Методы машинного обучения

Лекция 4

Линейная регрессия

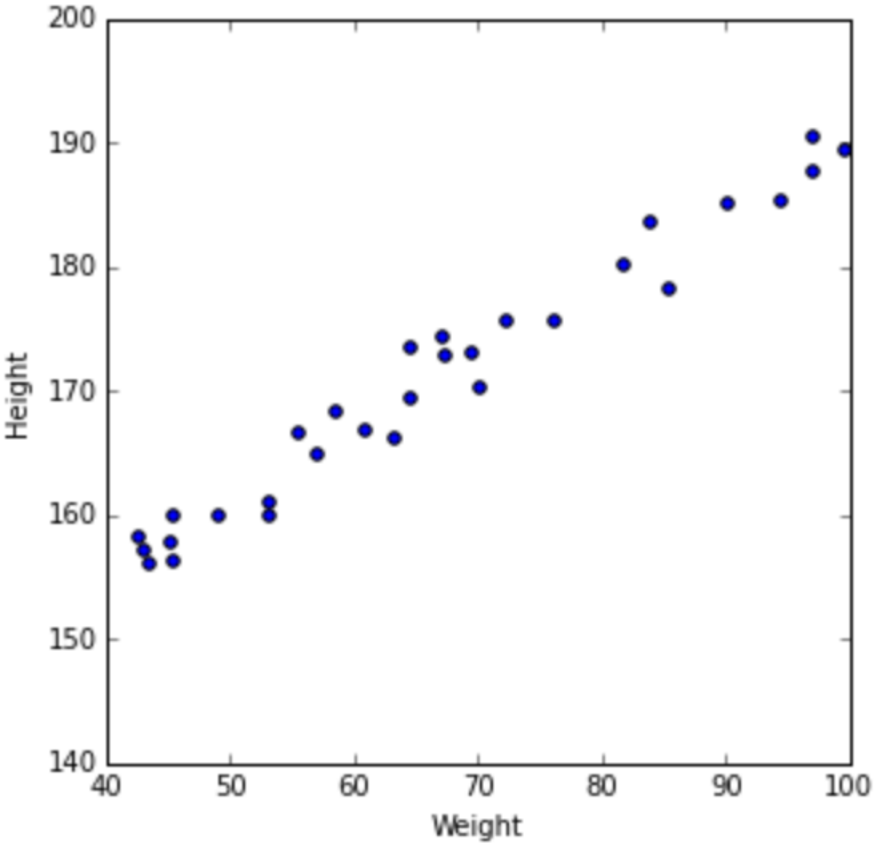
Эльвира Зиннурова

elvirazinnurova@gmail.com

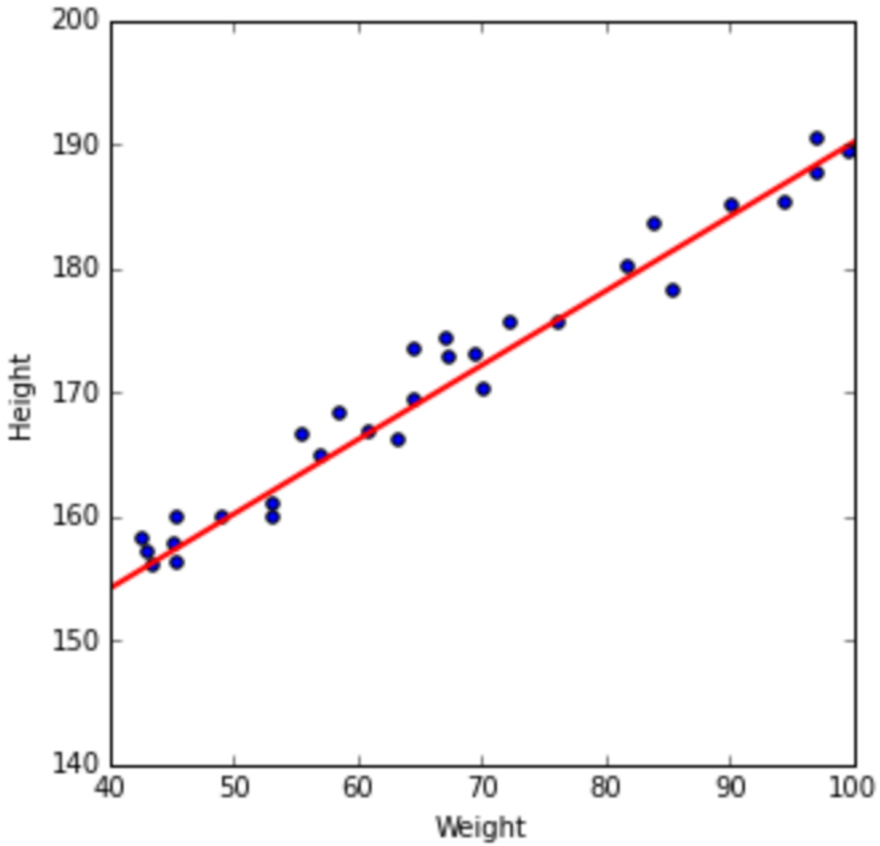
НИУ ВШЭ, 2019

Линейная регрессия

Одномерная выборка



Одномерная выборка



Парная регрессия

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

Линейная регрессия

- Взвешенная сумма признаков:

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

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

Линейная регрессия

- Взвешенная сумма признаков:

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

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

Единичный признак

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

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

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

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

Линейная регрессия

- Линейная модель: $a(x) = w_1 x^1 + \dots + w_d x^d = \langle w, x \rangle$
- Обучение:

$$\boxed{\sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2} \to \min_w$$

Функция с d аргументами

Умножение матриц и MSE

Векторы и матрицы

- Вектор размера d — тоже матрица
- Вектор-строка: $w = (w_1, \dots, w_d) \in \mathbb{R}^{1 \times d}$
- Вектор-столбец: $w = \begin{pmatrix} w_1 \\ \cdots \\ w_d \end{pmatrix} \in \mathbb{R}^{d \times 1}$

Линейная модель

- $a(x) = w_1 x^1 + \dots + w_d x^d$
- Как применить модель к целой выборке?

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix}$$

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

$$\begin{pmatrix} \sum_{i=1}^{d} w_i x_{1i} \\ \sum_{i=1}^{d} w_i x_{2i} \\ \vdots \\ \sum_{i=1}^{d} w_i x_{\ell i} \end{pmatrix}$$

Матричное умножение

- Только для матриц $A \in \mathbb{R}^{m \times k}$ и $A \in \mathbb{R}^{k \times n}$
- Результат: $AB = C \in \mathbb{R}^{m \times n}$
- Правило:

$$c_{ij} = \sum_{p=1}^{k} a_{ip} b_{pj}$$

Пример

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} & & \\ & & \\ & & \end{pmatrix}$$

Пример

$$\begin{pmatrix} \boxed{1 \quad 2} \\ 0 \quad 1 \\ 10 \quad 0 \end{pmatrix} * \begin{pmatrix} \boxed{\begin{matrix}1\\0\end{matrix}} & \begin{matrix}0\\0\end{matrix} & \begin{matrix}0\\2\end{matrix} \end{pmatrix} = \begin{pmatrix} \boxed{1} & & \\ & & \\ & & \end{pmatrix}$$

Пример

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & \boxed{0} & \\ & & \\ & & \end{pmatrix}$$

Пример

$$\begin{pmatrix} \boxed{1 \quad 2} \\ 0 \quad 1 \\ 10 \quad 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & \boxed{0} \\ 0 & 0 & \boxed{2} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \boxed{4} \\ & & \\ & & \end{pmatrix}$$

Пример

$$\begin{pmatrix} 1 & 2 \\ \boxed{0 \quad 1} \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 4 \\ \boxed{0} & & \\ & & \end{pmatrix}$$

Линейные преобразования

- Умножение на матрицу — линейная функция:
  - $A(x_1 + x_2) = Ax_1 + Ax_2$
  - $A(\alpha x) = \alpha Ax$
- Любая линейная функция описывается некоторой матрицей

Линейная модель

- Как применить модель к целой выборке?

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix}$$

$$w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

$$\begin{pmatrix} \sum_{i=1}^{d} w_i x_{1i} \\ \sum_{i=1}^{d} w_i x_{2i} \\ \vdots \\ \sum_{i=1}^{d} w_i x_{\ell i} \end{pmatrix}$$

Линейная модель

- Как применить модель к целой выборке?

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix}$$

$$w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

$$\begin{pmatrix} \sum_{i=1}^{d} w_i x_{1i} \\ \sum_{i=1}^{d} w_i x_{2i} \\ \vdots \\ \sum_{i=1}^{d} w_i x_{\ell i} \end{pmatrix} = Xw$$

Векторный вид MSE

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

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

Производная и градиент

Скорость роста

- Численность населения:

1950	1960	1970	1980	1990	2000
2,525,778,669	3,026,002,942	3,691,172,616	4,449,048,798	5,320,816,667	6,127,700,428

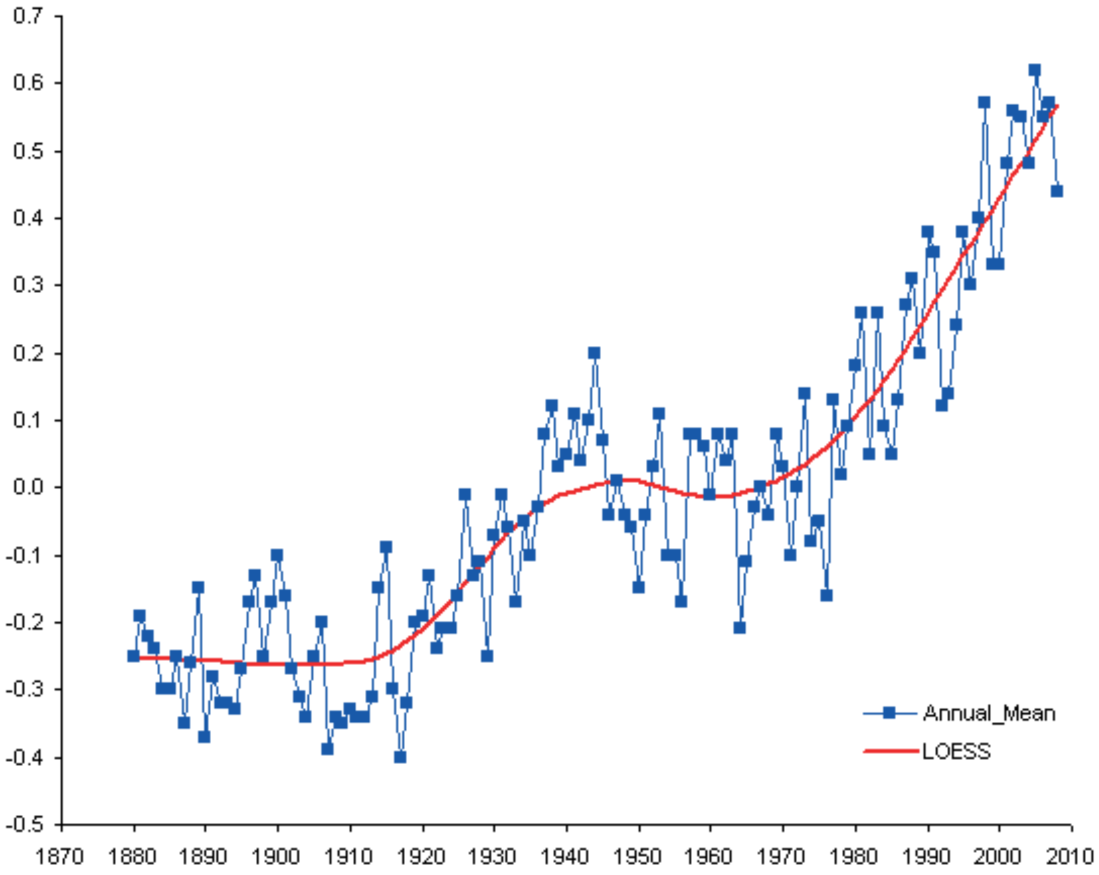
- Скорость роста между 1990 и 2000:

$$\frac{6127700428-5320816667}{10} = 80{,}688{,}376$$

- Дискретная величина

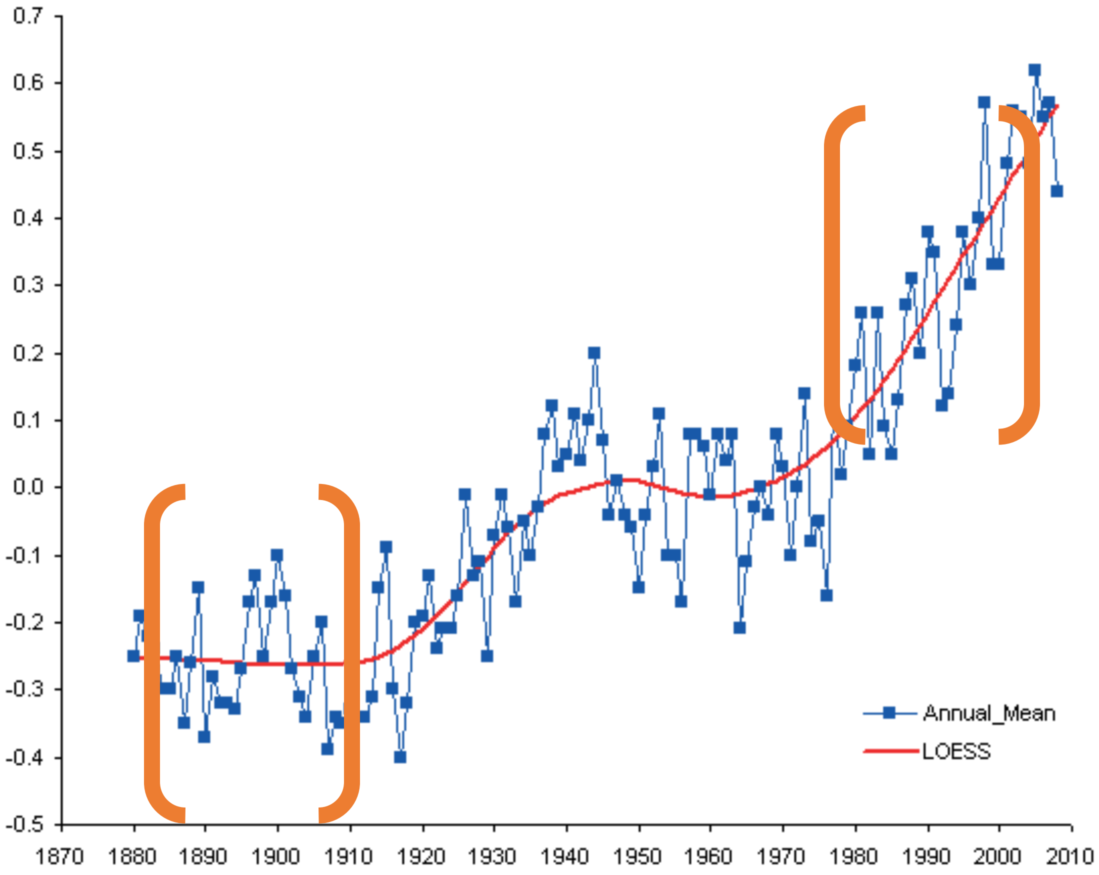
Скорость роста

- Отклонение температуры от нормы (непрерывная величина):



Скорость роста

- Отклонение температуры от нормы:



Высокая скорость

Низкая скорость

Скорость роста

- Можем измерить скорость на интервале $[x_0, x]$:

$$\frac{f(x) - f(x_0)}{x - x_0}$$

- Как измерить мгновенную скорость в конкретный момент x_0?
- Устремим x к x_0!

$$\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)$$

Скорость роста

- Можем измерить скорость на интервале $[x_0, x]$:

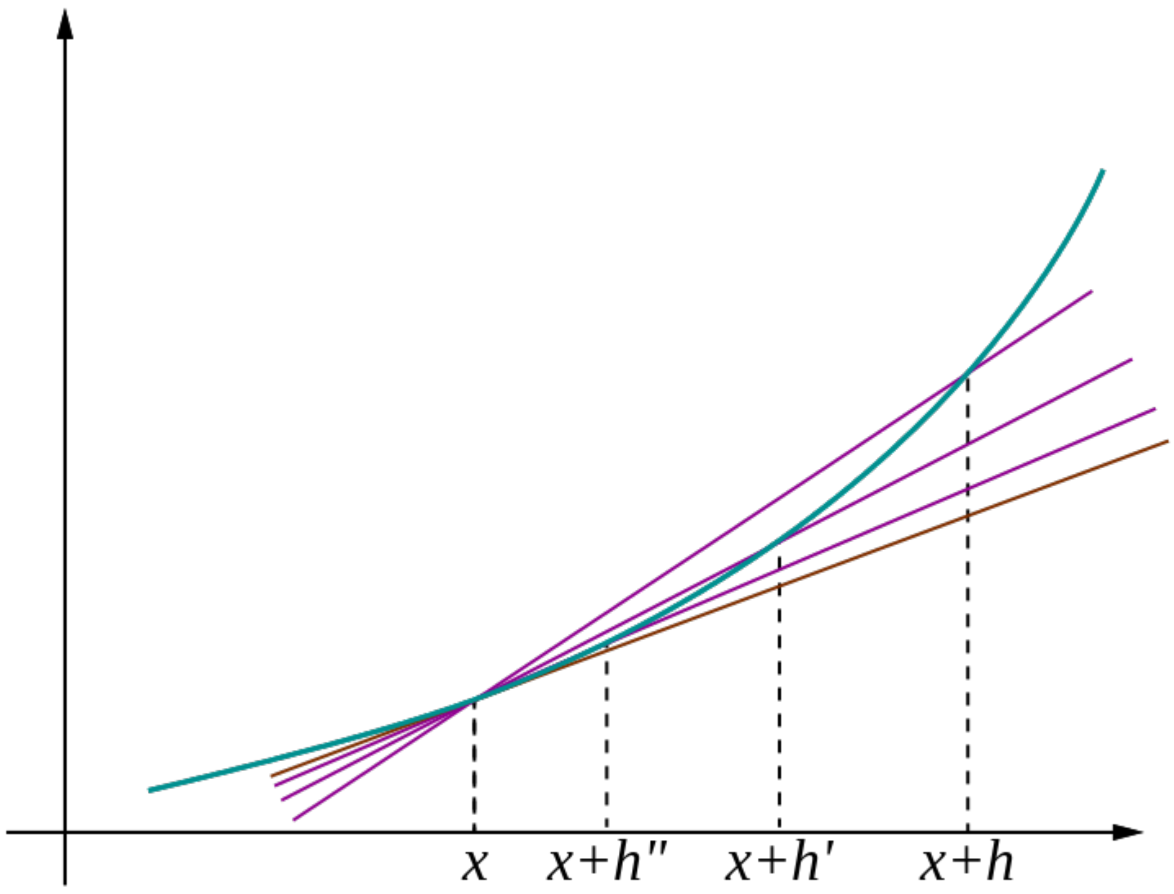
$$\frac{f(x) - f(x_0)}{x - x_0}$$

- Как измерить мгновенную скорость в конкретный момент x_0?
- Устремим x к x_0!

$$\boxed{\lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} = f'(x_0)}$$

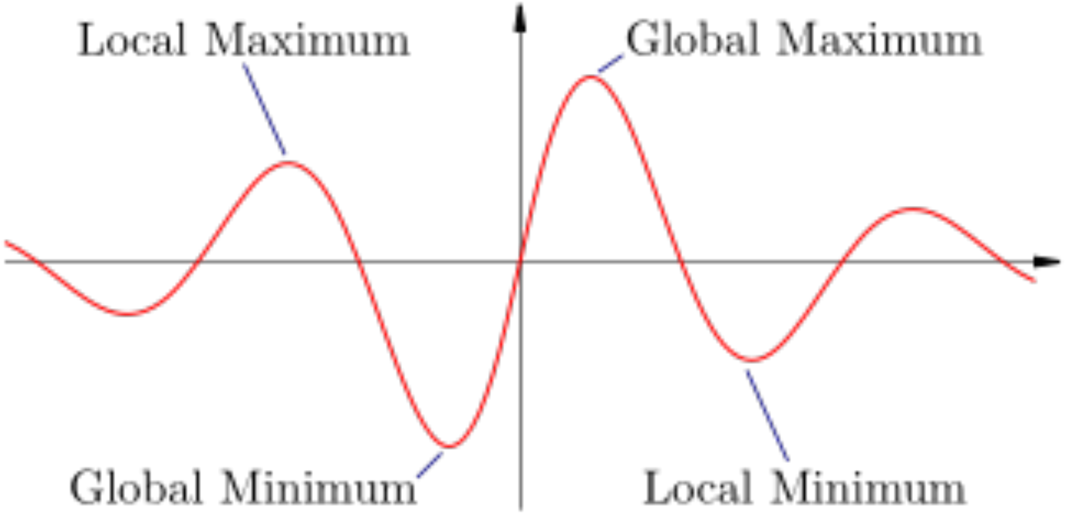
Производная

Производная



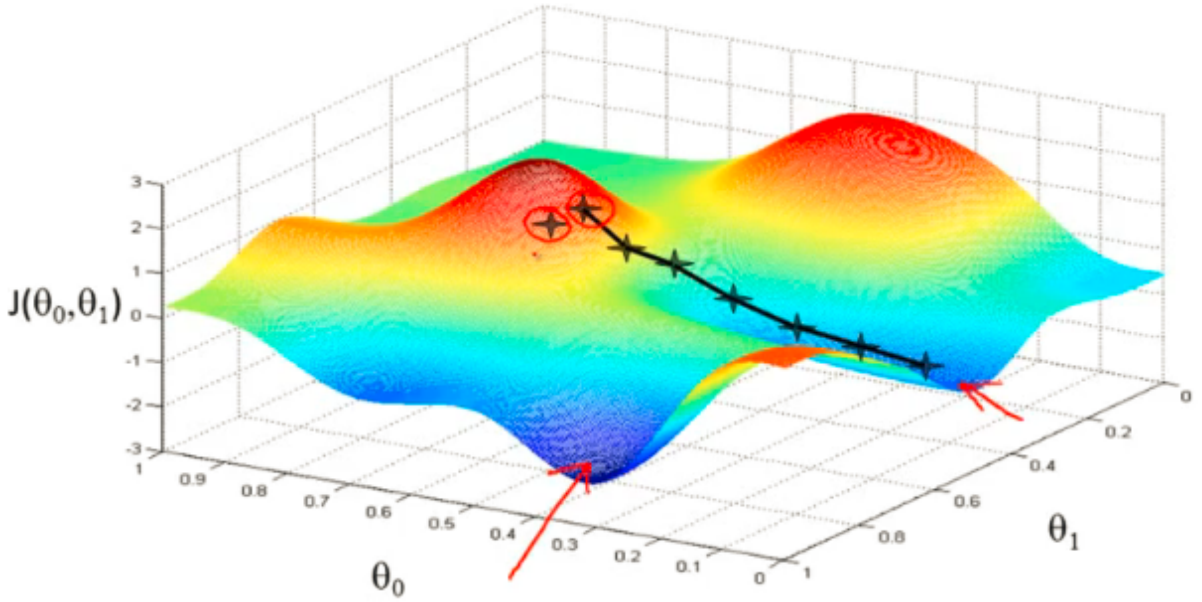
Экстремумы

- Экстремум — минимум или максимум
- Локальный минимум — меньше всех значений в некоторой окрестности
- Глобальный минимум — меньше всех значений



Экстремумы

- Локальные минимумы — одна из главных проблем в машинном обучении



Andrew Ng

Условие оптимальности

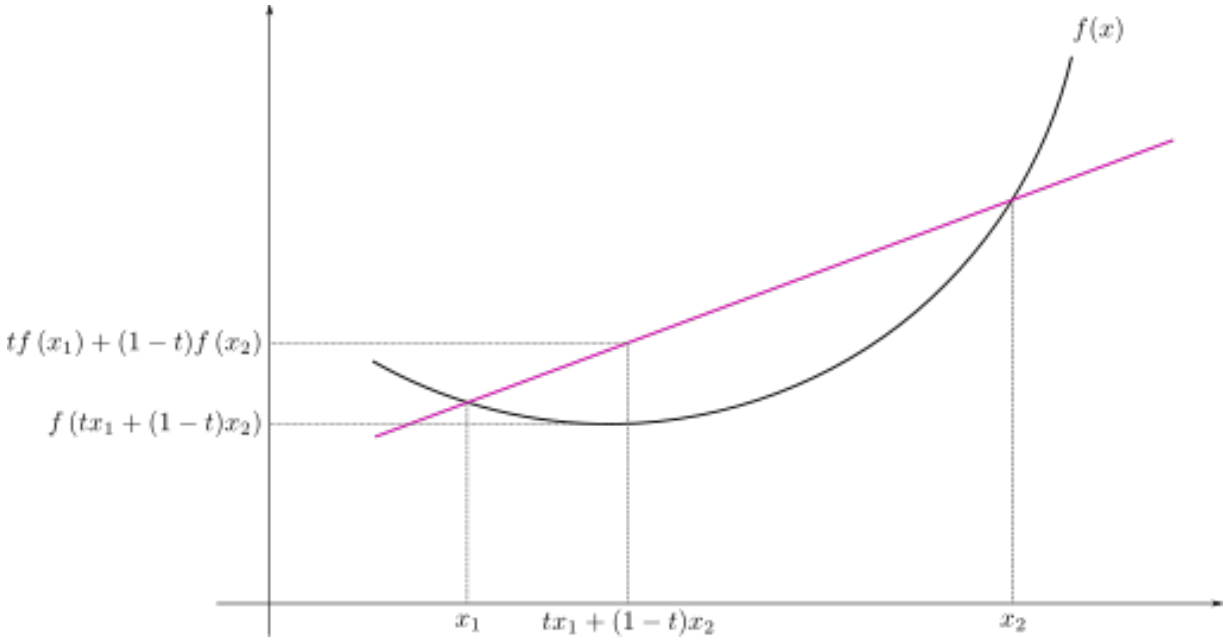
- Как понять, является ли точка x_0 экстремумом?
- Теорема Ферма: если точка x_0 — экстремум, и в ней существует производная, то $f'(x_0) = 0$

- Если функция везде имеет производную: решаем $f'(x) = 0$
- Если с производной проблемы: не повезло

- Даже если производная есть, то что делать с локальными экстремумами?

Выпуклые функции

- Функция выпуклая, если ее график лежит ниже любого отрезка, соединяющего две точки



Выпуклые функции

- Функция выпуклая, если во всех точках $f''(x) \geq 0$

- Важное свойство: любой локальный экстремум выпуклой функции является глобальным
- Решая уравнение $f'(x) = 0$, получим глобальные экстремумы
- Вывод: будем стараться выбирать выпуклые функционалы!

Пример

- Функционал качества линейной регрессии:

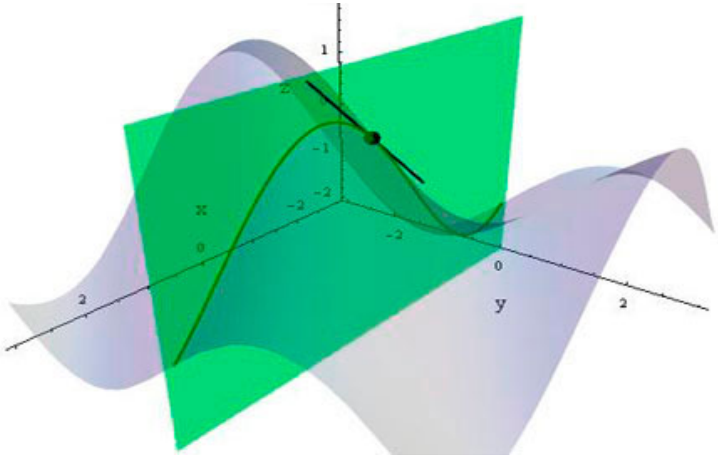
$$Q(w_1, \dots, w_d) = \sum_{i=1}^{\ell} \left(w_1 x^1 + \cdots + w_d x^d - y_i\right)^2$$

- Многомерная функция (т.е. от нескольких аргументов)
- Как искать ее минимум?

Частные производные

- С какой скоростью функция меняется вдоль переменной x_i?
- Частная производная по x_i:

$$\frac{\partial f}{\partial x_i} = \lim_{t \to 0} \frac{f(x_1, \dots, x_i + t, \dots, x_d) - f(x_1, \dots, x_i, \dots, x_d)}{t}$$



Градиент

- Градиент — вектор из частных производных:

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_d}\right)$$

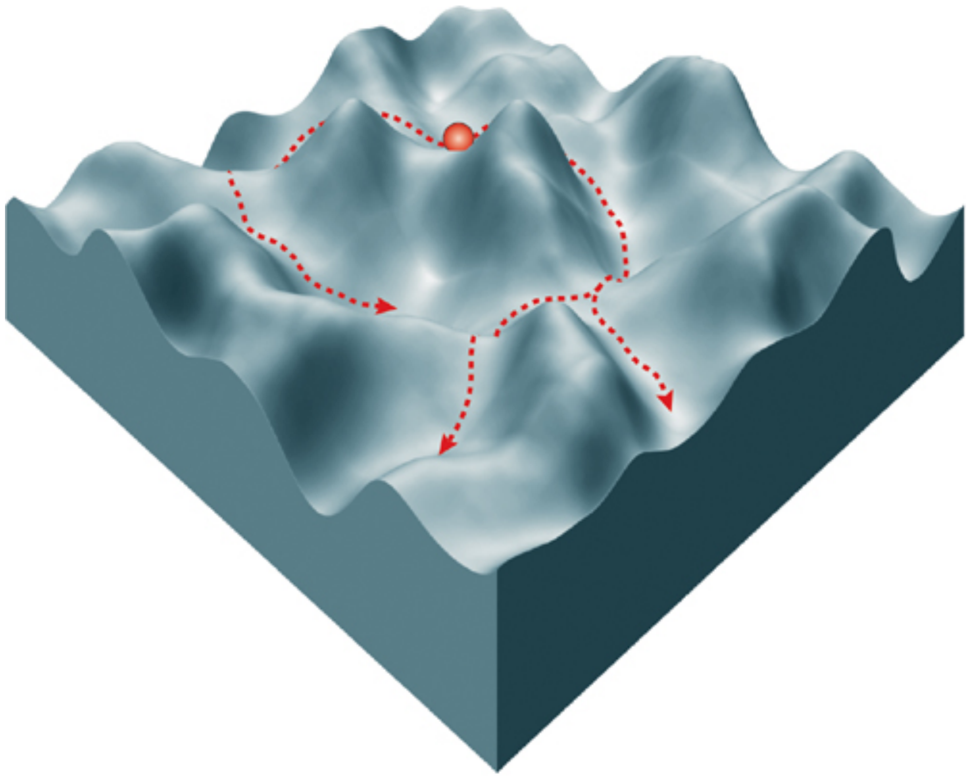
Условие оптимальности

- Как понять, является ли точка x_0 экстремумом?
- Обобщение теоремы Ферма: если точка x_0 — экстремум, и в ней существует градиент, то $\nabla f(x_0) = 0$

- Если функция везде имеет градиент: решаем $\nabla f(x) = 0$ (теперь это система уравнений!)
- Если с градиентом проблемы: не повезло

Экстремумы

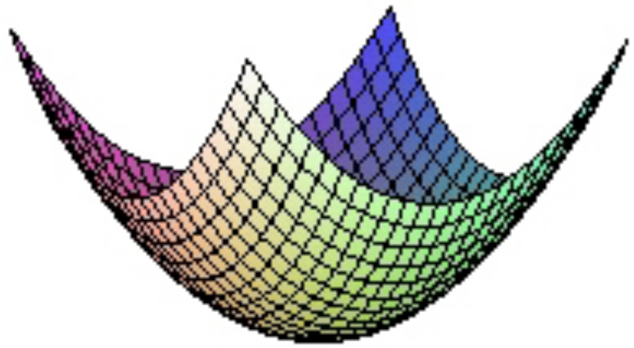
- Проблема с локальными экстремумами все еще актуальна

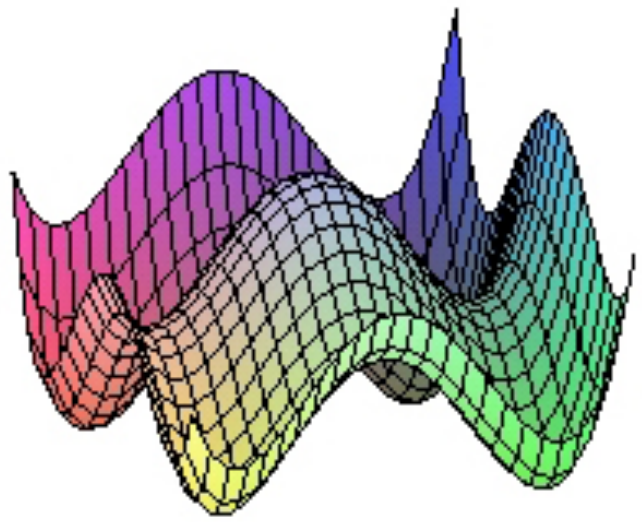


Nature Reviews | Molecular Cell Biology

Выпуклые функции

- Функция выпуклая, если ее график лежит ниже отрезка, соединяющего любые две точки

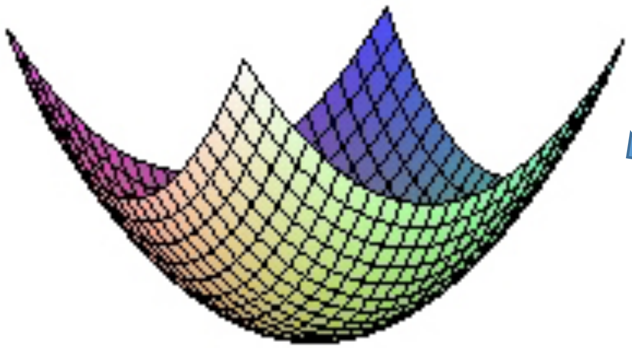


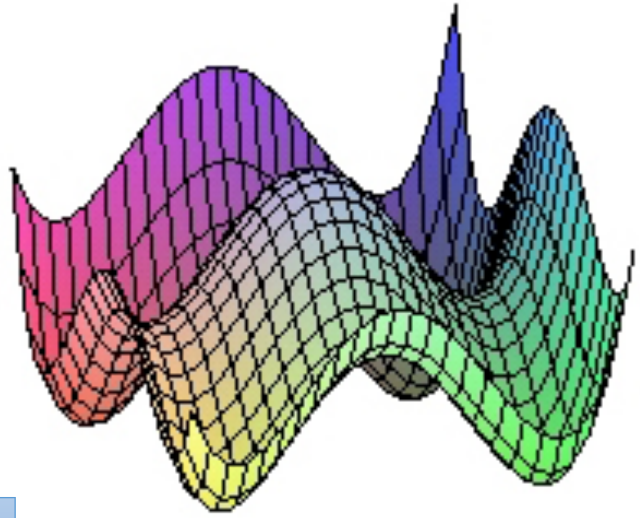


Выпуклые функции

- Функция выпуклая, если ее график лежит ниже отрезка, соединяющего любые две точки

Выпуклая функция

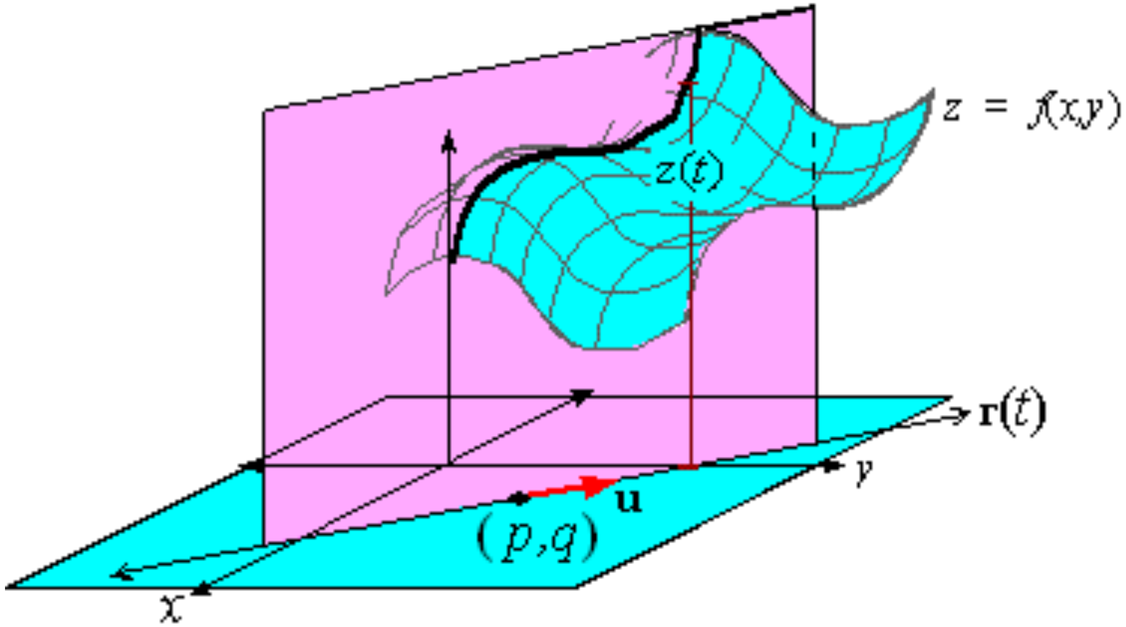




Невыпуклая функция

Производная по направлению

- Градиент — про скорость роста по конкретному аргументу
- С какой скоростью растет функция в конкретном направлении?



Производная по направлению

- Направление: v, причем $\|v\| = 1$
- Производная:

$$f'_v(x_0) = \lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t}$$

Связь с градиентом

- Зафиксируем точку x_0
- В каком направлении функция быстрее всего растет?

$$f'_v(x_0) \to \max_v$$

Угол между градиентом и направлением

- Связь производной по направлению и градиента:

$$f'_v(x_0) = \langle \nabla f(x_0), v \rangle = \|\nabla f(x_0)\| * \|v\| * \cos \boxed{\varphi}$$

Важное свойство градиента

- Производная по направлению максимальна, если направление совпадает с градиентом!
- **Градиент — направление наискорейшего роста функции**
- Антиградиент — направление наискорейшего убывания

Обучение линейной регрессии

Задача оптимизации

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

- Градиент существует в любой точке
- Выпуклая функция
- Единственный минимум (не всегда)

Градиент

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

Производные:

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

Векторный вид

- Векторная запись MSE:

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

- Условие минимума:

$$\nabla Q = \frac{2}{\ell} X^T (Xw - y) = 0$$

- Что, если попробуем решить эту систему уравнений?

Обратная матрица

- A^{-1}— обратная к A
- $AA^{-1} = A^{-1}A = I$
- I — единичная матрица
- Только для квадратных матриц

- Существует тогда и только тогда, когда $\det A \neq 0$

Обучение линейной регрессии

- Условие минимума решается аналитически!

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

- Но обращение матрицы — очень сложная операция
- А также некоторые другие проблемы
- Градиентный спуск гораздо быстрее — но об этом позже

Резюме

- Линейная регрессия — одна из самых простых моделей в машинном обучении
- Функционал качества: среднеквадратичная ошибка
- Обучение: аналитическая формула или градиентный спуск