Производная

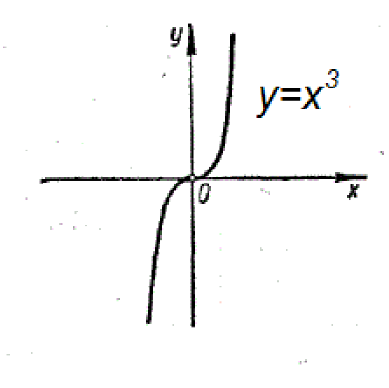
$$f^{`}(x)=\lim_{\Delta x\to 0}\frac{\Delta y}{\Delta x}=\lim_{\Delta x\to 0}\frac{f(x+\Delta x)-f(x)}{\Delta x}.$$

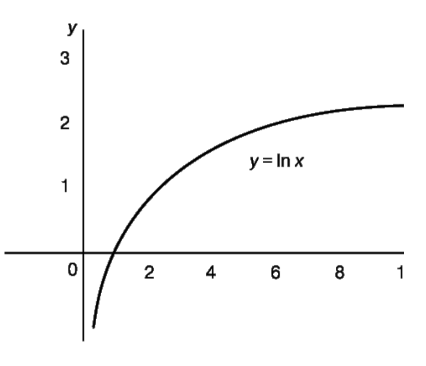
Эпсилон

Слишком маленький эпсилон может привести к тому, что разность будет настолько мала, что почти равна нулю.

f (x + eps) - f(x) ≈ 0

Слишком большой эпсилон приводит к тому, что мы не знаем, как поведет себя функция в точке x + eps.





Частная производная

$$\frac{\partial z}{\partial x}=f'_x(x,y)\ \text{(где } y=\text{const)},$$

$$\frac{\partial z}{\partial y}=f'_y(x,y)\ \text{(где } x=\text{const)}.$$

Градиент

вектор, который показывает направление наибольшего роста функции

Градиентный спуск

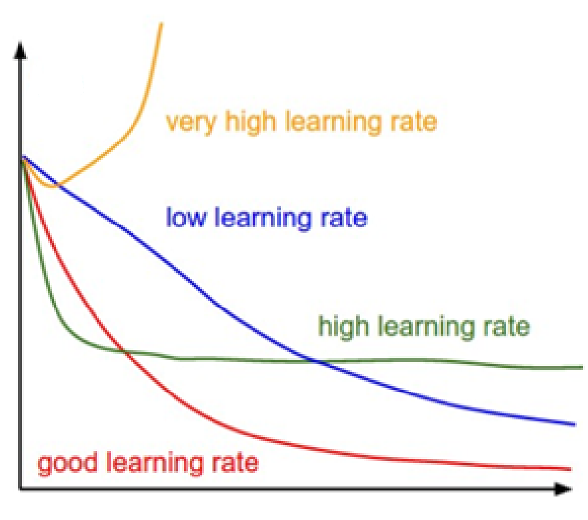
Задача: f(x) → min

Алгоритм:

1. Берем произвольный x и считаем в этой точке градиент.
2. Делаем цикл пока grad(x) < eps
3. Если он больше, то высчитываем новый x по формуле:

$$\theta=\theta-\eta\cdot\nabla_\theta J(\theta).$$

Learning rate



Градиентный спуск
в линейной регрессии

Задача линейной регрессии: минимизировать функцию потерь

$$\sum_{i=1}^{n}\left(\vec{\beta}\cdot\vec{x_i}-y_i\right)^2 \longrightarrow \quad \min$$

$$L\left(D,\vec{\beta}\right) \longrightarrow \quad \min$$

Аналитический способ:

1. Представить функцию потерь в матричном виде

$$Y^{\mathsf T}Y-Y^{\mathsf T}X\vec{\beta}-\vec{\beta}^{\mathsf T}X^{\mathsf T}Y+\vec{\beta}^{\mathsf T}X^{\mathsf T}X\vec{\beta}$$

2. Найти производную этой функции

$$-2X^{\mathsf T}Y+2X^{\mathsf T}X\vec{\beta}$$

3. Приравнять производную к нулю и выразить бэта

$$\vec{\beta}=\left(X^{\mathsf T}X\right)^{-1}X^{\mathsf T}Y$$

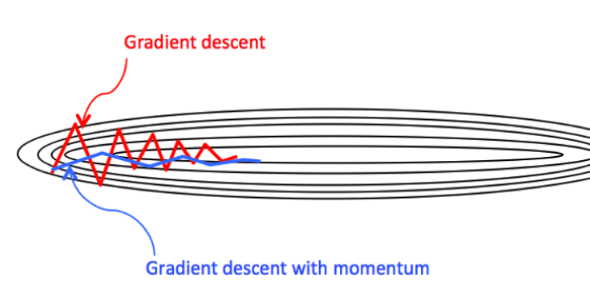
Стохастический градиентный спуск

Градиент оптимизируемой функции считается как градиент от случайно выбранного подмножества данных.

$$\theta=\theta-\eta\cdot\nabla_\theta J(\theta;x^{(i)};y^{(i)}).$$

Момент

$$v_t=\gamma v_{t-1}+\eta\nabla_\theta J(\theta)$$
$$\theta=\theta-v_t$$



Ускоренный градиент Нестерова

Дает нам приблизительное представление о следующем положении параметров.

$$v_t=\gamma v_{t-1}+\eta\nabla_\theta J(\theta-\gamma v_{t-1})$$
$$\theta=\theta-v_t$$

Метод отжига

Применяя его к градиентоному спуску, мы каждый раз уменьшаем понемногу ленинг рэйт по мере приближения к минимуму функции.

Adagrad

$$\theta_{t+1,i}=\theta_{t,i}-\frac{\eta}{\sqrt{G_{t,ii}+\epsilon}}\cdot g_{t,i}.$$

Adadelta и RMSprop

RMSprop и Adadelta были разработаны независимо друг от друга примерно в одно и то же время из-за необходимости решить проблему радикального снижения learning rate в Adagrad.

$$E[g^2]_t=\gamma E[g^2]_{t-1}+(1-\gamma)g_t^2.$$

$$\Delta\theta_t=-\frac{\eta}{\sqrt{E[g^2]_t+\epsilon}}g_t.$$

Adam

$$m_t=\beta_1 m_{t-1}+(1-\beta_1)g_t$$
$$v_t=\beta_2 v_{t-1}+(1-\beta_2)g_t^2$$

1. Вычисляем параметры

$$\hat{m}_t=\frac{m_t}{1-\beta_1^t}$$
$$\hat{v}_t=\frac{v_t}{1-\beta_2^t}$$

2. Считаем оценку параметров

$$\theta_{t+1}=\theta_t-\frac{\eta}{\sqrt{\hat{v}_t}+\epsilon}\hat{m}_t.$$

3. Подставялем в формулу