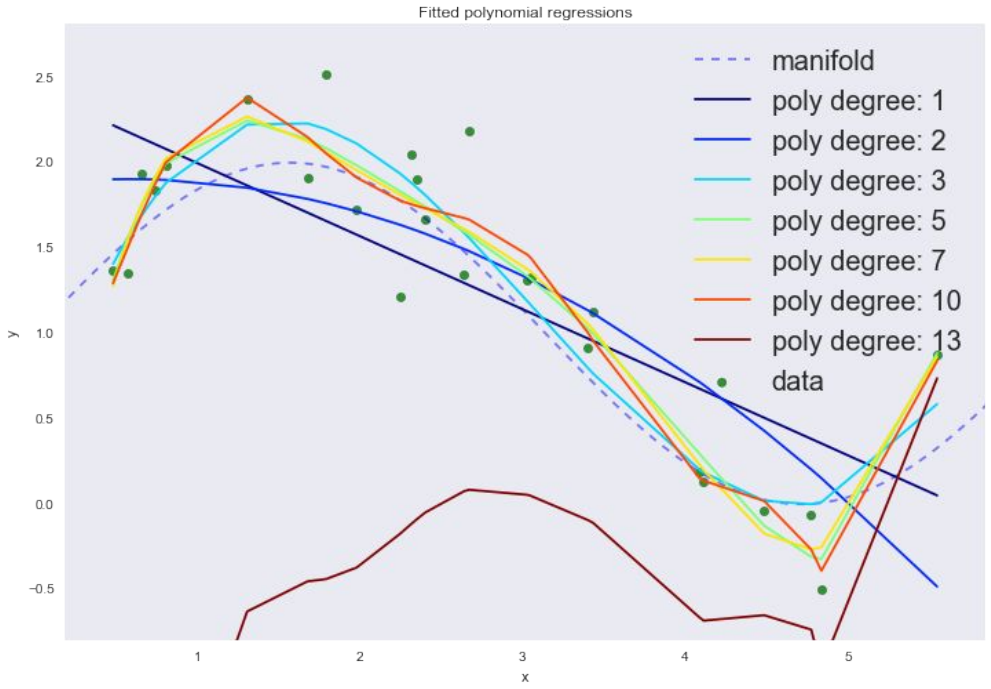
Примеры полиномиальной регрессии

Уравнения полиномиальных регрессий приведены на рисунке. С полиномом 13-й степени случилась фигня, так как при вычислении его коэффициентов возникла система с нулевым определителем.



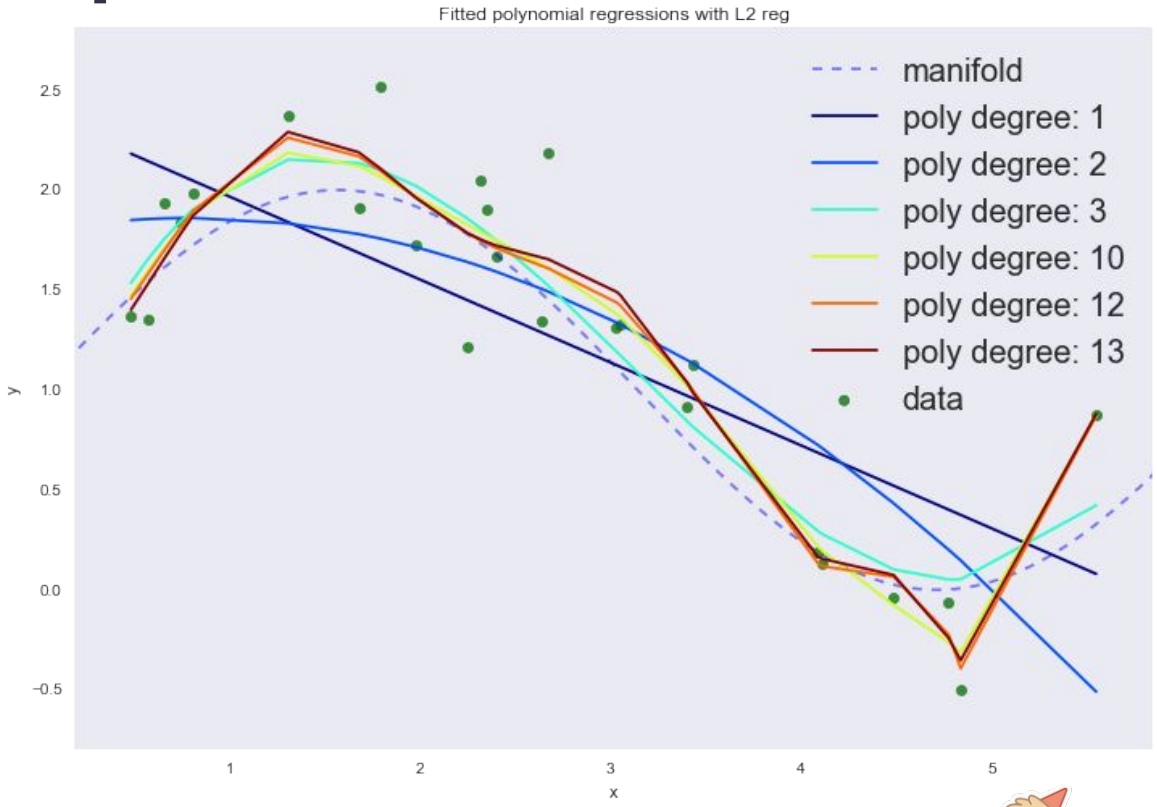
Примеры полиномиальной регрессии

Вывод: увеличение степени полинома не обязательно увеличивает точность предсказания!!!



Полиномы+регуляризация

Показаны уравнения полиномиальной регрессии с регуляризацией.



Использованная литература

1. https://habrahabr.ru/company/ods/blog/323890/#1-lineynaya-regressiya
2. сайт http://www.machinelearning.ru, статьи «Регрессия» и «Мультиколлинеарность»
3. https://habrahabr.ru/company/ods/blog/322076/ (последние 3 графика оттуда)