DEEP LEARNING

Введение

Святослав Елизаров, Борис Коваленко

8 сентября 2018

Высшая школа экономики

ОСНОВНЫЕ ПОНЯТИЯ

ВЕРОЯТНОСТЬ

Вероятностным пространством называется тройка вида:

$$(\Omega, \mathcal{A}, \mathbb{P})$$

Где:

- Ω – множество элементарных событий
- $\mathcal{A}$ – сигма-алгебра на Ω
- $\mathbb{P}$ – вероятностная мера, такая что $\mathbb{P}(\Omega) = 1$

СЛУЧАЙНАЯ ВЕЛИЧИНА

Случайная величина – математический термин, использующийся для представления объектов или их свойств.

Случайной величиной называется некоторая функция $X : \Omega \to \mathbb{R}$, измеримая относительно $\mathcal{A}$. Случайные величины бывают непрерывными и дискретными.

СЛУЧАЙНАЯ ВЕЛИЧИНА

Случайная величина – математический термин, использующийся для представления объектов или их свойств.

Случайной величиной называется некоторая функция $X : \Omega \to \mathbb{R}$, измеримая относительно $\mathcal{A}$. Случайные величины бывают непрерывными и дискретными.

Примеры:

- Случайная величина X отвечает времени ожидания автобуса.
- X отвечает всем возможным изображениям разрешения 640x480

РАСПРЕДЕЛЕНИЕ

Распределение – это некоторый закон, ставящий в соответствие реализациям (значениям) случайной величины вероятности их появления.

Обычно обозначается $p_X(\cdot)$, когда речь идёт о случайной величине X или просто $p(\cdot)$, когда это понятно из контекста и называется **плотностью вероятности**.

РАСПРЕДЕЛЕНИЕ

Так же существует кумулятивная функция распределения, которая показывает вероятность всех событий меньших или равных данному.

$$P(x) = \mathbb{P}(X < x)$$

Связь с плотностью распределения:

$$P(x) = \int_{-\infty}^{x} p(t)dt$$

ПРИМЕРЫ РАСПРЕДЕЛЕНИЙ

- Биномиальное распределение

$$p(x) = \binom{n}{x} \mu^x (1-\mu)^{n-x}$$

  Обозначается $B(n, \mu)$

- Гауссовское распределение (нормальное)

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

  Обозначается $N(\mu, \sigma^2)$

- Распределение Лапласа

$$p(x) = \frac{1}{2b} \exp\left(-\frac{|(x-\mu)|}{b}\right)$$

  Обозначается $L(\mu, b)$

УСЛОВНОЕ РАСПРЕДЕЛЕНИЕ

Условным называется распределение случайной величины при условии, что другая случайная величина приняла определённое значение.

$p(x, y)$ – совместное распределение случайных величин X и Y. Тогда условным распределением будет

$$p(x|y) = \frac{p(x, y)}{p(y)}$$

ЧАСТОТНЫЙ ПОДХОД

Важным является вопрос о том, как интерпретировать вероятности.

Частотным называется подход, в котором вероятность определяется частотой некоторого события, которе повторяется многократно в серии экспериментов.

МАКСИМИЗАЦИЯ ПРАВДОПОДОБИЯ

Представим, что мы имеем выборку $x_1, x_2, \ldots, x_n$, состоящую из реализаций случайной величины X.

Что делать, чтобы восстановить распределение X?

МАКСИМИЗАЦИЯ ПРАВДОПОДОБИЯ

Предположим, что распределение X принадлежит некоторому параметрическому семейству распределений(например, нормальному) с параметрами $\theta \in \Theta$. Можно записать это как условное распределение:

$$p(x|\theta)$$

Данное условное распределение называется **правдоподобием**, его иногда обозначают символом $\mathcal{L}$.

МАКСИМИЗАЦИЯ ПРАВДОПОДОБИЯ

Выпишем правдоподобие для имеющейся выборки:

$$p(x_1, x_2, \ldots, x_n|\theta)$$

И если считать объекты незавимыми, то

$$p(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

МАКСИМИЗАЦИЯ ПРАВДОПОДОБИЯ

Выпишем правдоподобие для имеющейся выборки:

$$p(x_1, x_2, \ldots, x_n|\theta)$$

И если считать объекты незавимыми, то

$$p(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

Или что эквивалентно:

$$\ln p(x_1, x_2, \ldots, x_n|\theta) = \sum_{i=1}^{n} \ln p(x_i|\theta)$$

МАКСИМИЗАЦИЯ ПРАВДОПОДОБИЯ

Необходимо найти такие параметры θ при которых правдоподобие будет максимальным.

$$\theta_{mle} = \arg\max_{\theta \in \Theta} p(x_1, x_2, \ldots, x_n | \theta)$$

Где Θ множество параметров для данного семейства.
Максимизация правдоподобия является основным механизмом статистического вывода в частотном подходе.

ПРИМЕР МАКСИМИЗАЦИИ ПРАВДОПОДОБИЯ

Представим, что мы подбрасываем монету.

Будем считать, что данные имеют распределение Бернулли.

$$Bern(x|\mu) = \mu^x(1-\mu)^{1-x}$$

Параметр μ – вероятность выпадения орла, $x \in \{0, 1\}$

ПРИМЕР МАКСИМИЗАЦИИ ПРАВДОПОДОБИЯ

Предположим, что данные независимы.

Выпишем логарифмическое правдоподобие:

$$\ln p(x_1, x_2, \ldots, x_n|\mu) = \sum_{i=1}^{n}(x_i \ln \mu + (1 - x_i) \ln(1 - \mu))$$

ПРИМЕР МАКСИМИЗАЦИИ ПРАВДОПОДОБИЯ

Пусть выборка данных состоит из трёх наблюдений и все они равны 1 (выпал орёл).

Из формулы логарифмического правдоподобия видно, что параметр μ оценивается следующим образом:

$$\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

ПРИМЕР МАКСИМИЗАЦИИ ПРАВДОПОДОБИЯ

Пусть выборка данных состоит из трёх наблюдений и все они равны 1 (выпал орёл).

Из формулы логарифмического правдоподобия видно, что параметр μ оценивается следующим образом:

$$\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Всё ли в порядке с данным решением?

ПРИМЕР МАКСИМИЗАЦИИ ПРАВДОПОДОБИЯ

Пусть выборка данных состоит из трёх наблюдений и все они равны 1 (выпал орёл).

Из формулы логарифмического правдоподобия видно, что параметр μ оценивается следующим образом:

$$\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Всё ли в порядке с данным решением?

$\mu_{ML} = 1$ и наша монета обязана всегда приземляться орлом вверх!

БАЙЕСОВСКИЙ ПОДХОД

Байесовский подход трактует вероятность как меру неопределенности.

Так же многие события являются уникальными и не могут быть многократно повторены для того, чтобы измерить частоту того или иного исхода. Например, исчезновение лесов в дельте Амазонки. Байесовский подход позволят использовать **априорное знание** для определения вероятности таких событий.

БАЙЕСОВСКИЙ ПОДХОД

Основная идея байесовского подхода заключается в том, что параметры семейства тоже являются случайной величиной, именно она определяет априорное знание. Для вывода используется формула Байеса:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

Где: $p(x|\theta)$ – правдоподобие

$p(\theta)$ – априорное распределение параметров или prior

$p(x)$ – вероятность данных или evidence

$p(\theta|x)$ – апостериорное распределение.

MAP

Важным преимуществом Байесовского подхода является то, что можно использовать $p(\theta|x)$ в в качестве априорного распределения в дальнейшем и постепенно дообучать модель.

Главный инструмент статистического вывода максимизация апостериорной вероятности или MAP:

$$\theta_{map} = \arg\max_{\theta \in \Theta} p(\theta|x)$$

MAP

Важным преимуществом Байесовского подхода является то, что можно использовать $p(\theta|x)$ в в качестве априорного распределения в дальнейшем и постепенно дообучать модель.

Главный инструмент статистического вывода максимизация апостериорной вероятности или MAP:

$$\theta_{map} = \arg\max_{\theta \in \Theta} p(\theta|x)$$

Проблемы?

СОПРЯЖЁННОЕ РАСПРЕДЕЛЕНИЕ

Априорное распределение $p(\theta)$ называется сопряжённым к $p(x|\theta)$, если апостериорное распределение $p(\theta|x)$ имеет ту же функциональную форму, что и $p(\theta)$.

Примеры сопряженных распределений:

- Нормальное распределение является сопряженным к нормальному
- Бета-распределение является сопряженным к распределению Бренулли.

$$p(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} = const \cdot x^{\alpha-1}(1-x)^{\beta-1}$$

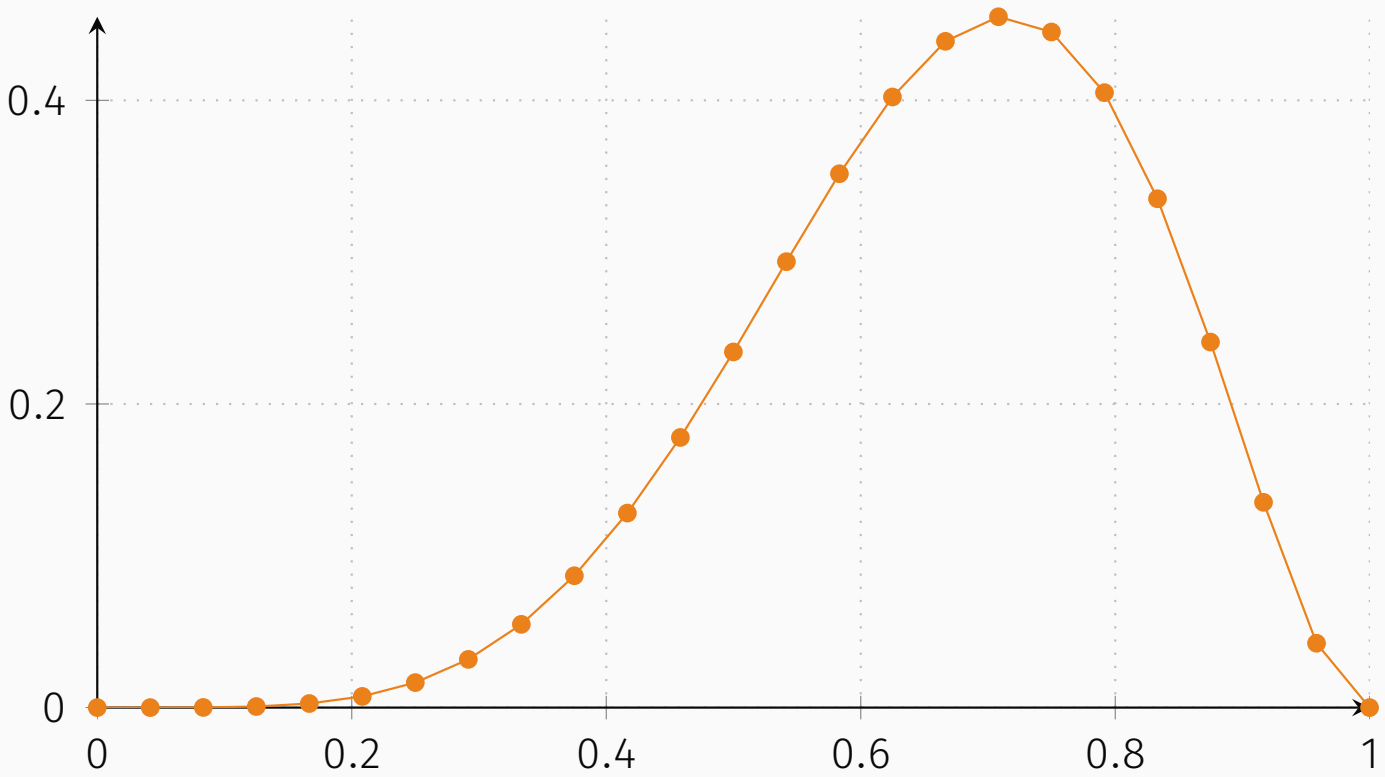
ПРИМЕР MAP

Вернёмся к примеру с монетой, но на этот раз зададим априорное распределение исходя из предположения, что монета честная.

В случае распределения Бернулии, нам необходимо выбрать симметричное бета-распределение. Возмём в качестве $\alpha = \beta = 3$

$$p(\mu|x) = const \cdot \mu^5(1-\mu)^2$$

ПРИМЕР MAP



СРАВНЕНИЕ ЧАСТОТНОГО И БАЙЕСОВСКОГО ПОДХОДА

	Частотный	Байесовский
Случайность	Объективная неопределенность	Субъективное незнание
Переменные	Случайные или детерминированные	Случайные
Инференс	Метод максимального правдоподобия	Теорема Байеса
Оценки	ML-оценки	Апостериорные или MAP-оценки
Размер выборки	$n >> 1$	$\forall n$

ПОСТАНОВКА ЗАДАЧИ

Пусть Ω – множество всех объектов. Обозначим через X некоторое подмножество этого множества, $X \subset \Omega$

Множество Y – это множество значений целевого признака.

Функция $\tilde{f} : \Omega \to Y$ ставит в соответствие каждому объекту некоторое значение $y \in Y$.

ПОСТАНОВКА ЗАДАЧИ

Дано:

- Множество X
- Значения функции $\tilde{f}$ на множестве X

Задача: Предсказать значения $\tilde{f}$ для всего множества Ω, или, другими словами, восстановить функцию f. Восстановленную функцию будем обозначать просто f.

Такая задача называется **обучением с учителем** или **supervised learning**

ПОСТАНОВКА ЗАДАЧИ

В случае если значения функции $\tilde{f}$ на множестве X неизвестны, то такая задача называется **обучением без учителя** или **unsupervised learning**

ПОСТАНОВКА ЗАДАЧИ

В данном курсе мы столкнёмся со множеством частных случаев каждой из этих задач:

С учителем:

1. Классификация
2. Регрессия
3. Сегментация изображений

Без учителя:

1. Word or sentence embeddings
2. Кластеризация
3. Style transfer

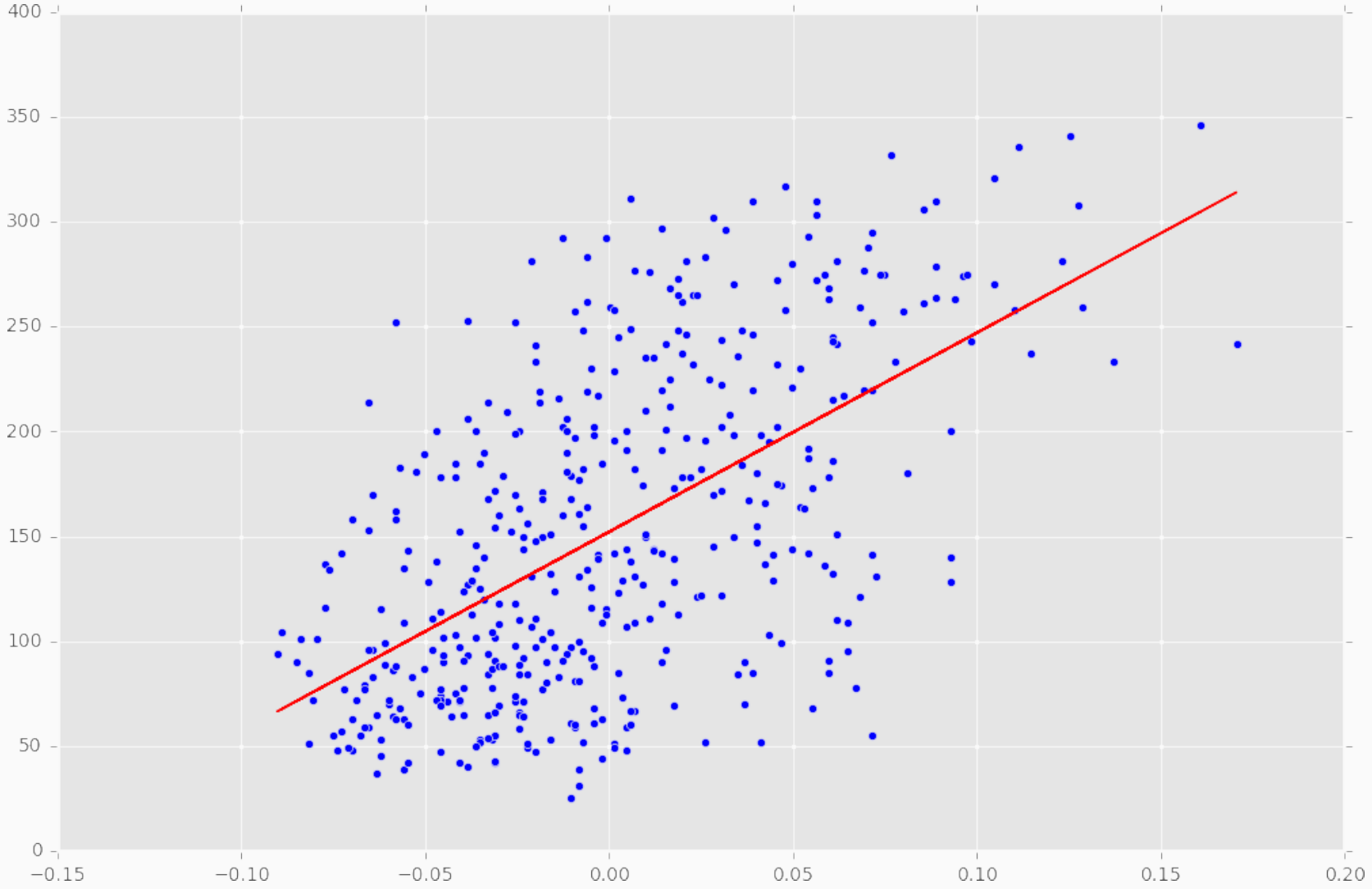
ЛИНЕЙНЫЕ МОДЕЛИ

ЛИНЕЙНАЯ РЕГРЕССИЯ

Простейшая линейная регрессия:

$$f(x) = \theta_1 x + \theta_0$$

ЛИНЕЙНАЯ РЕГРЕССИЯ



ЛИНЕЙНАЯ РЕГРЕССИЯ

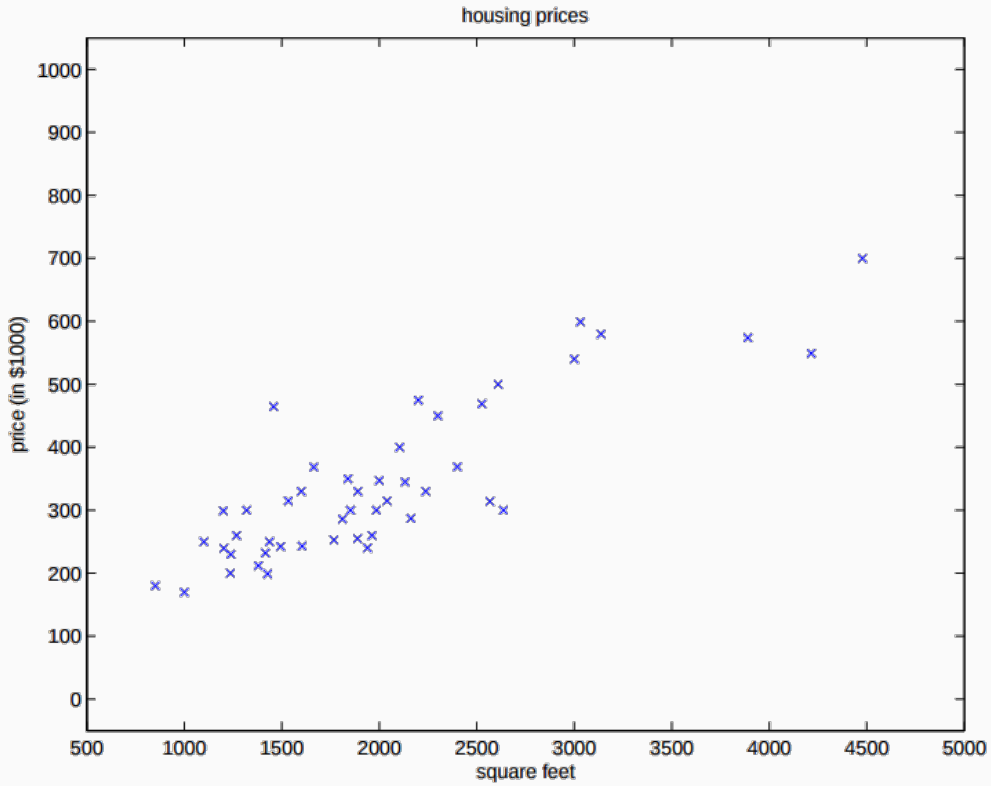
Сколько стоит квартира, если в ней x_1 комнат и ее площадь равна x_2

Количество комнат	Общая площадь, m^2	Цена, руб.
3	60	10 млн
2	100	7 млн
2	50	6 млн

...

ЛИНЕЙНАЯ РЕГРЕССИЯ

Построим график зависимости целевой переменной от одного из признаков:



Линейная зависимость?

ЛИНЕЙНАЯ РЕГРЕССИЯ

Будем моделировать зависимоть с помощью линейной функции:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 = \theta^T x$$

θ - вектор весов (параметров)

Будем измерять качество модели с помощью функционала:

$$J(\theta) = \frac{1}{2} \sum_i^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

Как подобрать θ?

ОБУЧЕНИЕ ЛИНЕЙНОЙ РЕГРЕССИИ

Функция потерь выпуклая. Выпуклая функция обладает множеством замечательных свойств, наиболее важными из которых для нас являются:

1. Функция непрерывна и дифференцируема на всём интервале за исключением не более чем счётного множества точек и дважды дифференцируема почти всюду.
2. Локальный минимум является глобальным.

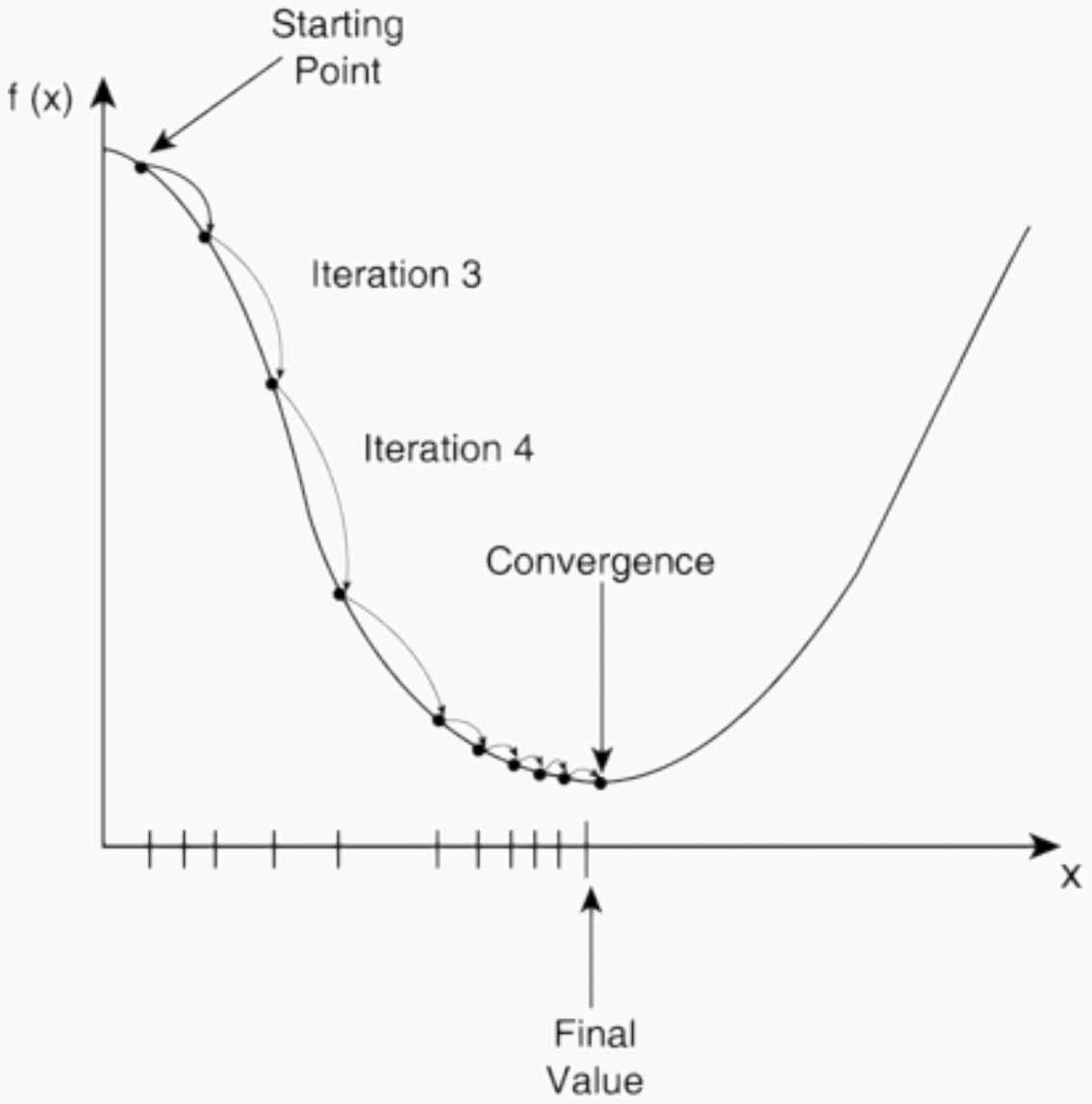
Таким образом мы можем применять основанные на вычислении градиента методы, не боясь застрять в локальном минимуме.

ГРАДИЕНТНЫЙ СПУСК

Градиент – обобщение производной на многомерный случай. Это вектор, показывающий направления роста функции и по модулю равный скорости роста. Обозначается $\nabla f(x)$.

Градиентный спуск – простейший метод численной оптимизации: суть метода в последовательном движении в направлении противоположном градиенту.

ГРАДИЕНТНЫЙ СПУСК



ГРАДИЕНТНЫЙ СПУСК

$$\theta_i = \theta_{i-1} - \alpha \nabla f_\theta$$

Где θ_i – вектор параметров функции f на итерации i.

f – целевая функция.

λ – learning rate, может быть как константой, так и функцией от номера итерации.

Рассмотрим алгоритм градиентного спуска для задачи линейной регрессии:

Инициализируем вектор параметров θ случайным образом и будем двигаться по градиенту функционала ошибки:

$$\theta_i = \theta_{i-1} - \alpha \nabla_\theta J$$

Найдём градиент функционала:

$$\nabla_\theta J(\theta) = \frac{\partial}{\partial \theta} \left[\frac{1}{2} \sum_i^m (h_\theta(x^{(i)}) - y^{(i)})^2 \right] = \sum_i^m (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)}$$

Алгоритм имеет вид:

1. Инициализировать θ
2. Повторять до схождения:

$$\theta_i = \theta_{i-1} - \alpha \sum_{i}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x^{(i)}$$

Можно ли использовать другие подходы?

ЛИНЕЙНАЯ РЕГРЕССИЯ

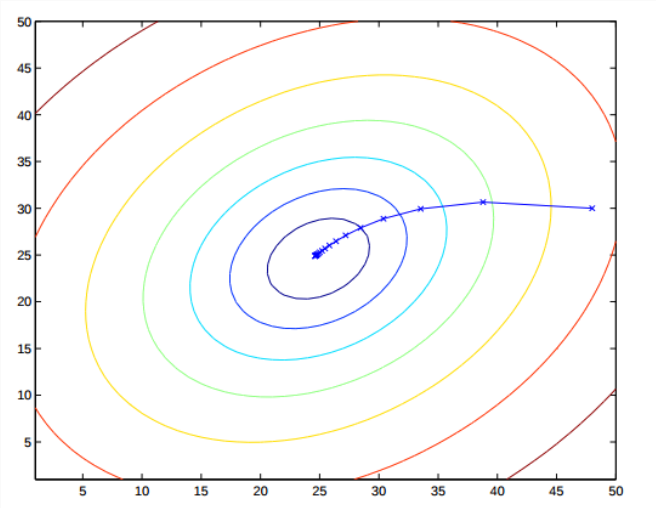


Рис. 1: Функция потерь выпуклая

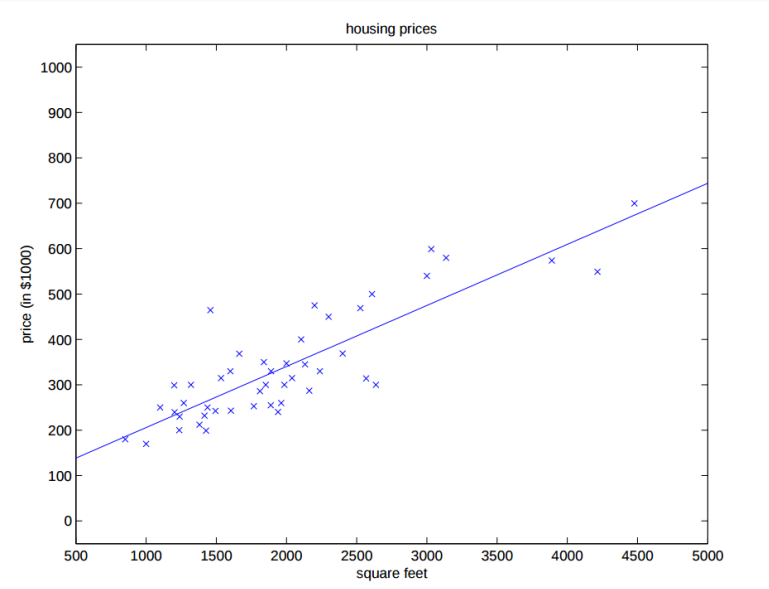


Рис. 2: Найденная функция

Решим задачу в матричной форме:

Форма функциональной зависимости:

$$h_\theta(X) = X\theta$$

Функционал ошибки:

$$J(\theta) = \frac{1}{2}(X\theta - y)^T(X\theta - y)$$

ЛИНЕЙНАЯ РЕГРЕССИЯ

Градиент функционала ошибки:

$$\nabla_\theta J(\theta) = \nabla_\theta \left[\frac{1}{2}(X\theta - y)^T(X\theta - y)\right] = \ldots = X^T X\theta - X^T y$$

$$X^T X\theta = X^T y$$

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

Теперь решим задачу в вероятностной постановке:

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^i$$

$$p(\epsilon^i) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(\epsilon^i)^2}{2\sigma^2}) \Rightarrow$$

$$\Rightarrow p(y^i)|x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2})$$

Выпишем функцию правдоподобия:

$$l(\theta) = \log L(\theta) = \log \prod_i^m p(y^i)|x^{(i)};\theta) =$$

$$= \log \prod_i^m \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) =$$

$$= \sum_i^m \log\left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)\right] =$$

$$= m \log \frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\frac{1}{2}\sum_i^m (y^{(i)} - \theta^T x^{(i)})^2$$

Чтобы максимизировать $l(\theta)$, надо минимизировать

$$\frac{1}{2}\sum_{i}^{m}(y^{(i)} - \theta^T x^{(i)})^2$$