Чудесный мир регрессии

Машинное обучение, 20!7

Спасибо К. В. Воронцову, МФТИ, Data Factory Яндекса, O.D.S. и кофеину.

Малютин Е. А.

Содержание

Сегодня в программе:

Регрессия. Начало

Рис.: Мемасик для вдохновения

Регрессия. Начало

Back to the roots

Пусть $\exists \ \{x_1, x_2 ... x_l\} \ \in X$ – множество объектов; $\exists \ \{y_i\}_{i=1}^{l} \in Y$ – множество допустимых ответов Пары (x_i, y_i) – называются прецедентами, а совокупность пар $X^l = (x_i, y_i)_{i=1}^{l}$ – обучающая выборка, а так же существует зависимость (алгоритм): $y^* : X \to Y$ – его-то и необходимо восстановить.

Регрессия

- $a(x) = x_0 + \sum\limits_{i=1}^{l} \omega_i x_i \ \to a(x) = \sum\limits_{i=0}^{l} \omega_i x_i \to \overrightarrow{y} = X\overrightarrow{\omega} + \epsilon$
  - $\overrightarrow{y} \in R^n$ – объясняемая (или целевая) переменная; ϵ – случайная переменная.
  - ω – вектор параметров модели (в машинном обучении эти параметры часто называют весами);
  - X – матрица наблюдений и признаков размерности n строк на $m+1$ столбцов (включая фиктивную единичную колонку слева) с полным рангом по столбцам: $rank(X) = m+1$;

Регрессия. Начало

Условности

Можно выписать модель регрессии явным образом для отдельного объекта:

$$y_i = \sum_{j=0}^{n} w_j X_{ij} + \epsilon$$

И, соответственно, мы приходим к некоторым условиям:

- матожидание: $\forall i : E[\epsilon_i] = 0$;
- гомоскедастичность: $\forall : Var(\epsilon_i) = \sigma^2 < \infty$
- некоррелированны: $\forall i \neq j : Cov(\epsilon_i, \epsilon_j) = 0$

Регрессия. Начало

Веса

Оценка $\overline{\omega_i}$ весов w_i называется **линейной**, если:

$$\overline{\omega_i} = \omega_{1i}y_1 + \omega_{2i}y_2 + \ldots + \omega_{ni}y_n;$$

Несмещённая оценка: $E[\overline{\omega_i}] = \omega_i;$

МНК: $\mathcal{L}(X, y, \omega) = \frac{1}{2n}\sum\limits_{i=1}^{n}(y_i - \omega^T x_i)^2;$

Найдем минимум данной ошибки:

$$\frac{d\mathcal{L}}{d\omega} = \frac{1}{2n}(-2X^T y + 2X^T X\omega)$$

И, соответственно:

$$\frac{d\mathcal{L}}{d\omega} = 0 \Leftrightarrow \omega = (X^T X)^{-1} X^T y$$

Регрессия. Начало

Метод максимального правдоподобия

Правдоподобие: $L = \Pi p(y_i x_i, \alpha)$
Log-likelihood: $W(\alpha) = log(L) = \sum ln(P(y|X, \alpha))$
Положим ошибки нормально-распределёнными: $\epsilon_i \backsim \mathcal{N}(0, \sigma^2)$
И тогда для нашей модели: $p(y_i|x_i, w) = \mathcal{N}(\sum\limits_j w_j X_{ij}, \sigma^2)$

$$log(p(X, y|w)) = log(\Pi\ \mathcal{N}(\sum_j w_j X_{ij}, \sigma^2)) = -\frac{n}{2} log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum(y_i - \omega^T x_i)^2$$

И, максимизируя это вот всё:

$$w = arg\max_w p(X, y|w) = arg\max_\omega -\mathcal{L}(X, y, \omega)$$

Вывод: минимизация МНК эквивалентна максимизации МП.

Регрессия. Начало

Разложение ошибки на смещение и разброс

- истинное значение целевой переменной складывается из некоторой детерминированной функции $f(\overrightarrow{x})$ и случайной ошибки $\epsilon y = f(\overrightarrow{x}) + \epsilon$;
- ошибка распределена нормально с центром в нуле и некоторым разбросом: $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- мы пытаемся приблизить детерминированную, но неизвестную функцию $f(\overrightarrow{x})$ линейной функцией от регрессоров $f^{(}\overrightarrow{x})$, которая, в свою очередь, является точечной оценкой функции f в пространстве функций.

Регрессия. Начало

Разложение ошибки

Рассмотрим ошибку в точке $\overrightarrow{x}$:

$E[\overrightarrow{x}] = E[(x - \widehat{f}(\overrightarrow{x}))] = E[y^2] + E[\widehat{f}^2] - 2E[y\widehat{f}]$;

По отдельности:

$E[y^2] = Var(y) + E[y]^2 = \sigma^2 + f^2$; $\;E[\widehat{f}^2] = Var(\widehat{f}) + E[f]^2$

$E[y\widehat{f}] = E[(f + \epsilon)\widehat{f}] = f\;E[\widehat{f}]$

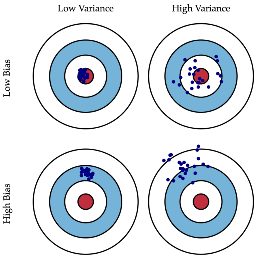
$Err(\overrightarrow{x}) = E[(y - \widehat{f}(\overrightarrow{x}))^2] = \sigma^2 + f^2 + Var(\widehat{f}) + E[\widehat{f}^2] - 2fE[f] =$

$= (f - E[\widehat{f}])^2 + Var(\widehat{f}) + \sigma^2 = Bias(\widehat{f}^2) + Var(\widehat{f}) + \sigma^2$

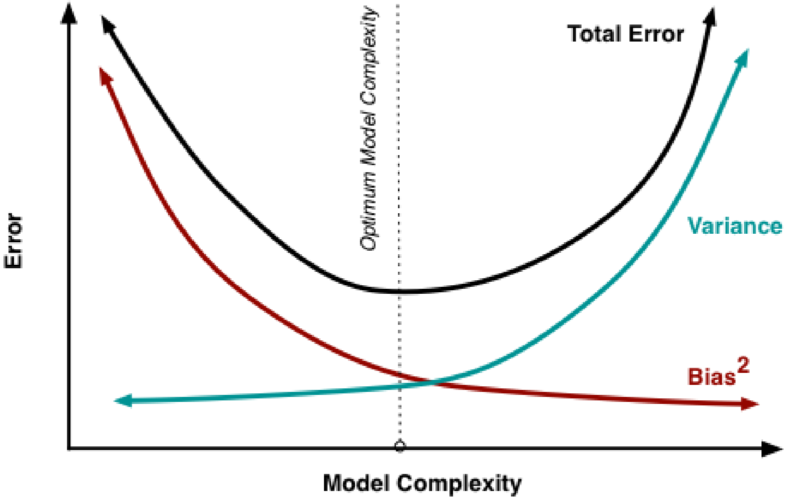
Регрессия. Начало

Ошибки:

- квадрат смещения: $Bias(\widehat{f})$ – средняя ошибка по всевозможным наборам данных;
- дисперсия: $Var(\widehat{f})$ – на сколько ошибка будет отличаться, если обучать модель на разных наборах данных;
- неустранимой ошибки: σ^2



Регрессия. Начало



Регрессия. Начало

Регуляризация

- При мультиколлинеарности в данных получаем $(X^T X)^{-1}$ – экстремально большие значения собственных чисел $(\frac{1}{\lambda_i})$

Регрессия. Начало

Регуляризация

- При мультиколлинеарности в данных получаем $(X^T X)^{-1}$ – экстремально большие значения собственных чисел $(\frac{1}{\lambda_i})$
- Решение? Регуляризация по Тихонову:

$$\mathcal{L}(X, \overrightarrow{y}, \overrightarrow{\omega}) = \frac{1}{2n}||\overrightarrow{y} - X\overrightarrow{\omega}||^2 + ||\mathcal{G}\overrightarrow{\omega}||^2$$

Регрессия. Начало

Регуляризация

- При мультиколлинеарности в данных получаем $(X^T X)^{-1}$ – экстремально большие значения собственных чисел $(\frac{1}{\lambda_i})$
- Решение? Регуляризация по Тихонову:

$$\mathcal{L}(X, \overrightarrow{y}, \overrightarrow{\omega}) = \frac{1}{2n}||\overrightarrow{y} - X\overrightarrow{\omega}||^2 + ||\mathcal{G}\overrightarrow{\omega}||^2$$

- Часто используется в таком виде: $\mathcal{G} = \frac{\lambda}{2}E$

Регрессия. Начало

Регуляризация

- При мультиколлинеарности в данных получаем $(X^T X)^{-1}$ – экстремально большие значения собственных чисел $(\frac{1}{\lambda_i})$
- Решение? Регуляризация по Тихонову:

$$\mathcal{L}(X, \overrightarrow{y}, \overrightarrow{\omega}) = \frac{1}{2n}||\overrightarrow{y} - X\overrightarrow{\omega}||^2 + ||\mathcal{G}\overrightarrow{\omega}||^2$$

- Часто используется в таком виде: $\mathcal{G} = \frac{\lambda}{2}E$
- Точное решение:

$$\overrightarrow{\omega} = (X^T X + \lambda E)^{-1} X^T \overrightarrow{y}$$

Регрессия. Начало

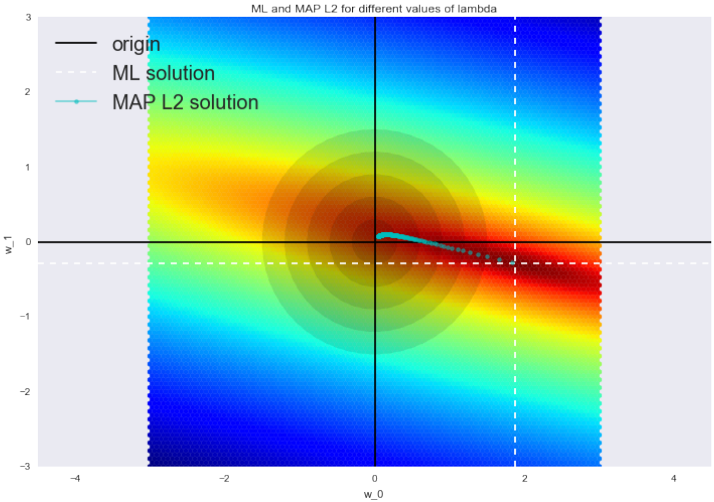
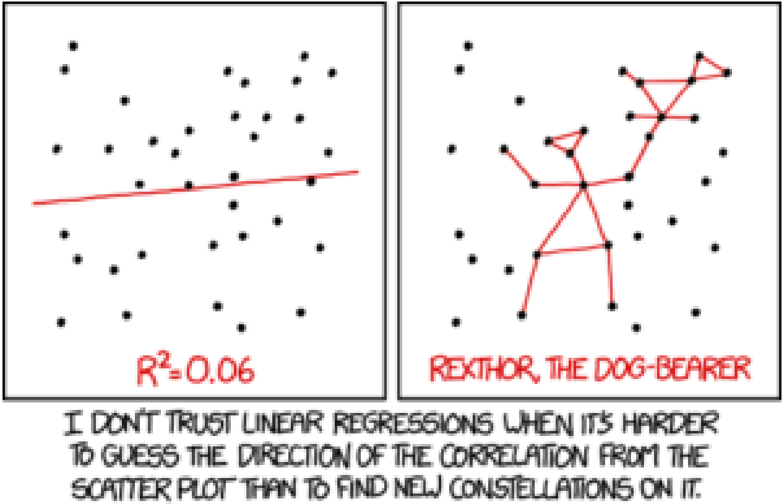


Рис.: Особенности регуляризации

Мемасик для вдохновения



Логистическая регрессия

бинарный классификатор на основе регрессии: $a(\overrightarrow{x}) = sign(\overrightarrow{\omega}^T x)$

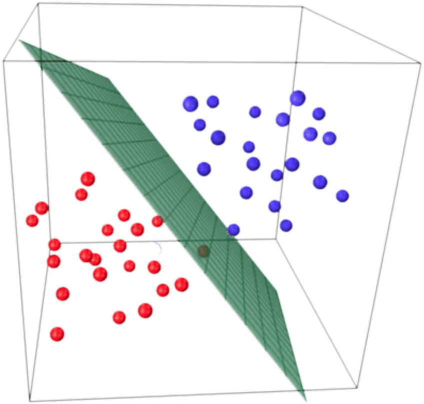


Рис.: Линейно-разделимая выборка

Логистическая регрессия

$$p_+ = P(y_i = 1 | \overrightarrow{x_i}, \overrightarrow{\omega})$$

Клиент	Вероятность невозврата
Mike	0.78
Jack	0.45
Larry	0.13
Kate	0.06
William	0.03
Jessica	0.02

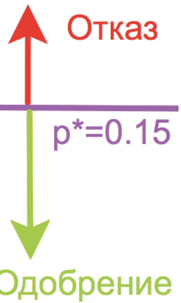
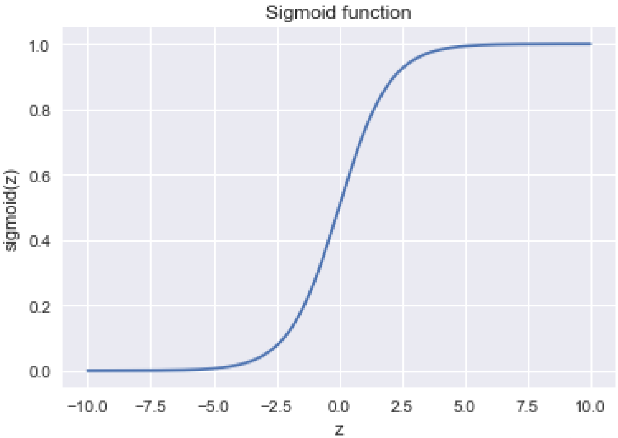


Рис.: Пример бинаризации

Логистическая регрессия

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



Логистическая регрессия

$\exists X, P(X); OR(X) \equiv \frac{P(X)}{1-P(X)}$; (отношение вероятностей) $P(X) \in [0,1]; OR(X) \in R$

Вычисляем лог. регрессию

- Вычислить значение $w_0 + w_1x_1 + w_2x_2 + ... = w^T\overrightarrow{x}$. (уравнение $\overrightarrow{w}^T\mathbf{x} = 0$ задает гиперплоскость, разделяющую примеры на 2 класса);
- Вычислить логарифм отношения шансов: $log(OR_+) = \overrightarrow{w}^T\overrightarrow{x}$.
- Вычисляем вероятность: $p_+ = \frac{OR_+}{1+OR_+} = \frac{exp^{\overrightarrow{w}^T\overrightarrow{x}}}{1+exp^{\overrightarrow{w}^T\overrightarrow{x}}} = \frac{1}{1-exp^{-\overrightarrow{w}^T\overrightarrow{x}}} = \sigma(\overrightarrow{w}^T\overrightarrow{x})$

Логистическая регрессия

Начало

- $p_+ = P(y_i = 1|\overrightarrow{x_i}, \overrightarrow{\omega}) = \sigma(\overrightarrow{w}^T, \overrightarrow{x})$

Логистическая регрессия

Начало

- $p_+ = P(y_i = 1|\overrightarrow{x_i}, \overrightarrow{\omega}) = \sigma(\overrightarrow{w}^T, \overrightarrow{x})$
- $p_- = P(y_i = -1|\overrightarrow{x_i}, \overrightarrow{\omega}) = 1 - \sigma(\overrightarrow{w}^T, \overrightarrow{x}) = \sigma(-\overrightarrow{w}^T, \overrightarrow{x})$

Логистическая регрессия

Начало

- $p_+ = P(y_i = 1|\overrightarrow{x_i}, \overrightarrow{\omega}) = \sigma(\overrightarrow{w}^T, \overrightarrow{x})$
- $p_- = P(y_i = -1|\overrightarrow{x_i}, \overrightarrow{\omega}) = 1 - \sigma(\overrightarrow{w}^T, \overrightarrow{x}) = \sigma(-\overrightarrow{w}^T, \overrightarrow{x})$
- следим за руками...

Логистическая регрессия

Начало

- $p_+ = P(y_i = 1|\overrightarrow{x_i}, \overrightarrow{\omega}) = \sigma(\overrightarrow{w}^T, \overrightarrow{x})$
- $p_- = P(y_i = -1|\overrightarrow{x_i}, \overrightarrow{\omega}) = 1 - \sigma(\overrightarrow{w}^T, \overrightarrow{x}) = \sigma(-\overrightarrow{w}^T, \overrightarrow{x})$
- следим за руками...
- $P(y = y_i|\overrightarrow{x_i}, \overrightarrow{\omega}) = \sigma(y_i \overrightarrow{w}^T, \overrightarrow{x})$

Логистическая регрессия

Начало

- $p_+ = P(y_i = 1|\overrightarrow{x_i}, \overrightarrow{\omega}) = \sigma(\overrightarrow{w}^T, \overrightarrow{x})$
- $p_- = P(y_i = -1|\overrightarrow{x_i}, \overrightarrow{\omega}) = 1 - \sigma(\overrightarrow{w}^T, \overrightarrow{x}) = \sigma(-\overrightarrow{w}^T, \overrightarrow{x})$
- следим за руками...
- $P(y = y_i|\overrightarrow{x_i}, \overrightarrow{\omega}) = \sigma(y_i \overrightarrow{w}^T, \overrightarrow{x})$
- $M(\overrightarrow{x_i}) = y_i \overrightarrow{\omega} \overrightarrow{x}$ – отступ

Логистическая регрессия

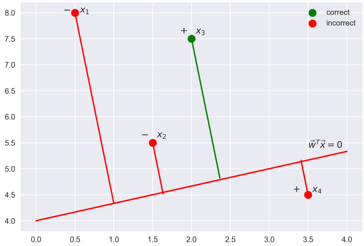


Рис.: Отсутпы на объектах

Отступы

- $M >> 0$ – метка поставлена правильно; $M << 0$ – выброс, шум
- $M \simeq 0$ – ну не знаю...

Логистическая регрессия

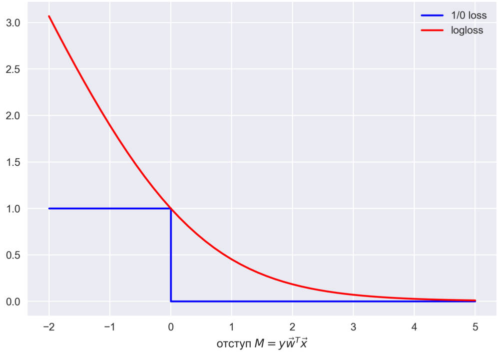
Угадайте что?

$$P\left(\vec{y} \mid X, \vec{w}\right) = \prod_{i=1}^{\ell} P\left(y = y_i \mid \vec{x}_i, \vec{w}\right),$$

$$\begin{aligned}
\log P\left(\vec{y} \mid X, \vec{w}\right) &= \log \prod_{i=1}^{\ell} P\left(y = y_i \mid \vec{x}_i, \vec{w}\right) \\
&= \log \prod_{i=1}^{\ell} \sigma\left(y_i \vec{w}^T \vec{x}_i\right) \\
&= \sum_{i=1}^{\ell} \log \sigma\left(y_i \vec{w}^T \vec{x}_i\right) \\
&= \sum_{i=1}^{\ell} \log \frac{1}{1 + \exp^{-y_i \vec{w}^T \vec{x}_i}} \\
&= -\sum_{i=1}^{\ell} \log\left(1 + \exp^{-y_i \vec{w}^T \vec{x}_i}\right)
\end{aligned}$$

$$\mathcal{L}_{\updownarrow\wr\}}(X, \vec{y}, \vec{w}) = \sum_{i=1}^{\ell} \log\left(1 + \exp^{-y_i \vec{w}^T \vec{x}_i}\right).$$

Логистическая регрессия



Градиентный спуск

Обощенный алгоритм градиентного спуска:

- Минимизируем эмпирический риск: $Q(w) = \sum\limits_{i=1}^{l} \mathcal{L}_i(\omega) \to \min\limits_{\omega}$

Градиентный спуск

Обощенный алгоритм градиентного спуска:

- Минимизируем эмпирический риск: $Q(w) = \sum\limits_{i=1}^{l} \mathcal{L}_i(\omega) \to \min\limits_{\omega}$
- Начальное приближение: $w^0 = start$

Градиентный спуск

Обощенный алгоритм градиентного спуска:

- Минимизируем эмпирический риск: $Q(w) = \sum_{i=1}^{l} \mathcal{L}_i(\omega) \to \min\limits_{\omega}$
- Начальное приближение: $w^0 = start$
- $t = 1 \dots n$ итеративно пересчитываем:

$$w^{(t+1)} = w^{(t)} - h * \nabla Q(w^{(t)}), \qquad \nabla Q(w) = (\frac{\partial Q(w)}{\partial w_i})_{j=0}^{n}$$

Градиентный спуск

Обощенный алгоритм градиентного спуска:

- Минимизируем эмпирический риск: $Q(w) = \sum\limits_{i=1}^{l} \mathcal{L}_i(\omega) \to \min\limits_{\omega}$
- Начальное приближение: $w^0 = start$
- $t = 1 \dots n$ итеративно пересчитываем:

$$w^{(t+1)} = w^{(t)} - h * \nabla Q(w^{(t)}), \qquad \nabla Q(w) = (\frac{\partial Q(w)}{\partial w_i})_{j=0}^{n}$$

- Всё вместе:

$$w^{(t+1)} = w^{(t)} - h * \sum_{i=1}^{l} \nabla \mathcal{L}_i(w^{(t)})$$

где h – это шаг градиента(learning rate).

Градиентный спуск

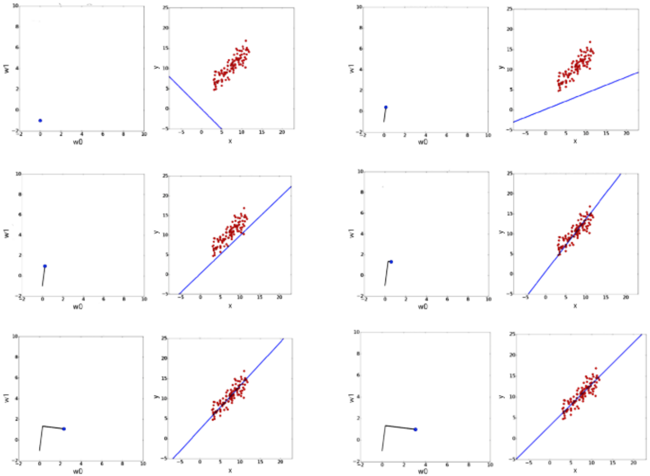


Рис.: Пример Gradient Descent на одномерной регрессии

Градиентный спуск

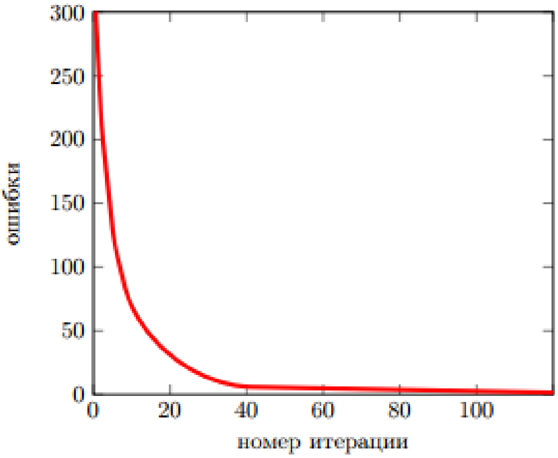


Рис.: Ошибка Gradient Descent

Градиентный спуск

Инициализация весов:

- как выбрать инициализацию?

Градиентный спуск

Инициализация весов:

- как выбрать инициализацию?
- Случайно: $w_j = random(-\frac{1}{2n}, \frac{1}{2n})$

Градиентный спуск

Инициализация весов:

- как выбрать инициализацию?
- Случайно: $w_j = random(-\frac{1}{2n}, \frac{1}{2n})$
- Эвристика:

$$w_j = \frac{\langle y, f_j \rangle}{\langle f_j, f_j \rangle}$$

Градиентный спуск

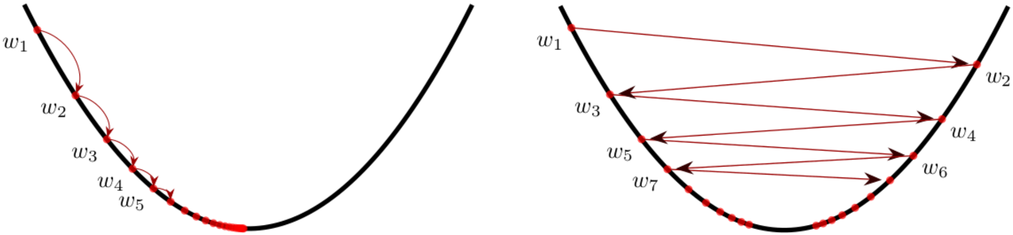


Рис.: Случай маленького и большого шага

$$h_t = \frac{k}{t} \text{— адаптивный шаг}$$

Градиентный спуск

Пример с регрессией

- Задача оптимизации:

$$Q(\omega, X) = \frac{1}{l}||X\omega - y||^2 \to \min_{\omega}$$

- Градиентный спуск:

$$\nabla_{\omega} Q(\omega, X) = \frac{2}{l} X^T (X\omega - y)$$

- $||X\omega - y||^2$ – вектор ошибок

Гридентный спуск

problems?

- j-я компонента для градиента:

$$\frac{\partial Q}{\partial \omega_j} = \frac{2}{l} \sum_{i=1}^{l} x_i^j (\langle w_i, x_i \rangle - y_i)^2$$

Гридентный спуск

problems?

- j-я компонента для градиента:

$$\frac{\partial Q}{\partial \omega_j} = \frac{2}{l} \sum_{i=1}^{l} x_i^j (\langle w_i, x_i \rangle - y_i)^2$$

- Выход – стохастический градиент: $w^0 = 0$

Гридентный спуск

problems?

- j-я компонента для градиента:

$$\frac{\partial Q}{\partial \omega_j} = \frac{2}{l} \sum_{i=1}^{l} x_i^j (\langle w_i, x_i \rangle - y_i)^2$$

- Выход – стохастический градиент: $w^0 = 0$
- Шаг: $\omega^t = \omega^t - h_t \nabla Q(w^{t-1}, \{x_i\})$

Гридентный спуск

problems?

- j-я компонента для градиента:

$$\frac{\partial Q}{\partial \omega_j} = \frac{2}{l}\sum_{i=1}^{l} x_i^j(\langle w_i, x_i\rangle - y_i)^2$$

- Выход – стохастический градиент: $w^0 = 0$
- Шаг: $\omega^t = \omega^t - h_t \nabla Q(w^{t-1}, \{x_i\})$
- Остановка: $||w^t - w^{t-1}|| < \epsilon$

Градиентный спуск

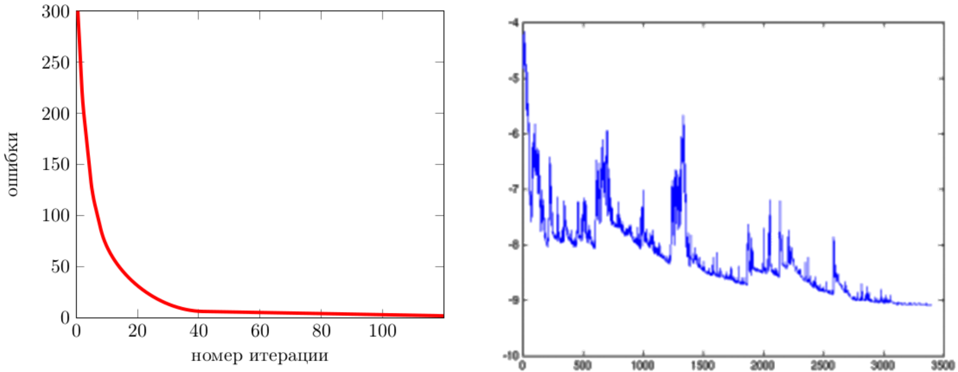


Рис.: Сходимость классического и стохастического градиетного спуска

Градиентный спуск

Преимущества и недостатки

Преимущества:

- Легко обобщается
- Потоковый
- Работает с большими данными
- Можно заканчивать, даже не предъявив всю выборку

Недостатки:

Градиентный спуск

Преимущества и недостатки

Преимущества:

- Легко обобщается
- Потоковый
- Работает с большими данными
- Можно заканчивать, даже не предъявив всю выборку

Недостатки:

- Многоэкстремальный
- Переобучение
- Если функция потерь имеет горизонтальные асимптоты, то процесс может попасть в состояние «паралича».

Градиентный спуск

Крутые эвристики

- **Нормализация данных:**

$$x_j = \frac{x_j - x_m in}{x_m ax - x_m in} \quad \text{либо} x_j = \frac{x_j - x_{med}}{x_{var}}$$

Градиентный спуск

Крутые эвристики

- **Нормализация данных:**

$$x_j = \frac{x_j - x_min}{x_max - x_min} \quad \text{либо} x_j = \frac{x_j - x_{med}}{x_{var}}$$

- **Порядок предъявления объектов:**

Градиентный спуск

Крутые эвристики

- **Нормализация данных:**

$$x_j = \frac{x_j - x_m in}{x_m ax - x_m in} \quad \text{либо} x_j = \frac{x_j - x_{med}}{x_{var}}$$

- **Порядок предъявления объектов:**
  - перетасовка объектов (shuffling)

Градиентный спуск

Крутые эвристики

- **Нормализация данных:**

$$x_j = \frac{x_j - x_min}{x_max - x_min} \quad \text{либо} x_j = \frac{x_j - x_{med}}{x_{var}}$$

- **Порядок предъявления объектов:**
  - перетасовка объектов (shuffling)
  - предъявлять те объекты, на которых была допущена ошибка

Градиентный спуск

Крутые эвристики

- **Нормализация данных:**

$$x_j = \frac{x_j - x_min}{x_max - x_min} \quad \text{либо} x_j = \frac{x_j - x_{med}}{x_{var}}$$

- **Порядок предъявления объектов:**
  - перетасовка объектов (shuffling)
  - предъявлять те объекты, на которых была допущена ошибка
  - сравнить величину ошибки на предъявленном объекте с некоторым порогом

Градиентный спуск

Квадратичная регуляризация:

- Штраф за норму: $Q_\tau(w) = Q(w) + \tau/2||w||^2$
- $w = w(1 - h * \tau) + h * Q'(w)$

Ещё эвристики

- выбор величины шага
  - От t: $h = 1/t$
  - Решить оптимизацию: $Q(w - hQ'(w)) \to \min\limits_{\omega}$
- выбивание из локальных минимумов
- ранний останов.

Регрессии

To be continued...