

Введение в анализ данных

Лекция 7

Линейная регрессия

Евгений Соколов

sokolov.evg@gmail.com

НИУ ВШЭ, 2016

Напоминание

- \mathbb{X} — пространство объектов, \mathbb{Y} — пространство ответов
- $x = (x^1, \dots, x^d)$ — признаковое описание
- $X = (x_i, y_i)_{i=1}^{\ell}$ — обучающая выборка
- $a(x)$ — алгоритм, модель
- $Q(a, X)$ — функционал ошибки алгоритма a на выборке X
- Обучение: $a(x) = \arg \min_{a \in \mathcal{A}} Q(a, X)$

План на сегодня

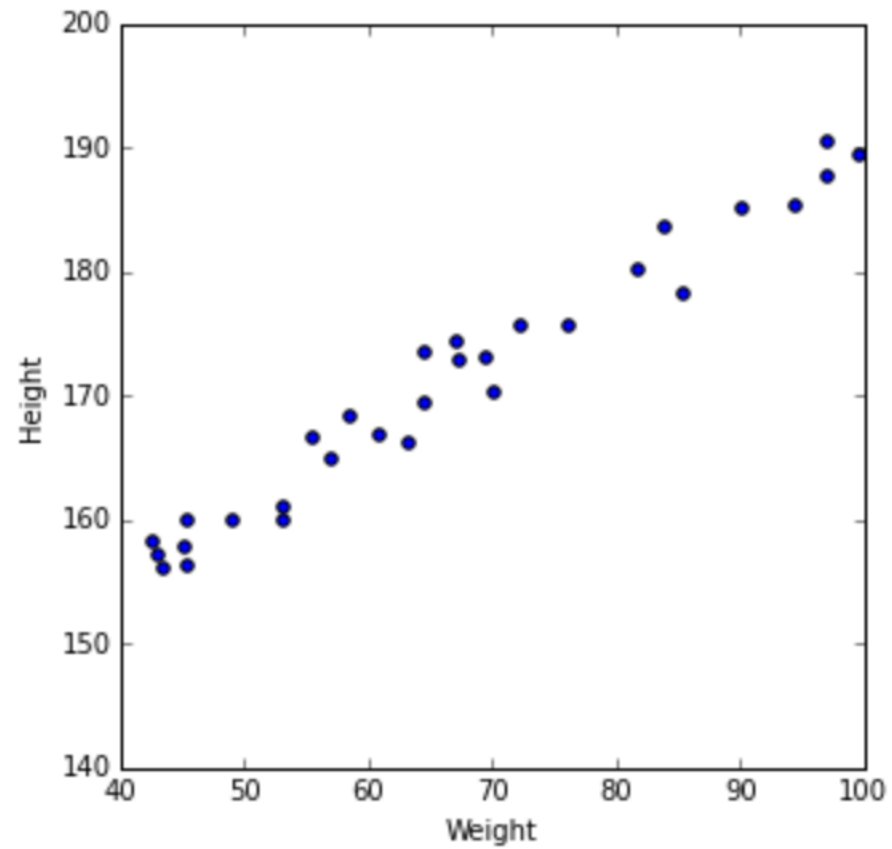
- Линейные модели в регрессии
- Среднеквадратичная ошибка
- Обучение линейной регрессии
- Переобучение и регуляризация
- Подготовка данных

Модель линейной регрессии

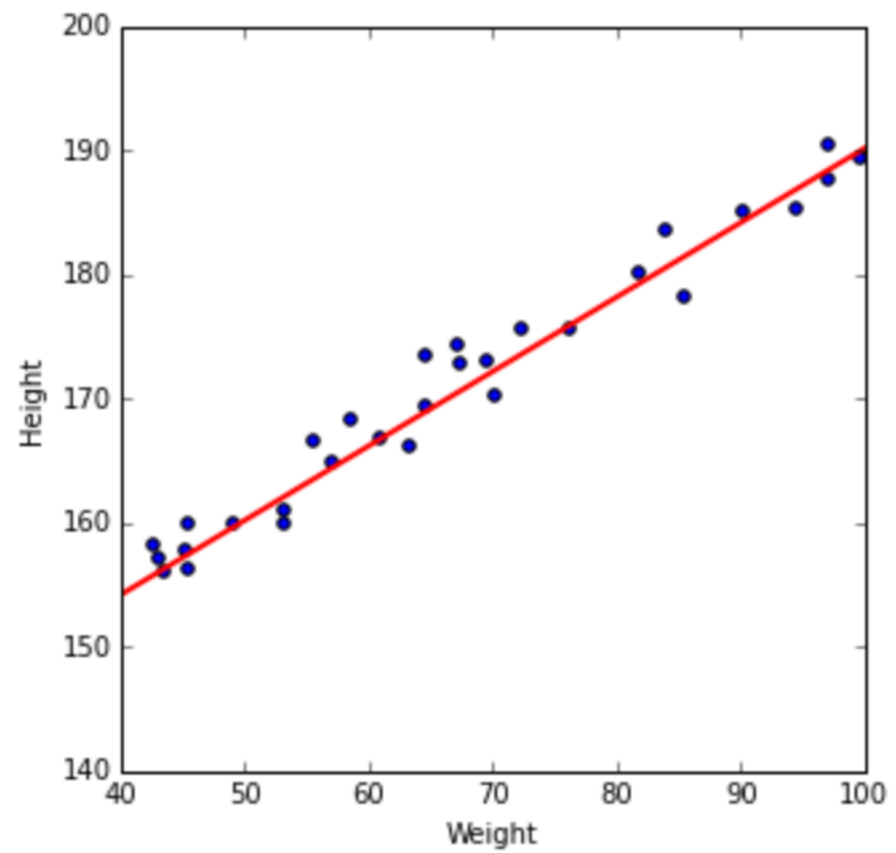
Задача регрессии

- Вещественные ответы: $\mathbb{Y} = \mathbb{R}$
- Определение возраста пользователя
- Предсказание цены квартиры
- Предсказание рейтинга фильма
- Предсказание числа ретвитов

Одномерная выборка



Одномерная выборка



Парная регрессия

- Простейший случай: один признак
- Модель: $a(x) = w_1x + w_0$
- Два параметра: w_1 и w_0
- Одна из простейших моделей

Линейная регрессия

- Взвешенная сумма признаков:


$$a(x) = w_0 + w_1x^1 + \dots + w_dx^d$$

- x^1, x^2, \dots, x^d — значений признаков
- $w_0, w_1, w_2, \dots, w_d$ — параметры
- w_0 — смещение

Линейная регрессия

- Взвешенная сумма признаков:

$$a(x) = w_0 + w_1x^1 + \dots + w_dx^d$$

- x^1, x^2, \dots, x^d — значений признаков
 - $w_0, w_1, w_2, \dots, w_d$ — параметры
 - w_0 — смещение
- 

Единичный признак

$$a(x) = w_0 * 1 + w_1 x^1 + \dots + w_d x^d$$

- w_0 — как бы коэффициент при единичном признаке
- Добавим его!

$$\begin{pmatrix} x_{11} & \dots & x_{1d} & 1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{\ell 1} & \dots & x_{\ell d} & 1 \end{pmatrix}$$

Линейная регрессия

- Везде далее считаем, что среди признаков есть единичный

$$a(x) = w_1 x^1 + \dots + w_d x^d = \langle w, x \rangle$$



Скалярное
произведение

Обучение линейной регрессии

- Как измерить качество алгоритма $a(x) = \langle w, x \rangle$?
- Как выбрать лучший алгоритм с точки зрения этого качества?

Среднеквадратичная ошибка

Функционал ошибки

$a(x)$	y	отклонение
11	10	1
9	10	-1
20	10	10
1	10	-9

Функционал ошибки

- Ошибку надо минимизировать
- Минимизация отклонения ($a(x) - y$) приведёт к провалу

$a(x)$	y	отклонение
11	10	1
9	10	-1
20	10	10
1	10	-9

Функционал ошибки

- Возьмём модуль: $|a(x) - y|$
- Не имеет производной

$a(x)$	y	$ a(x) - y $
11	10	1
9	10	1
20	10	10
1	10	9

Функционал ошибки

- Возведём в квадрат: $(a(x) - y)^2$

$a(x)$	y	$(a(x) - y)^2$
11	10	1
9	10	1
20	10	100
1	10	81

Среднеквадратичная ошибка

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

- MSE (Mean Squared Error)

Среднеквадратичная ошибка

$$Q(w, X) = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2}$$

- RMSE (Root Mean Squared Error)
- В тех же единицах измерения, что и ответы
- Сложные производные из-за корня

Векторный вид MSE

$$Q(w, X) = \frac{1}{\ell} \|Xw - y\|^2$$

- X — матрица объекты-признаки
- y — вектор ответов на обучающей выборке

Обучение линейной регрессии

Задача оптимизации

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \rightarrow \min_w$$

- Гладкая функция
- Выпуклая функция
- Единственный минимум (не всегда)

Градиентный спуск

- Повторять до сходимости:

$$w^t = w^{t-1} - \eta \nabla Q(w^{t-1})$$

- Сходимость: $\|w^t - w^{t-1}\| < \varepsilon$

Градиент

$$\nabla Q(w, X) = \left(\frac{\partial Q}{\partial w_1}, \dots, \frac{\partial Q}{\partial w_d} \right)$$

Производные:

$$\frac{\partial Q}{\partial w_j} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_i^j (\langle w, x_i \rangle - y_i)$$

Нюансы

- Выбор длины шага η — пробуем разные значения
- Выборка должна быть масштабирована
- Признаки не должны коррелировать

Страшная правда

- Векторная запись MSE:

$$Q(w, X) = \frac{1}{\ell} \|Xw - y\|^2$$

- Условие минимума:

$$\nabla Q(w, X) = 0$$

- Что, если попробуем решить эту систему уравнений?

Страшная правда

- Она решается аналитически!

$$w = (X^T X)^{-1} X^T y$$

- Но обращение матрицы — очень сложная операция
- Градиентный спуск гораздо быстрее

Мультиколлинеарность

Мультиколлинеарность

- Наличие зависимостей между признаками
- Приводит к тому, что решений бесконечное число
- Далеко не все из них имеют хорошую обобщающую способность

Линейная зависимость

- Худший случай — линейно зависимые признаки
- Для любого объекта:

$$\alpha_1 x^1 + \dots + \alpha_d x^d = \langle \alpha, x \rangle = 0$$

Линейная зависимость

- Допустим, мы нашли решение w_*
- Модифицируем: $w_1 = w_* + t\alpha$
- (t — число)
- Ответ нового алгоритма на любом объекте:

$$\langle w_1, x \rangle = \langle w_* + t\alpha, x \rangle = \langle w_*, x \rangle + t\langle \alpha, x \rangle = \langle w_*, x \rangle$$

- w_1 — тоже решение!

Коррелирующие признаки

- Тоже плохо
- Сначала разберёмся с корреляцией

Коэффициент корреляции

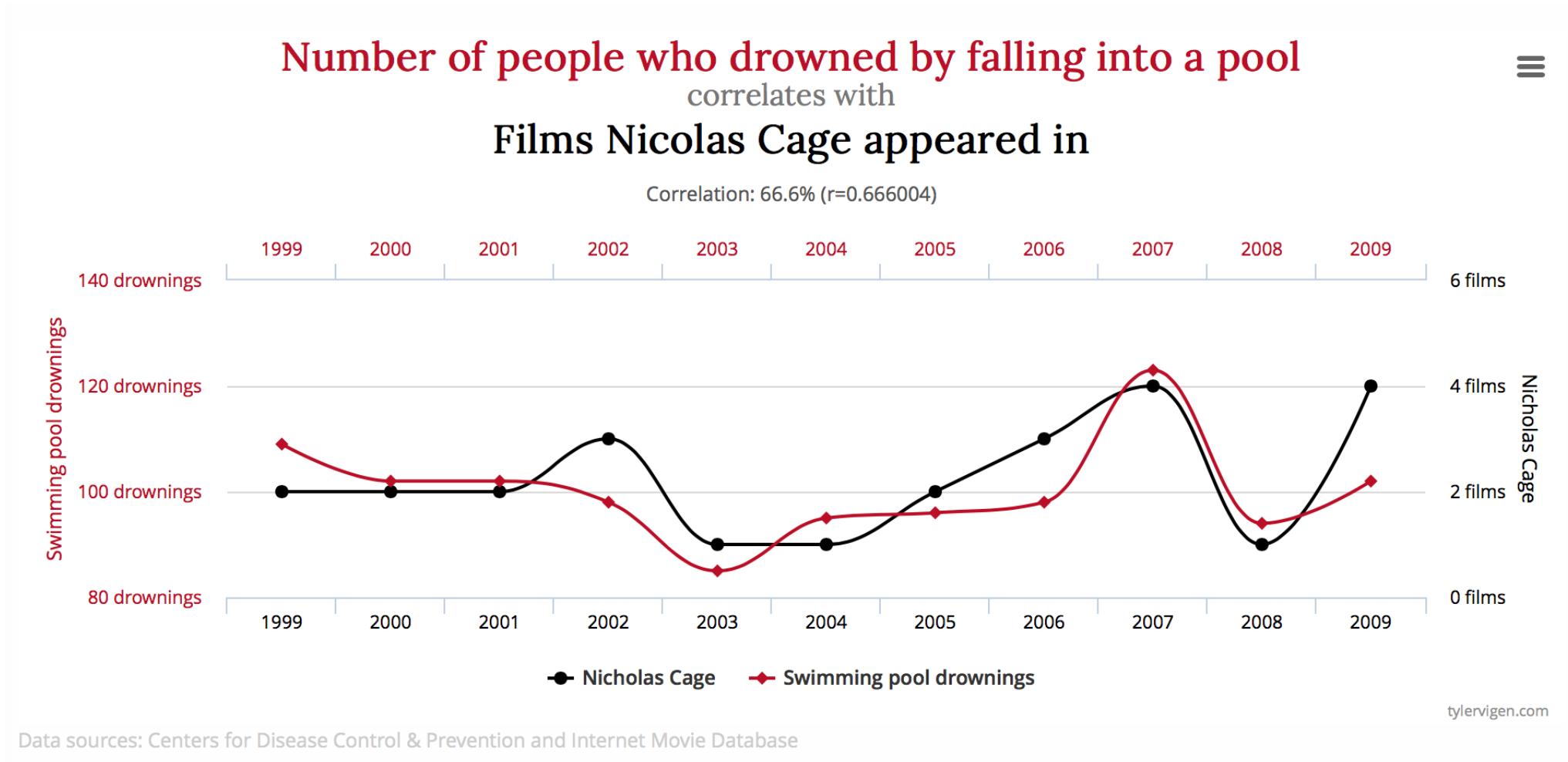
$$\rho(\xi, \eta) = \frac{\mathbb{E}(\xi - \mathbb{E}\xi)(\eta - \mathbb{E}\eta)}{\sqrt{\mathbb{D}\xi \mathbb{D}\eta}}$$

- $\rho(\xi, \eta) \in [-1, +1]$
- Очень грубо: чем ближе к +1 или -1, тем точнее выполнено уравнение

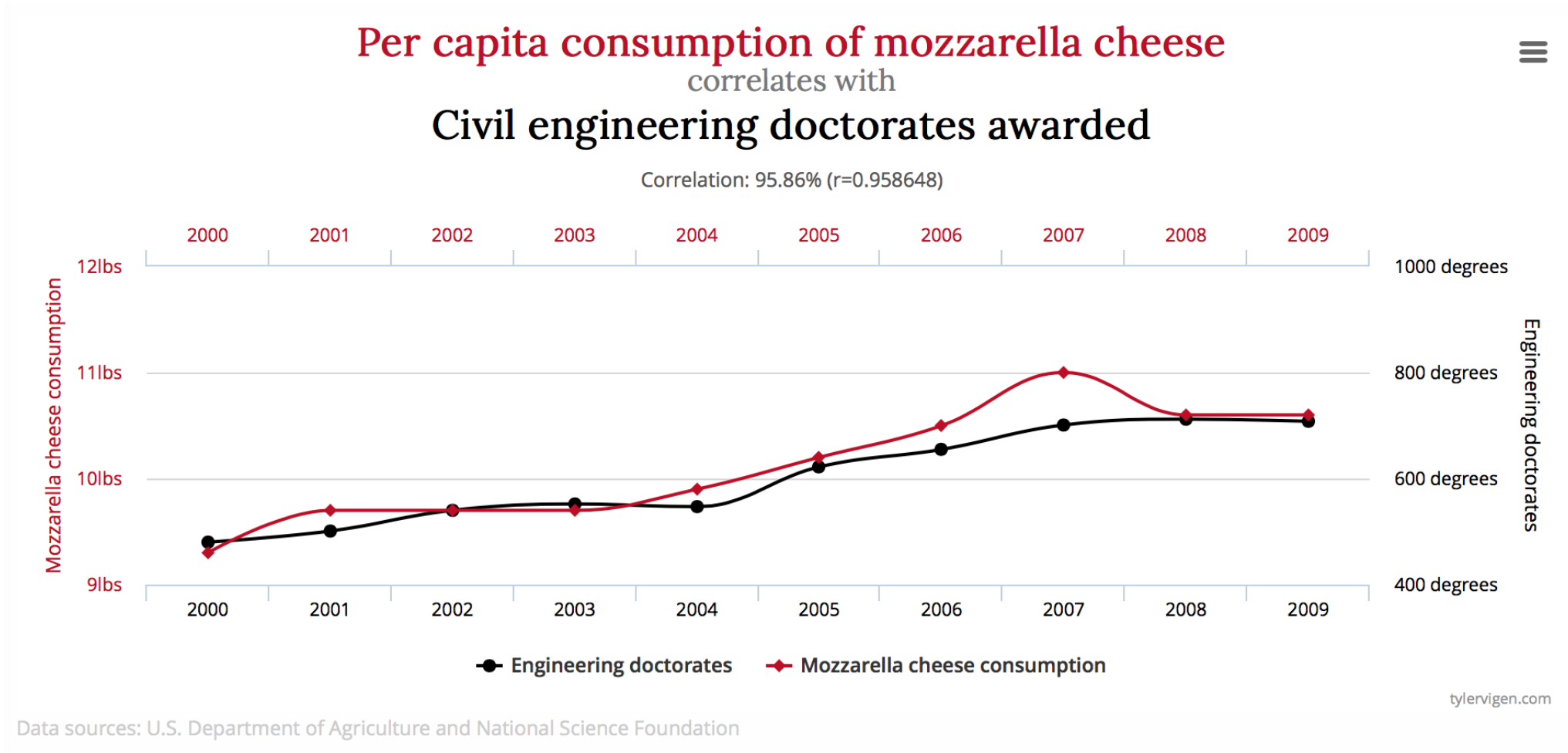
$$\xi = a\eta + b$$

- Мера линейной зависимости

Пример



Пример



Распространённое заблуждение

- Из корреляции следует причинно-следственная связь
 - Это не так!
 - Корреляция означает, что события часто происходят вместе
 - Но никак не следуют друг из друга
-
- Больше примеров: <http://tylervigen.com/spurious-correlations>

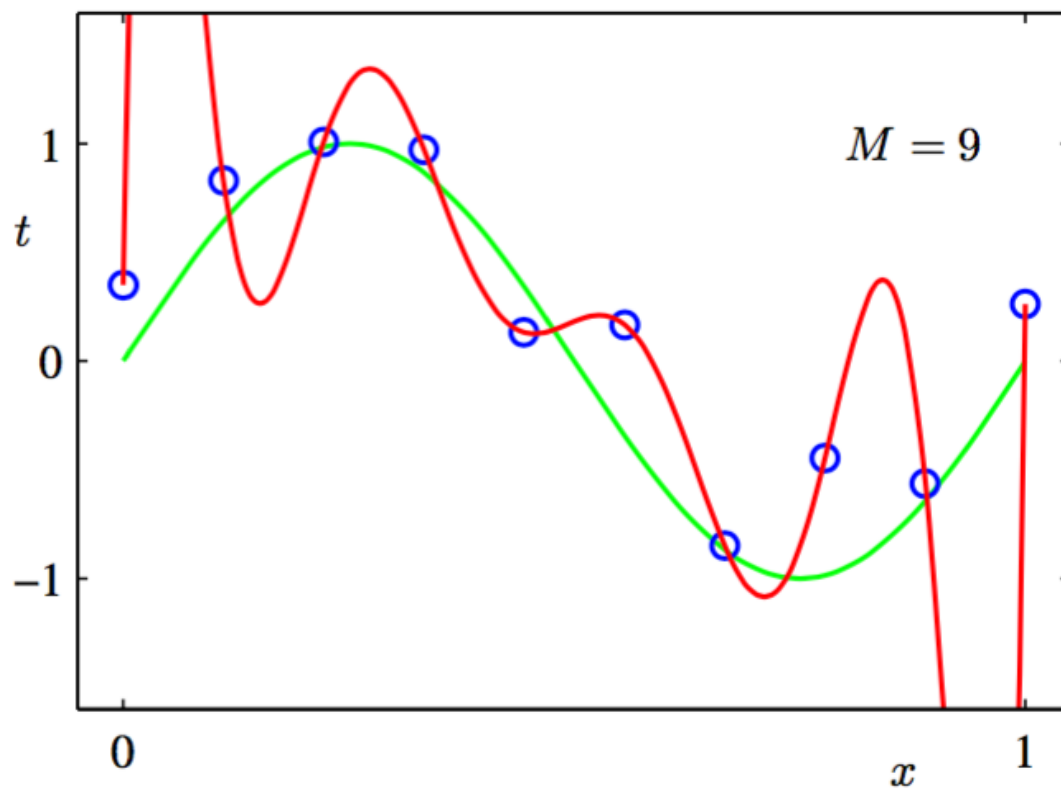
Коррелирующие признаки

- Плохо, если есть коррелирующие признаки
- Решение: отбор признаков или их декорреляция
- В следующих курсах

Переобучение и регуляризация

Пример

- Один признак x
- $a(x) = w_0 + w_1x + w_2x^2 + \dots + w_9x^9$



Пример

- Коэффициенты:

$$a(x) = 0.5 + 13458922x - 43983740x^2 + \dots + 2740x^9$$

- Большие коэффициенты — симптом переобучения
- (эмпирическое наблюдение)

Симптом переобучения

- Большие коэффициенты в линейной модели — это плохо
- Пример: предсказание роста по весу
 - $a(x) = 698x - 41714$
- Изменение веса на 0.01 кг приведет к изменению роста на 7 см
- Не похоже на правильную зависимость

Регуляризация

- Будем штрафовать за большие веса!
- Функционал:

$$Q(w, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \rightarrow \min_w$$

- Регуляризатор:

$$\|w\|^2 = \sum_{j=1}^d w_j^2$$

Регуляризация

- Регуляризованный функционал ошибки:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|^2 \rightarrow \min_w$$

- Всё ещё гладкий и выпуклый

Коэффициент регуляризации

- λ — новый параметр, надо подбирать
- Высокий λ — простые модели
- Низкий λ — риск переобучения
- Нужно балансировать
- Подбор λ — с помощью кросс-валидации (далее в лекциях)

Смысл регуляризации

- Минимизация регуляризованного функционала равносильна решению условной задачи:

$$\begin{cases} \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \rightarrow \min_w \\ \|w\|^2 \leq C \end{cases}$$

L_1 -регуляризация

- L_1 -регуляризатор:

$$\|w\|_1 = \sum_{j=1}^d |w_j|$$

- Регуляризованный функционал ошибки:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \|w\|_1 \rightarrow \min_w$$

L_1 -регуляризация

- Функционал становится негладким
- Сложнее оптимизировать
- Зато производится отбор признаков
- Часть весов в решении будут нулевыми

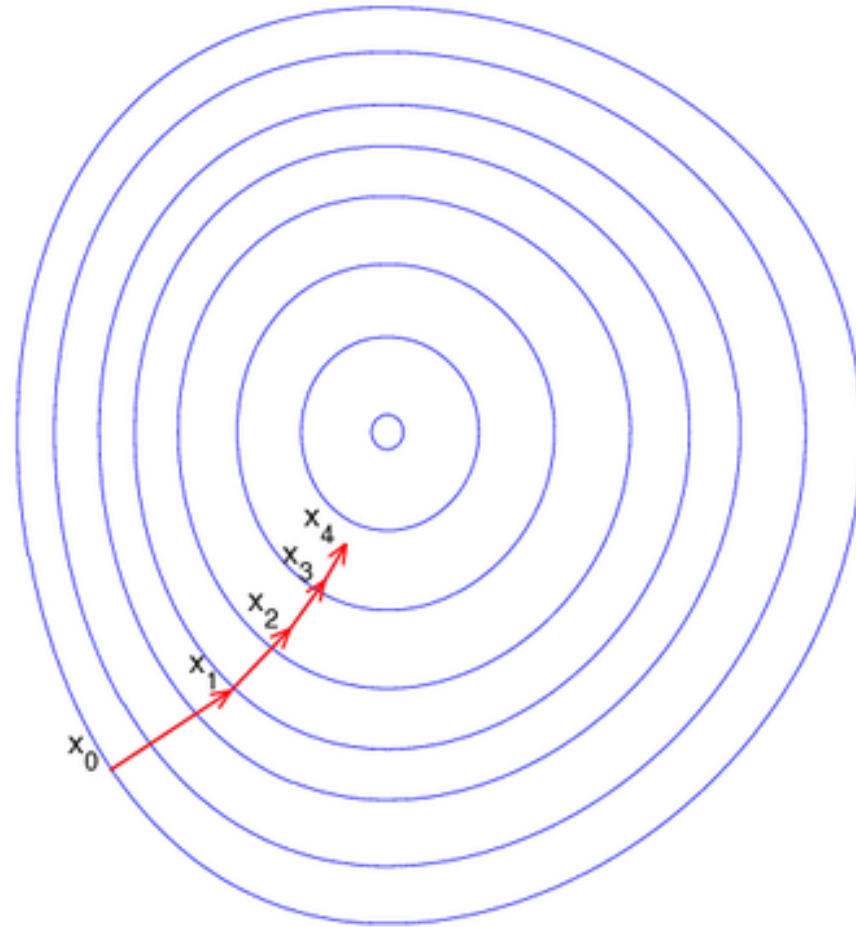
Подготовка признаков

Краткая инструкция

- Вещественные признаки — масштабирование
- Категориальные признаки — one-hot-кодирование

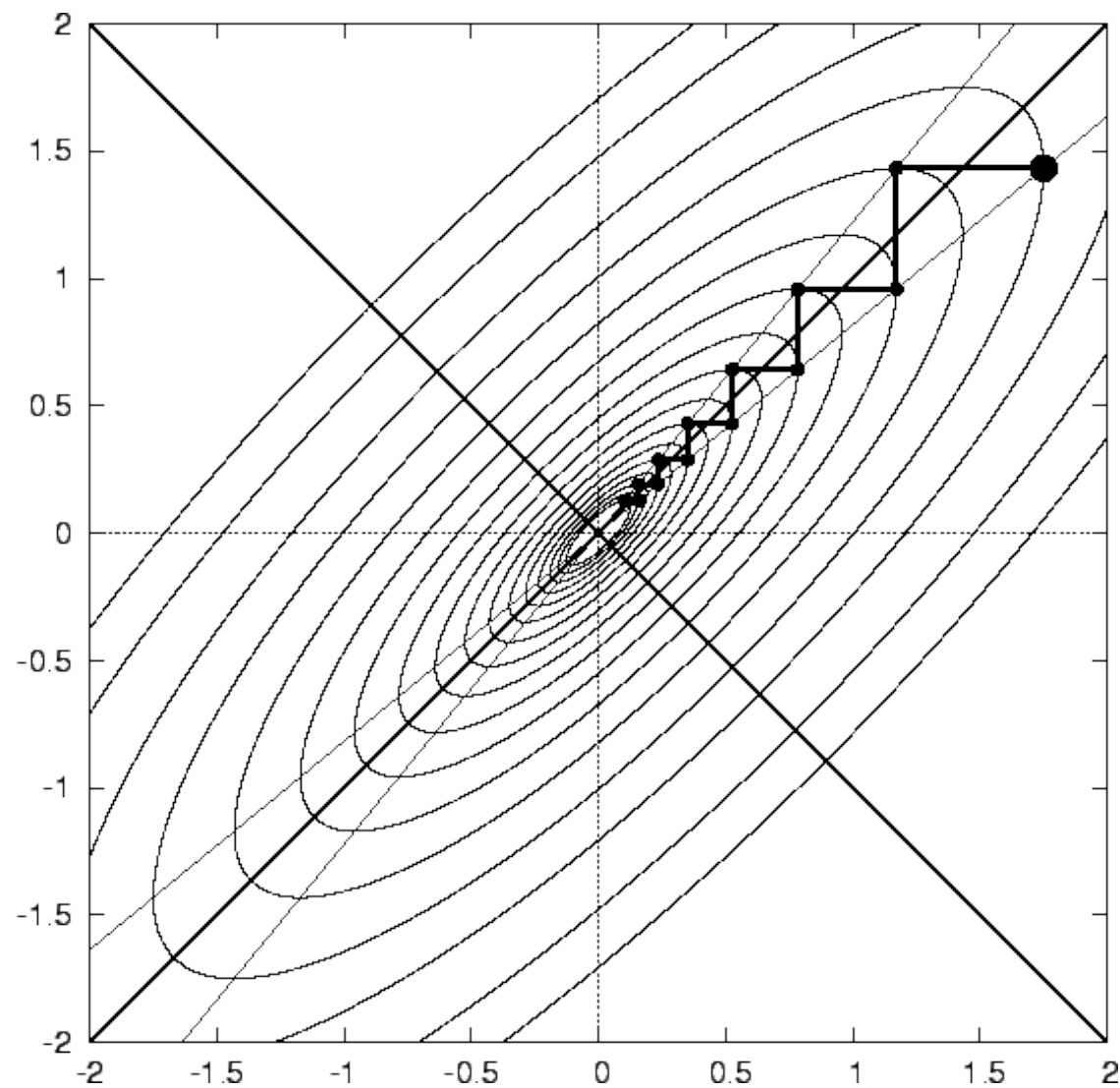
Масштабирование выборки

Хороший случай



Масштабирование выборки

Плохой случай



Масштабирование выборки

- Задача: одобряют ли заявку на грант?
- 1-й признак: сколько успешных заявок было до этого у заявителя
- 2-й признак: год рождения заявителя
- Масштаб: единицы и тысячи
- Все признаки должны иметь одинаковый масштаб

Масштабирование выборки

- Отмасштабируем j -й признак
- Вычисляем среднее и стандартное отклонение признака на обучающей выборке:

$$\mu_j = \frac{1}{\ell} \sum_{i=1}^{\ell} x_i^j$$

$$\sigma_j = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} (x_i^j - \mu_j)^2}$$

Масштабирование выборки

- Отмасштабируем j -й признак
- Вычтем из каждого значения признака среднее и поделим на стандартное отклонение:

$$x_i^j := \frac{x_i^j - \mu_j}{\sigma_j}$$

Категориальные признаки

- Пример: город клиента банка
- Три объекта со значениями [Москва, Санкт-Петербург, Москва]
- Закодируем двумя числовыми признаками:

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

One-hot-кодирование

- Заводим столько новых признаков, сколько значений у категориального
- Каждый соответствует одному возможному значению
- Единице равен тот, который встретился на данном объекте

Практические аспекты

Важность признаков

- Если признаки масштабированы, то вес характеризует важность признака в модели

Term	Coefficient	Std. Error	Z Score
Intercept	2.46	0.09	27.60
lcavol	0.68	0.13	5.37
lweight	0.26	0.10	2.75
age	−0.14	0.10	−1.40
lbph	0.21	0.10	2.06
svi	0.31	0.12	2.47
lcp	−0.29	0.15	−1.87
gleason	−0.02	0.15	−0.15
pgg45	0.27	0.15	1.74

Квадратичные признаки

- Можно добавлять новые признаки, зависящие от исходных
- Модель может восстанавливать более сложные зависимости
- Пример: квадратичные признаки

[площадь, этаж, число комнат]

- Новые признаки:

[площадь, этаж, число комнат,

площадь², этаж², число комнат²,

площадь * этаж, площадь * число комнат, этаж * число комнат,]

Подбор коэффициента регуляризации

- Простейший способ: отложенная выборка
- Дана выборка X
- Делим на две части
- На первой обучаем модели с разными λ
- На второй вычисляем ошибку
- Выбираем λ , с которым ошибка была минимальной



Резюме

- Линейная регрессия — одна из самых простых моделей в машинном обучении
- Функционал качества: среднеквадратичная ошибка
- Обучение: градиентный спуск
- Борьба с переобучением: регуляризация
- Подготовка признаков: масштабирование и one-hot-кодирование