Машинное обучение

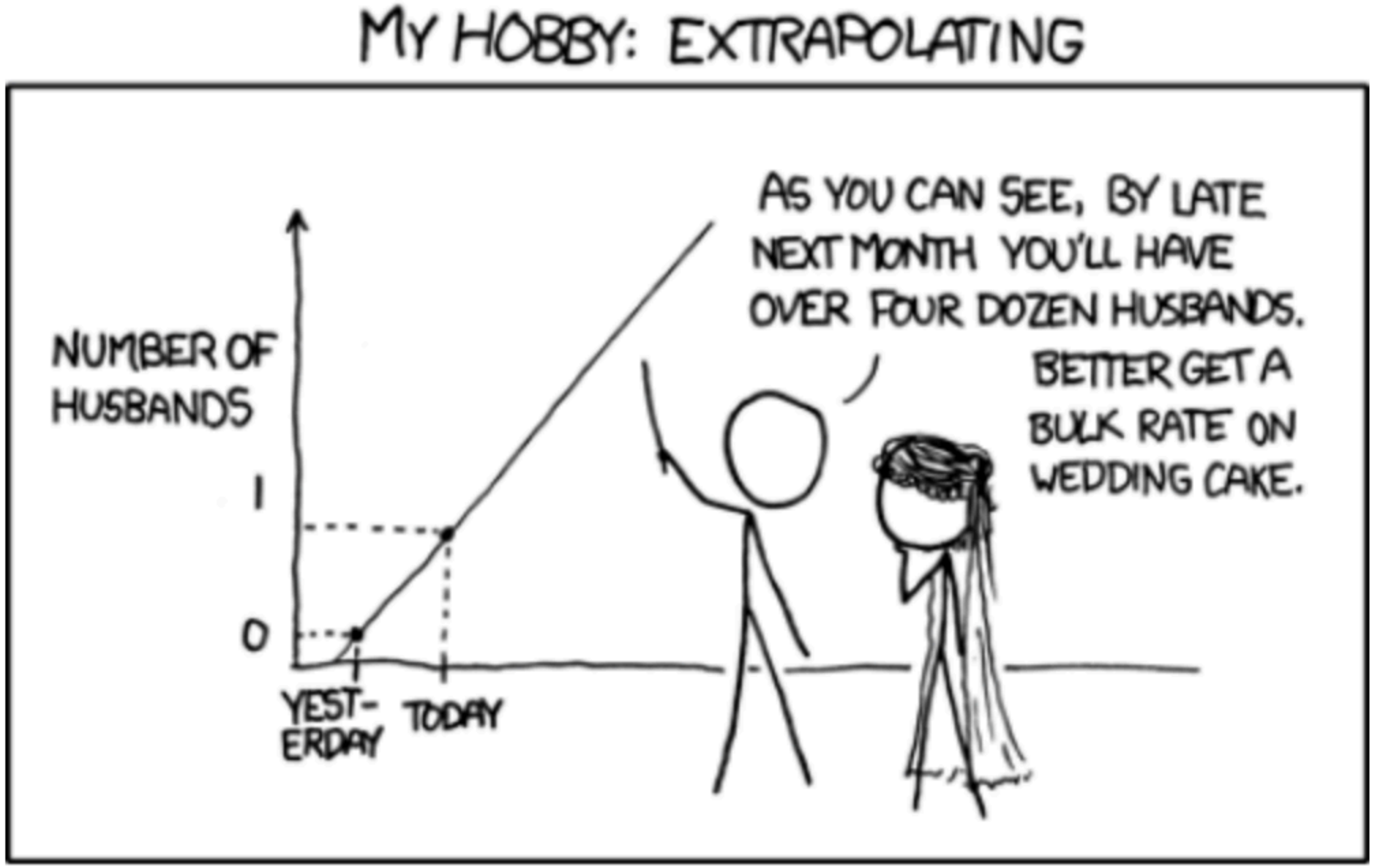
Лекция 3

Линейные модели

Власов Кирилл Вячеславович

2018

Линейные модели



Линейные модели

Линейная модель - взвешенная сумма признаков и член смещения (*bias term*), который также называют свободным членом (intercept term)

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

$\hat{y}$	Предсказываемое значение
n	Количество признаков
$x_1, x_2, \ldots, x_n$	Значения признаков
$\theta_0, \theta_1, \theta_2, \ldots, \theta_n$	Веса признаков и свободный член

Векторная форма:

$$\hat{y} = h_\theta(x) = \theta^T \cdot x$$

θ	Вектор весов и свободный член
x	Вектор значений признаков примера, где $x_0 = 0$

Линейные модели

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

Средняя квадратичная ошибка (Mean Squared Error, MSE):

$$MSE = \frac{1}{l} \sum_{i=1}^{l} \left(f(x_i) - y_i\right)^2$$

$$MSE = \frac{1}{l} \sum_{i=1}^{l} \left(\hat{y}^{(i)} - y_i\right)^2 \to \frac{1}{l} \sum_{i=1}^{l} \left(\theta^T \cdot x^{(i)} - y_i\right)$$

Метод наименьших квадратов

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

Средняя квадратичная ошибка (Mean Squared Error, MSE):

$$MSE = \frac{1}{l} \sum_{i=1}^{l} \left(f(x_i) - y_i\right)^2$$

$$MSE = \frac{1}{l} \sum_{i=1}^{l} \left(\hat{y}^{(i)} - y_i\right)^2 \rightarrow \frac{1}{l} \sum_{i=1}^{l} \left(\theta^T \cdot x^{(i)} - y_i\right)^2 \rightarrow min$$

Аналитический способ поиска оптимальных весов (нормальное уравнение):

$$\hat{\theta} = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot y$$

X Матрица объектов признаков

y Вектор целевой переменной

$\hat{\theta}$ Оптимальный вектор весов, который сводит к минимуму MSE

Подробный и понятный вывод: Открытый курс машинного обучения: Тема 4

Метод наименьших квадратов

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n$$

Средняя квадратичная ошибка (Mean Squared Error, MSE):

$$MSE = \frac{1}{l}\sum_{i=1}^{l}\left(f(x_i) - y_i\right)^2$$

$$MSE = \frac{1}{l}\sum_{i=1}^{l}\left(\hat{y}^{(i)} - y_i\right)^2 \to \frac{1}{l}\sum_{i=1}^{l}\left(\theta^T \cdot x^{(i)} - y_i\right)^2 \to min$$

Аналитический способ поиска оптимальных весов (нормальное уравнение):

$$\hat{\theta} = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot y$$

BLUE (Best Linear Unbiased Estimator)

X Матрица объектов признаков

y Вектор целевой переменной

$\hat{\theta}$ Оптимальный вектор весов, который сводит к минимуму MSE

Подробный и понятный вывод: Открытый курс машинного обучения: Тема 4

Метод наименьших квадратов

Теорема Гаусса — Маркова

оценки метода наименьших квадратов оптимальны в классе линейных несмещённых оценок

Условия для парной регрессии:

1. *модель данных правильно специфицирована*. Нет лишних переменных, или учтены все важные $Y = \theta_0 + \theta \cdot X + \epsilon$
2. *все X* детерминированы и не все равны между собой. Иными словами, переменные не должны быть постоянными.
3. Ошибки не носят систематического характера, то есть $E(\epsilon_i) = 0 \forall i$
4. Дисперсия ошибок одинакова (гомоскедастичность) и равна некоторой $\sigma^2 = const$
5. Ошибки некоррелированы, то есть $cov(\epsilon_i, \epsilon_j) = 0 \forall i, j$

Условия для Множественной регрессии:

1. *модель данных правильно специфицирована*. Нет лишних переменных, или учтены все важные
2. $rang(X) = m$
3. $E(\epsilon_i) = 0 \forall i$
4. $cov(\epsilon_i, \epsilon_j) = 0 \forall i, j$

Реализация в python

Проблемы нормального уравнения

$$\hat{\theta} = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot y$$

Проблемы нормального уравнения

$$\hat{\theta} = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot y$$

$\left(X^T \cdot X\right)$ *– матрица размером n x n, где n – количество признаков*

Вычислительная сложность для обратной матрицы: O(n^3)

Градиентый спуск

$$MSE = \frac{1}{l}\sum_{i=1}^{l}\left(\hat{y}^{(i)} - y_i\right)^2 \rightarrow \frac{1}{l}\sum_{i=1}^{l}\left(\theta^T \cdot x^{(i)} - y_i\right)^2 \rightarrow min$$

Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron

Градиентный спуск

$$MSE = \frac{1}{l}\sum_{i=1}^{l}\left(\hat{y}^{(i)} - y_i\right)^2 \to \frac{1}{l}\sum_{i=1}^{l}\left(\theta^T \cdot x^{(i)} - y_i\right)^2 \to min$$

Градиент - Вектор указывающий направление наибольшего возрастания функции, компоненты которого равны частным производным по всем её аргументам.

$$\nabla\varphi = \left(\frac{\partial\varphi}{\partial x_1}, \frac{\partial\varphi}{\partial x_2}, \ldots, \frac{\partial\varphi}{\partial x_n},\right)$$

Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron

Градиентный спуск

$$MSE = \frac{1}{l}\sum_{i=1}^{l}\left(\hat{y}^{(i)} - y_i\right)^2 \rightarrow \frac{1}{l}\sum_{i=1}^{l}\left(\theta^T \cdot x^{(i)} - y_i\right)^2 \rightarrow min$$

Градиент - Вектор указывающий направление наибольшего возрастания функции, компоненты которого равны частным производным по всем её аргументам.

$$\nabla\varphi = \left(\frac{\partial\varphi}{\partial x_1}, \frac{\partial\varphi}{\partial x_2}, \ldots, \frac{\partial\varphi}{\partial x_n},\right)$$

$$\frac{\partial MSE(\theta)}{\partial\theta_j} = \frac{2}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y^{(i)}) \cdot x_j^{(i)}$$

$$\nabla MSE(\theta) = \begin{pmatrix} \frac{\partial MSE(\theta)}{\partial\theta_0} \\ \frac{\partial MSE(\theta)}{\partial\theta_1} \\ \cdots \\ \frac{\partial MSE(\theta)}{\partial\theta_n} \end{pmatrix} = \frac{2}{l}X^T \cdot \left(X \cdot \theta - y\right)$$

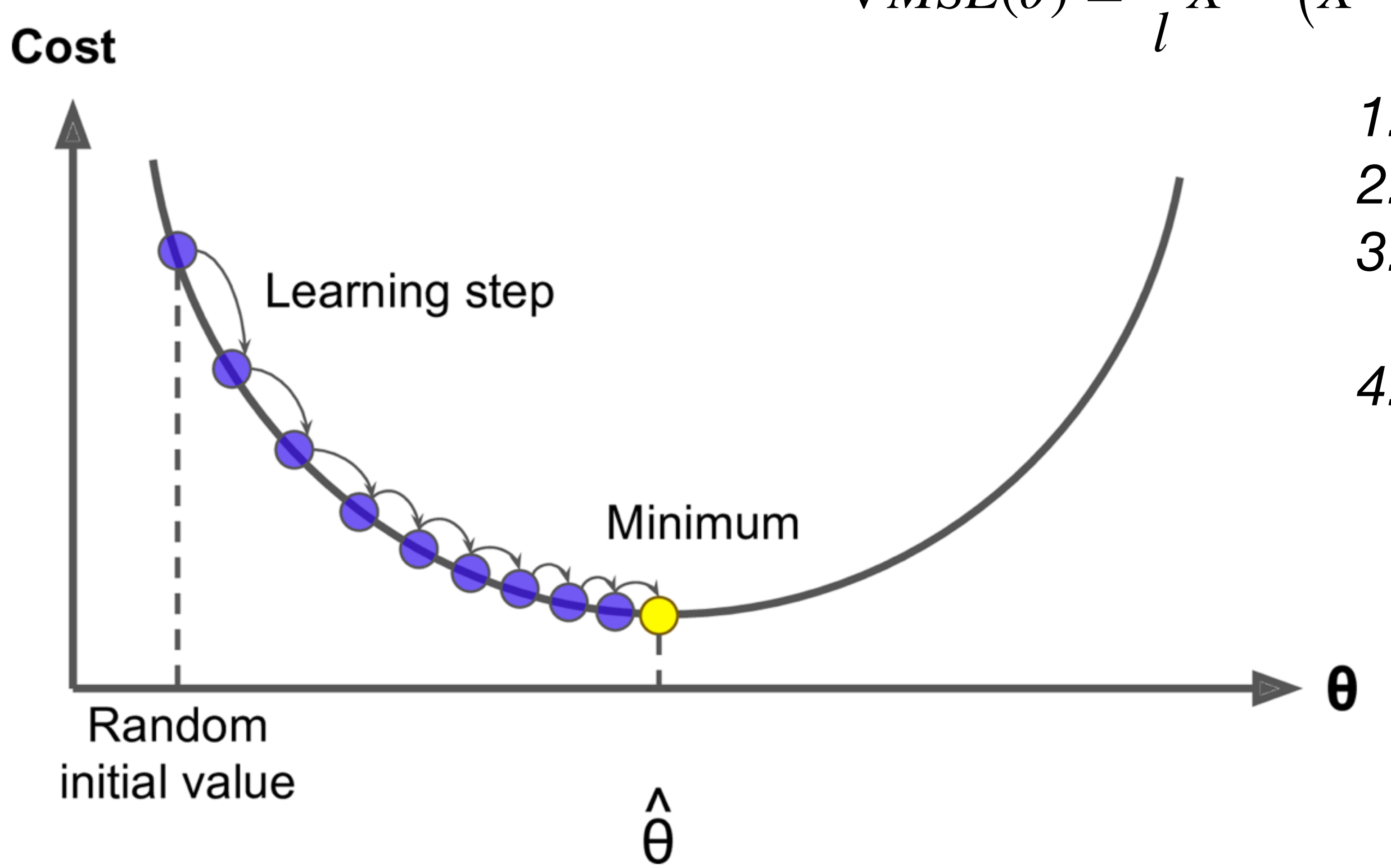
Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron

Градиентный спуск

$$MSE = \frac{1}{l}\sum_{i=1}^{l}\left(\hat{y}^{(i)} - y_i\right)^2 \rightarrow \frac{1}{l}\sum_{i=1}^{l}\left(\theta^T \cdot x^{(i)} - y_i\right)^2 \rightarrow min$$

Градиент - Вектор указывающий направление наибольшего возрастания функции, компоненты которого равны частным производным по всем её аргументам.

$$\nabla MSE(\theta) = \frac{2}{l} X^T \cdot \left(X \cdot \theta - y\right)$$

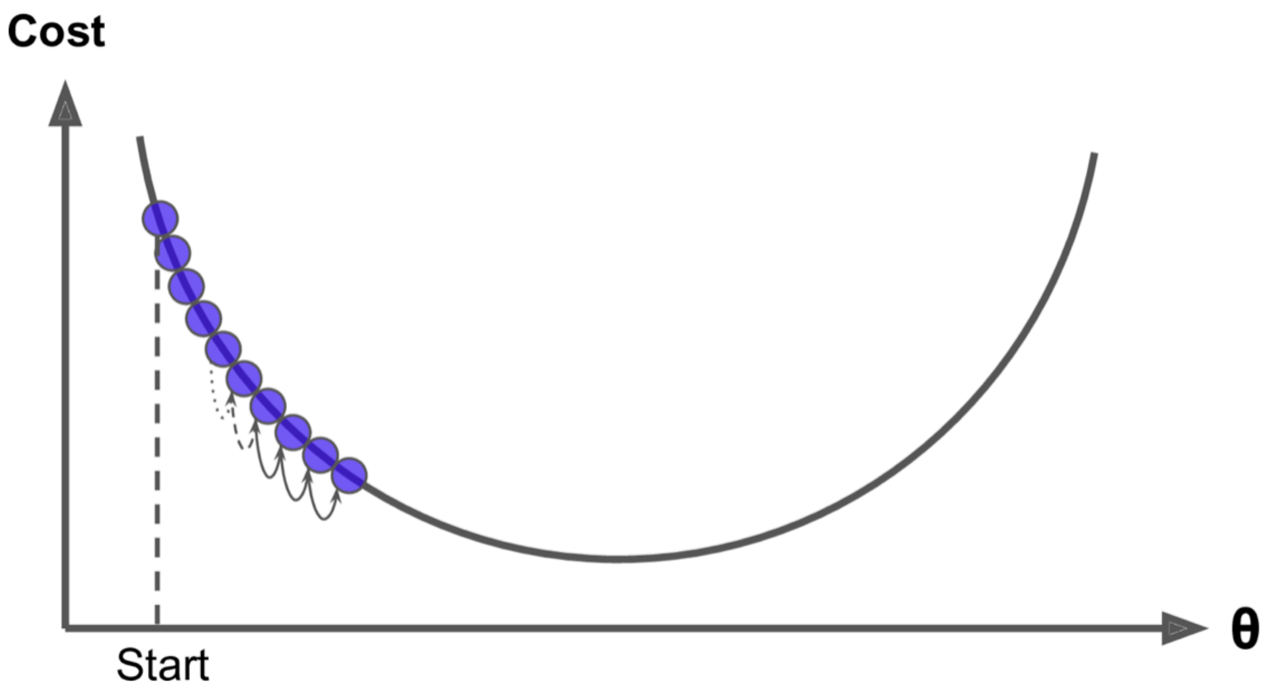


1. *Случайно задаем веса*
2. *Считаем градиент в точке*
3. *Изменяем веса путем вычитания градиента*
4. *Повторяем п.2*

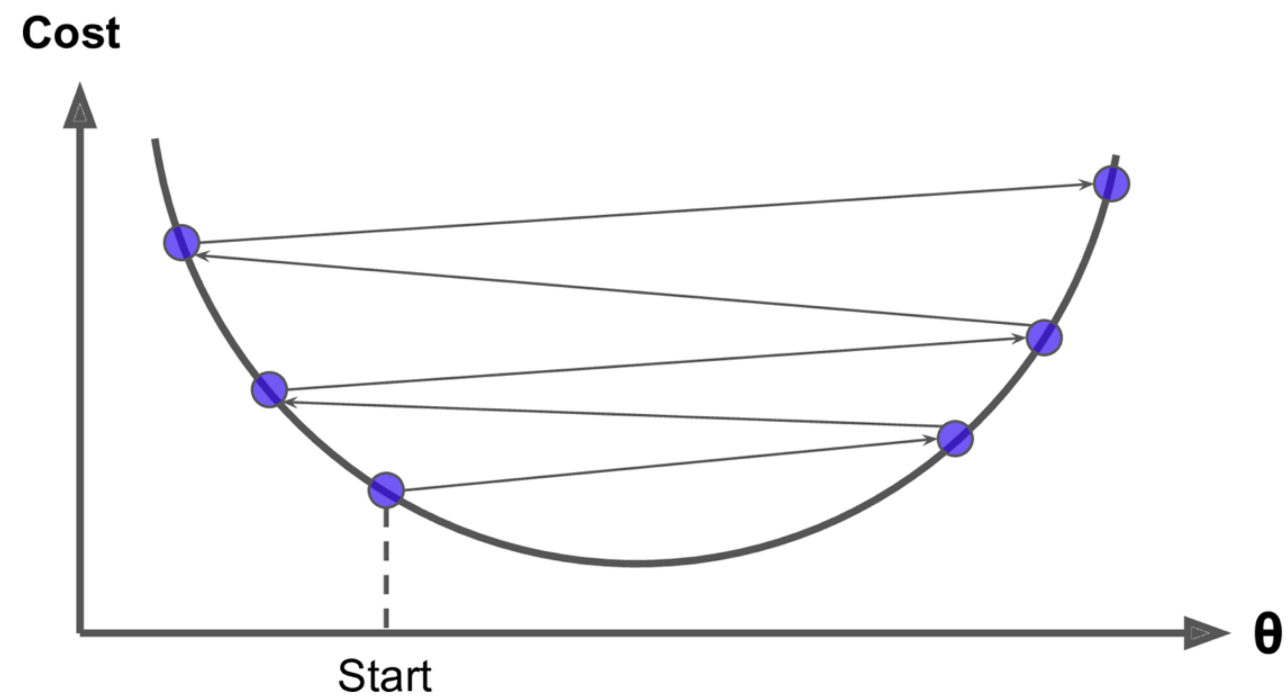
** Мы можем контролировать скорость обучения (learning rate) умножая градиент на шаг обучения*

Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron

Выбор шага обучения градиентного спуска



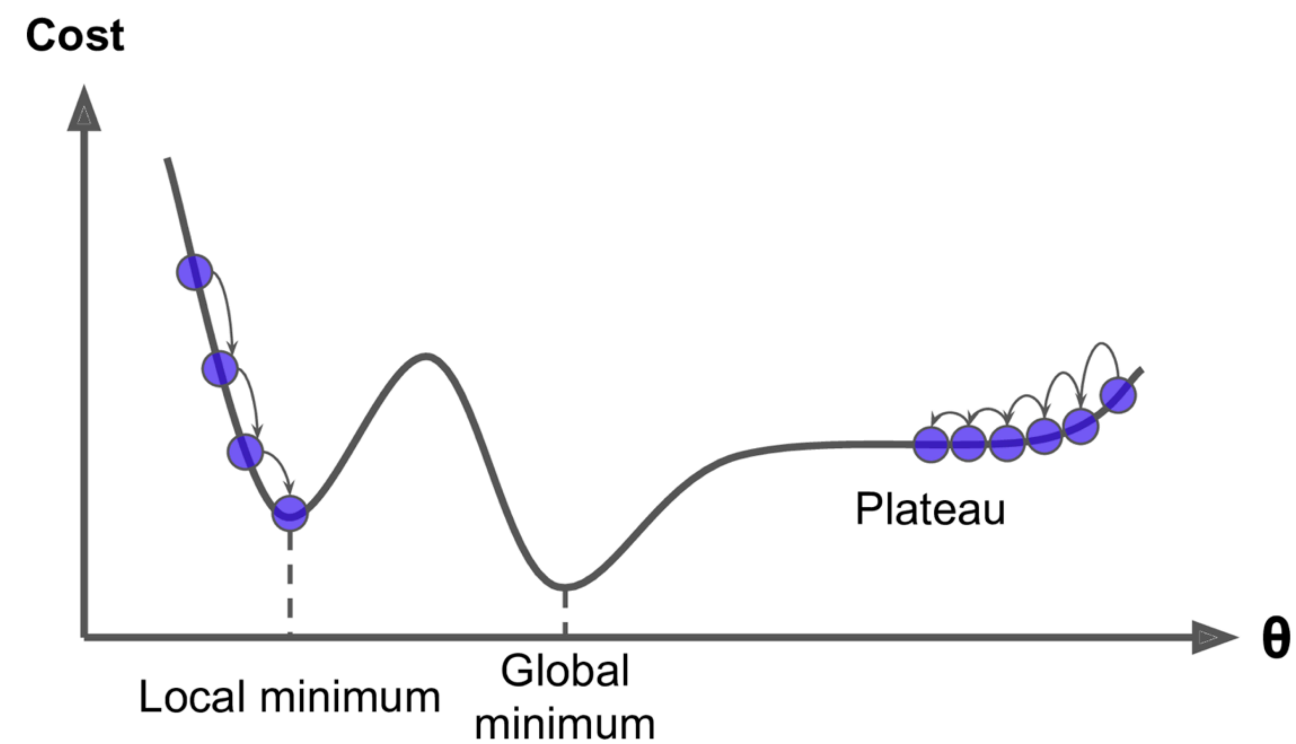
Слишком маленький шаг обучения
рискуем не дойти до минимума



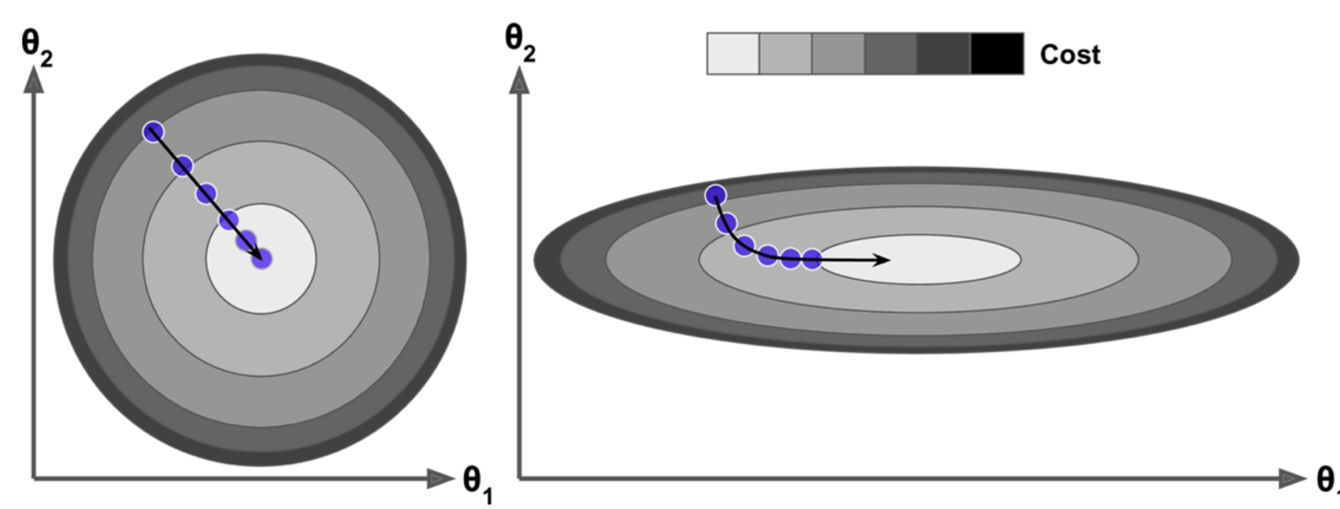
Слишком большой шаг обучения
рискуем проскочить минимум

Другие проблемы градиентного спуска

Не все функции потерь ~~одинаково полезны~~ имеют форму чаши (параболоид)



Важно масштабировать данные



Реализация в python

Стохастический градиентный спуск

$$MSE = \frac{1}{l}\sum_{i=1}^{l}\left(\hat{y}^{(i)} - y_i\right)^2 \rightarrow \frac{1}{l}\sum_{i=1}^{l}\left(\theta^T \cdot x^{(i)} - y_i\right)^2 \rightarrow min$$

$$\frac{\partial MSE(\theta)}{\partial \theta_j} = \frac{2}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y^{(i)}) \cdot x_j^{(i)}$$

Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron

Стохастический градиентный спуск

$$MSE = \frac{1}{l}\sum_{i=1}^{l}\left(\hat{y}^{(i)} - y_i\right)^2 \rightarrow \frac{1}{l}\sum_{i=1}^{l}\left(\theta^T \cdot x^{(i)} - y_i\right)^2 \rightarrow min$$

Пакетный градиентный спуск

$$\frac{\partial MSE(\theta)}{\partial \theta_j} = \frac{2}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y^{(i)}) \cdot x_j^{(i)}$$

$$\frac{\partial MSE(\theta)}{\partial \theta_j} = (\theta^T \cdot x^{(i)} - y^{(i)}) \cdot x^{(i)}$$

$$\nabla MSE(\theta) = x_i^T \cdot \left(x_i \cdot \theta - y\right)$$

Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron

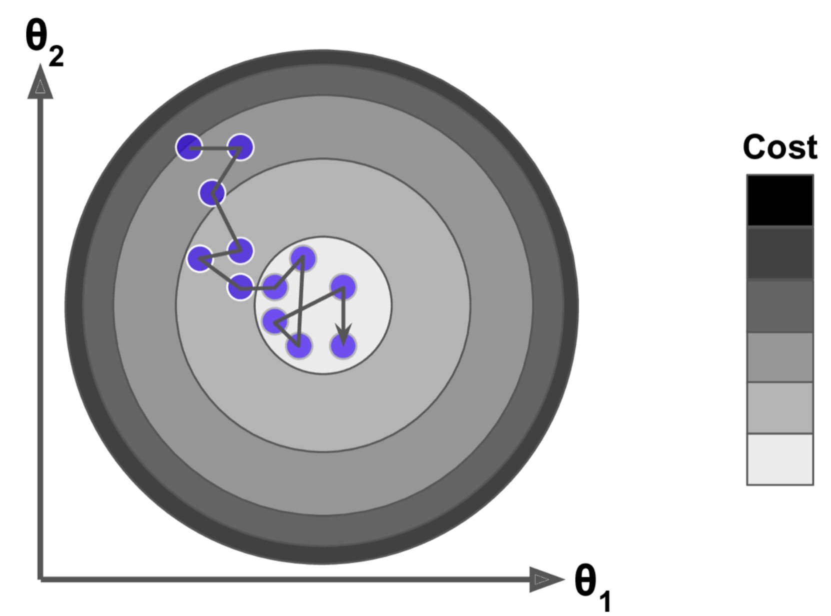
Стохастический градиентный спуск

$$MSE = \frac{1}{l}\sum_{i=1}^{l}\left(\hat{y}^{(i)} - y_i\right)^2 \rightarrow \frac{1}{l}\sum_{i=1}^{l}\left(\theta^T \cdot x^{(i)} - y_i\right)^2 \rightarrow min$$

$$\frac{\partial MSE(\theta)}{\partial \theta_j} = \frac{2}{l}\sum_{i=1}^{l}(\theta^T \cdot x^{(i)} - y^{(i)}) \cdot x_j^{(i)}$$

$$\frac{\partial MSE(\theta)}{\partial \theta_j} = (\theta^T \cdot x^{(i)} - y^{(i)}) \cdot x^{(i)}$$

$$\nabla MSE(\theta) = x_i^T \cdot \left(x_i \cdot \theta - y\right)$$

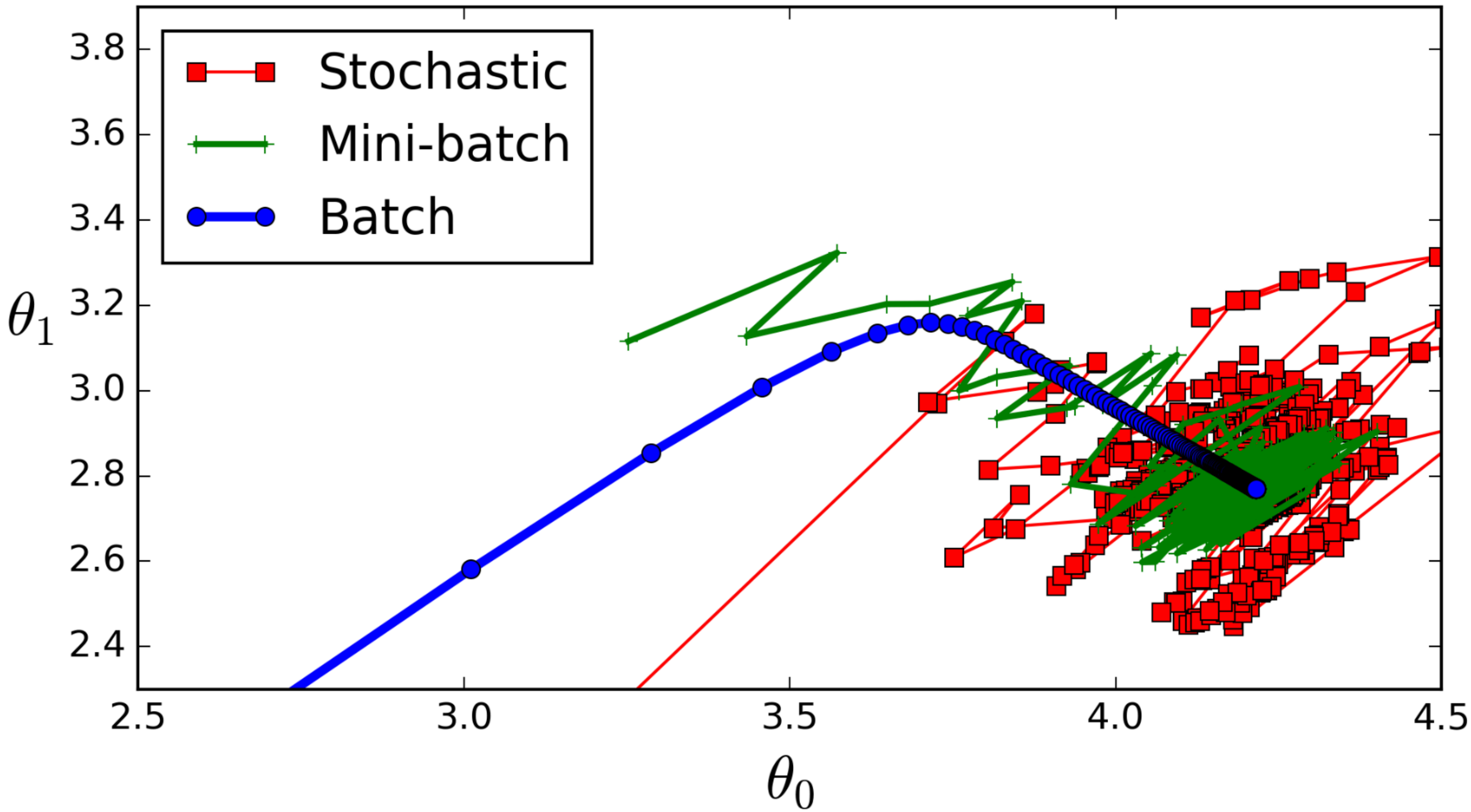


Случайно выбираем объект, и двигаемся к минимуму.
Каждый объект выборки может прогоняться несколько раз или быть не выбран вообще

Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron

Выбор метода градиентного спуска

Можно скрестить два похода: Стохастический и пакетный и выбирать случайные подборки например из 100 объектов, тогда получится: метод называемый Mini-batch



Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron

Регуляризация линейных моделей

1. Одно из условий Гауса-Маркова: $rang(X) = m$

Если оно не выполняется, то решение МНК $\hat{\theta} = \left(X^T \cdot X\right)^{-1} \cdot X^T \cdot y$ не существует

так как Матрица $X^T \cdot X$ сингулярна (вырождена)

Значит нам нужно сделать так, чтобы сделать матрицу
вырожденной (регулярной)

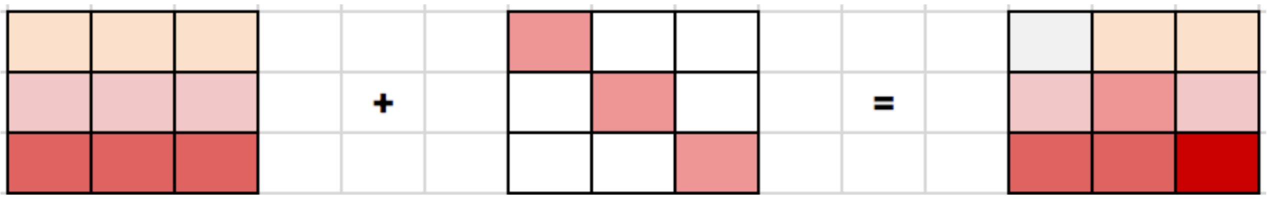
2. Мультиколлениарность

Собственные значения будут стремиться к 0, а при обращении матрицы к ∞

Регуляризация линейных моделей

L2 - регуляризация, гребневая регрессия, ridge

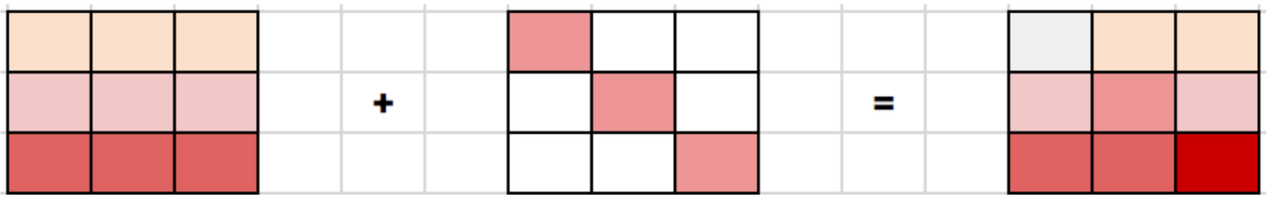
$$MSE(\theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2 \rightarrow min$$



Регуляризация линейных моделей

L2 - регуляризация, гребневая регрессия, ridge

$$MSE(\theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2 \rightarrow min$$



L1 - регуляризация, Lasso (Least absolute shrinkage and selection operator)

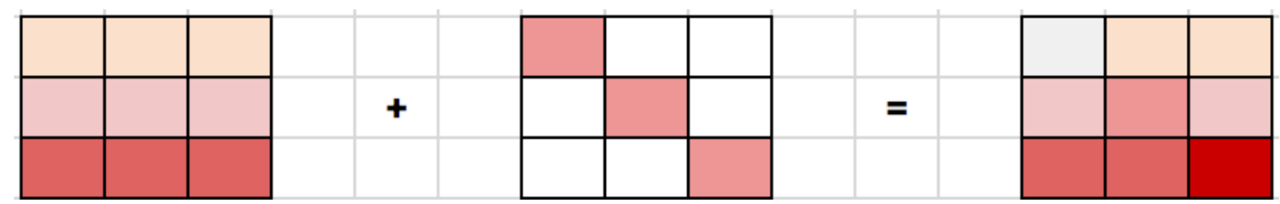
$$MSE(\theta) + \alpha \sum_{i=1}^{n} |\theta_i| \rightarrow min$$

введение ограничений на норму вектора коэффициентов модели приводит к обращению в 0 некоторых коэффициентов модели.

Регуляризация линейных моделей

L2 - регуляризация, гребневая регрессия, ridge

$$MSE(\theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2 \rightarrow min$$



L1 - регуляризация, Lasso (Least absolute shrinkage and selection operator)

$$MSE(\theta) + \alpha \sum_{i=1}^{n} |\theta_i| \rightarrow min$$

введение ограничений на норму вектора коэффициентов модели приводит к обращению в 0 некоторых коэффициентов модели.

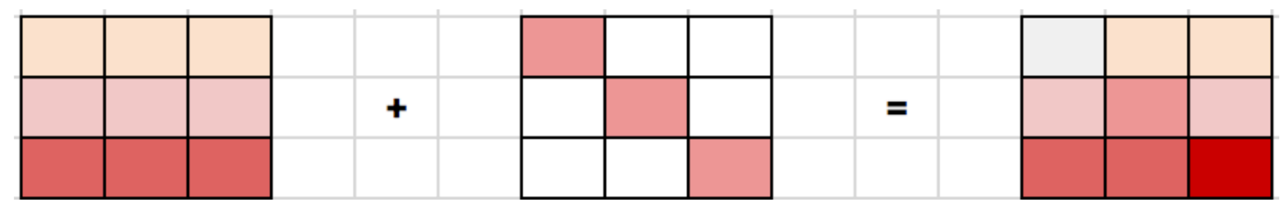
ElasticNet

$$MSE(\theta) + \alpha r \sum_{i=1}^{n} |\theta_i| + \alpha \frac{1-r}{2} \sum_{i=1}^{n} \theta_i^2 \rightarrow min$$

Регуляризация линейных моделей

L2 - регуляризация, гребневая регрессия, ridge

$$MSE(\theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} \theta_i^2 \to min$$



L1 - регуляризация, Lasso (Least absolute shrinkage and selection operator)

$$MSE(\theta) + \alpha \sum_{i=1}^{n} |\theta_i| \to min$$

введение ограничений на норму вектора коэффициентов модели приводит к обращению в 0 некоторых коэффициентов модели.

ElasticNet

$$MSE(\theta) + \alpha r \sum_{i=1}^{n} |\theta_i| + \alpha \frac{1-r}{2} \sum_{i=1}^{n} \theta_i^2 \to min$$