ECE-GT 6143

L3

1. Recap: Linear Regression

Today:

① Gradient descent

② Application: Linear Regression (again)

Fit a linear model:

$$y = w_0 + w_1 x$$

- Lose function: $MSE(w) = \frac{1}{n}\sum [y_i - (w_0 + w_1 x_i)]^2$

$$w_0 = \bar{y} - w_1 \bar{x}$$

$$w_1 = \frac{6xy}{6^2 x}$$

← Univariate 单变量 Core

Multivariate:

$$X = \begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(d)} \end{pmatrix} \quad \& \quad y$$

Linear model: $y = w_0 + w_1 x^{(1)} + w_2 x^{(2)} + \cdots + w_d x^{(d)}$

w_0: intercept; $w_1, \ldots, w_d$: coefficients

Assumption: Data & Label are zero mean. ignore w_0. $y = \sum_{i=1}^{d} w_i x_i = \langle w, x \rangle$

Step 1: $y = \langle w, x \rangle$

Step 2: $MSE = \frac{1}{2}\sum_{i=1}^{n} (y_i - \langle w, x \rangle)^2$

d个变量，n次测量 nxd matrix

Step 3 Minimize MSE with respect to W

nxd dx1

$$X = \begin{bmatrix} -x_1- \\ -x_2- \\ \vdots \\ -x_n- \end{bmatrix}$$

$$\begin{bmatrix} \langle w, x_1 \rangle \\ \langle w, x_2 \rangle \\ \vdots \\ \langle w, x_n \rangle \end{bmatrix} \quad XW = \begin{bmatrix} -x_1- \\ -x_2- \\ \vdots \\ -x_n- \end{bmatrix} \begin{bmatrix} | \\ w \\ | \end{bmatrix}$$

本来是$\frac{1}{n}$，n为样本数目是常数，设为$\frac{1}{2}$方便求导后有系数

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$MSE = \left(\frac{1}{2}\right) \|y - Xw\|_2^2$$

矩阵不能直接除

$$w^* = \left(\sum_{i=1}^{n} x_i x_i^T\right)^{-1} \left[\sum_{i=1}^{n} y_i x_i\right]$$

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

dx1 dxd dxn nx1

Multivariate Case

$n \geq d$ $X^T X$ 可逆

intuitive: we need $n \geq d$ equations to solve d variable

Bad! ↑ inversion takes $O(d^3)$ time

这样找 MSE 最小的系数太慢

So today:

Goal: Improve this running time sufficiently using

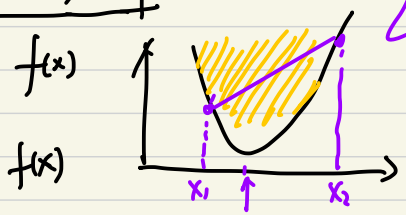
Gradient descent

梯度下降 ⇒ 找最优点（极值）

First:

convexity

Convex function (凸函数)



$\alpha x_1+(1-\alpha)x_2$ (if $\alpha=0.5 \Rightarrow$ mid point)

Words: straight line $(x_1, f(x_1))$ & $(x_2, f(x_2))$ should be above $f(x)$

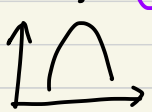
Integer Realm 实数空间

Math: 定义: d维映射到1维

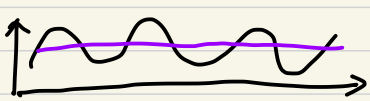
$f: \mathbb{R}^d \to \mathbb{R}$ for any $x, y \in \mathbb{R}^d$

$f(\alpha x+(1-\alpha)y) \le \alpha f(x)+(1-\alpha)f(y)$ all α in $[0,1]$

Opposite of convex: Concave 凹

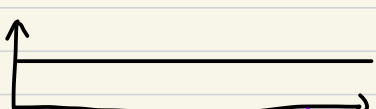


sinx



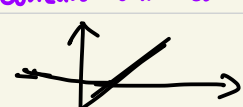
Neither concave nor convex

constant



Both Concave and convex

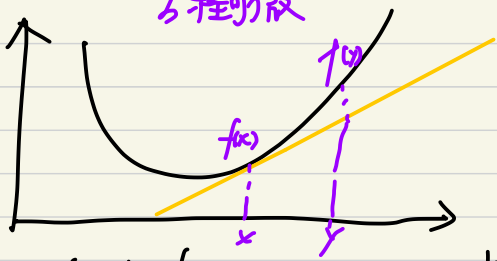
Straight



Assume: function has differentiability / smoothness

Property

方程形成



Words: tangent of f() at any point lies "below" the curve

Math:

Scalar: For any x, y, $f(y) \geq f(x) + \frac{df(x)}{dx}(y-x)$

Vector 向量:

$$\begin{bmatrix} \frac{\partial f}{\partial x^{(1)}} \\ \frac{\partial f}{\partial x^{(2)}} \\ \vdots \\ \frac{\partial f}{\partial x^{(d)}} \end{bmatrix} = \nabla f(x)$$

For all x, y: $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle$ x, y 横生排

Observation:

If $\nabla f(x) = 0$, then $f(y) \geq f(x)$ for all y

$\Rightarrow x$ is a global minimum of f

此时 x 点为最小值点

Gradient descent

寻找最小MSE对应系数

(MSE最小值所对应的)

goal: fix x^* that minimize $f(x)$

Approach: find Δ, $f(w+\Delta) < f(w)$ 往更下面方向找 最小点

- $w \leftarrow w+\Delta$
- Repeat

$\nabla f(x)$ points in the direction of maximum change in f (该点的梯度方向)

$w_k - \alpha_k \nabla f(w_k)$ defines how much I want to move on that direction

往负梯度方向走

"step size" "learning rate"

for $k = 0, 1, 2 \cdots, T$ T: number of epochs

terminate after T epochs 几元为更新次数

or when ∇f becomes small

$(AB)^T = B^T A^T$

$\frac{\partial Ax}{\partial x} = A^T$

Application of GD: Linear Regression

(MSE) $L(w) = \frac{1}{2} \|y - Xw\|_2^2$ → 对w求导，矩阵求导就是梯度

$\nabla L(w) = -X^T(y - Xw)$

$= X^T(Xw - y)$

GD: - initialize w_0

- iterate

~~往梯度下降方向~~ 负梯度方向

$w_{k+1} \leftarrow w_k - \alpha_k X^T(Xw - y)$ untill T or ∇ become small

一代一代更新 w　　α_k: step size

Running time: $O(nd) \cdot T$

- Set error parameter ε, terminate when $\|\nabla f(x)\|_2 \leq \epsilon$

- can prove that this happens after $T = \log_{1/\rho} \left[\frac{\|w_0\|_2}{\varepsilon} \right]$ 所以可以确定T (更新次数)

T为迭代次数时，上面的关系成立 (停止的条件)。

Where $\rho = \frac{1 - \frac{l}{L}}{1 + \frac{l}{L}}$ (ρ is a number between 0 ~ 1)

$\rho = \frac{L - l}{L + l}$

l & L are smallest and largest eigenvalue of $X^T X$

Eigenvalue: $Av = \lambda v$

特征值

A ↑ eigen vector　　λ ↑ eigen value