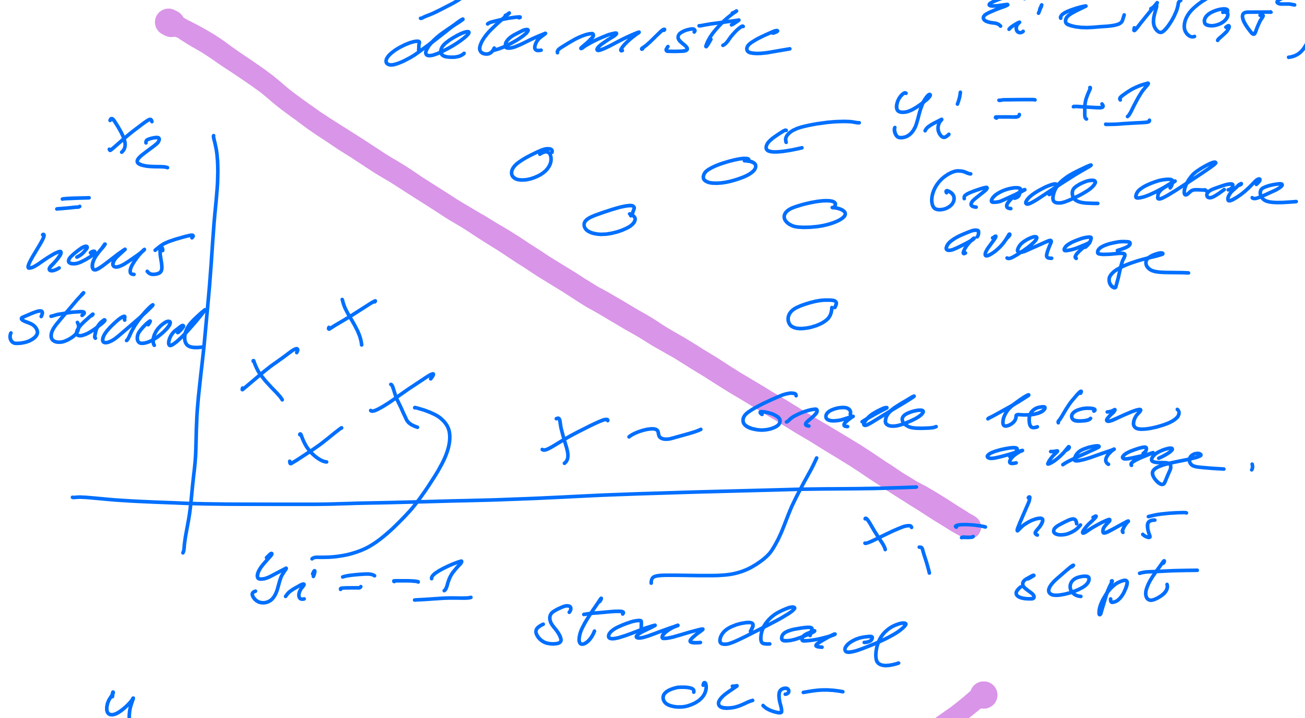
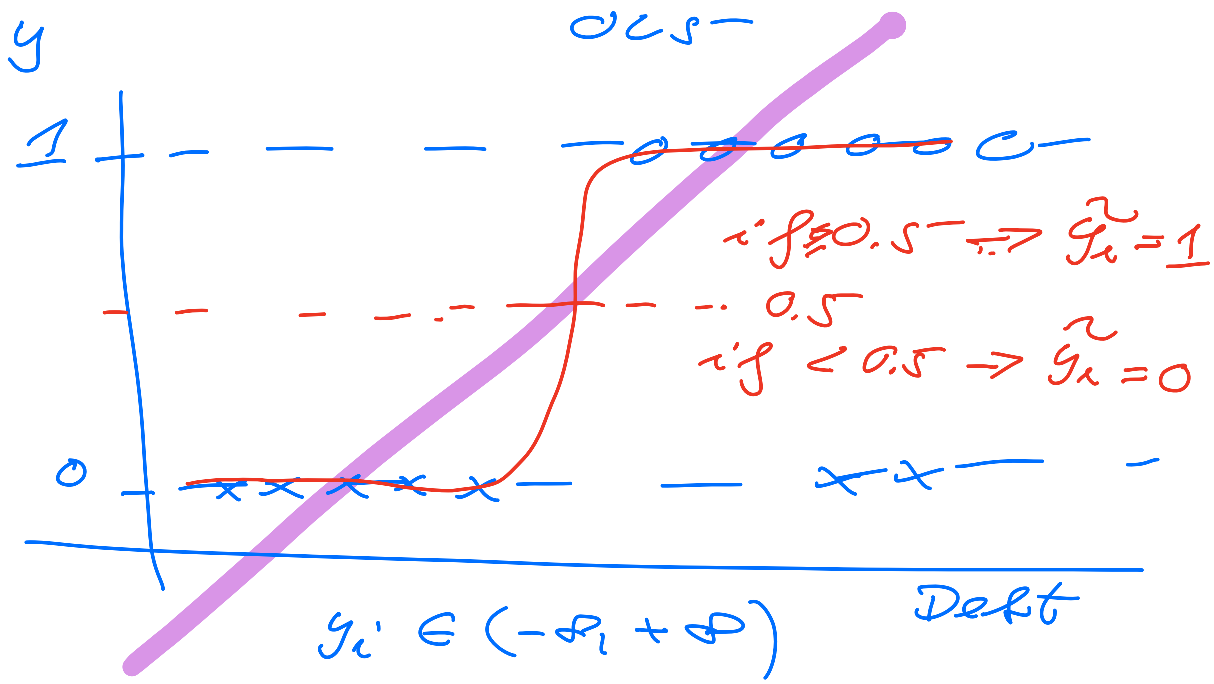
# Comp Sci, December 12, 2022

$$y_i = f(x_i) + \varepsilon_i$$

deterministic

$\varepsilon_i \sim N(0, \sigma^2)$




$y_i \in (-\infty, +\infty)$

$y_i \in \{0, 1\}$

Replace $f(x_i) \to p(x_i)$

$$0 \le p(x_i) \le 1$$

simple $p(x) = \dfrac{e^x}{1+e^x}$

$$\Rightarrow \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = p(x)$$

Two parameters, $\beta_0$, $\beta_1$

in principle $x \in (-\infty, +\infty)$

$p(x) \le 0.5$, then $y = 0$

$p(x) > 0.5$, then $y = 1$

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

$$p(y_i \mid x_i, \beta) = p(y_i = +1 \mid x_i \beta) = p_i$$

$$p(y_i = 0 \mid x_i \beta) = 1 - p_i$$

$$\sum_{i=0}^{1} p(y_i \mid x_i \beta) = 1$$

logit transformation

$$g(x_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

$$= \beta_0 + \beta_1 x_i$$

Linear Regression

$$y = X\beta + \varepsilon = \hat{y} + \varepsilon$$

$$y \sim N(X\beta, \sigma^2)$$

Replace with

$$y = p(x) + \varepsilon$$

which distribution does $y$ follow? which distribution does $\varepsilon$ follow?

$y = 1$, then $p(x) = P$

$$\left( \overset{\text{example}}{=} \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \right)$$

$y = 0$ has probability $1 - P$

if $y = 1$ has $\boxed{\underline{\varepsilon = 1 - P}\ \text{with probability } P}$

$$y = p(x) + \varepsilon$$

if $y = 0 \Rightarrow \varepsilon = -P$

with probability $1-p$

assume that $\varepsilon$ has mean value $= 0$

$$\mathbb{E}[\varepsilon] = (1-p)p - p(1-p) = 0$$

$$\mathbb{E}[x] = \sum_{i \in D} p(x_i) x_i$$

$$\mathrm{var}[\varepsilon^2] = (1-p)^2 p + (-p)^2 (1-p)$$

$$= p(1-p)$$

$\Rightarrow$ $\varepsilon$ follows a Binomial distribution,

assumption

$y_i$ are i.i.d.

independent and identically distributed

$y_i = 1$ has probability $p(x_i)$

$$= p_i = p(y_i = 1 | x_i, \beta)$$

$y_i = 0$ has probability

$$1 - p(x_i) = 1 - p_i$$

we assume binomial

$$p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

$$D = \{(x_0, y_0), (x_1, y_1) \ldots (x_{n-1}, y_{n-1})\}$$

$$y_i \in \{0, 1\}$$

$$p(D|\beta) = \prod_{i=0}^{n-1} p_i^{y_i}(1-p_i)^{1-y_i}$$

optimal $\beta$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\max}\, P(D|\beta)$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min}\left(-\log P(D|\beta)\right)$$

$$C(\beta) = -\log P(D|\beta)$$

$$= -\sum_{i=0}^{n-1}\Big[y_i \log p_i + (1-y_i)\log(1-p_i)\Big]$$

$$\frac{\partial C(\beta)}{\partial \beta} = 0$$

$$\log p_i = \beta_0 + \beta_1 x_i - \log(1+e^{\beta_0+\beta_1 x_i})$$

$$C(\beta) = -\sum_i\Big[y_i(\beta_0+\beta_1 x_i) - \log(1+e^{\beta_0+\beta_1 x_i})\Big]$$

$$\frac{\partial C}{\partial \beta_0} = 0 = -\sum_{i=0}^{n-1}(y_i - p_i)$$

$$= g_0$$

$$\frac{\partial C}{\partial \beta_1} = 0 = -\sum_{i=0}^{n-1} x_i(y_i - p_i)$$

in general we have

$$\frac{\partial C}{\partial \beta} = 0 = -X^T(y-p)$$

$$p, y \in \mathbb{R}^n$$

$$X \in \mathbb{R}^{n \times p}$$

Exercise: show that

$$\frac{\partial^2 C}{\partial \beta \, \partial \beta^T} = X^T W X = \text{Hessian}$$

W is a diagonal only matrix

$$W_{ii} = p_i(1-p_i)$$

Linear regression:

OLS $\quad \frac{\partial^2 C}{\partial \beta \, \partial \beta^T} = X^T X \cdot 2/n$

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

$$\frac{\partial C}{\partial \beta} = g = X^T(p-y) = 0$$

our model

$$p_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

Due to non-linearity of $\beta$, we don't get a simple analytical expression for $\beta$.

$$\frac{\partial C}{\partial \beta} = X^T(p-y) = 0$$

Basic-method: Newton-Raphson (iterative solution)

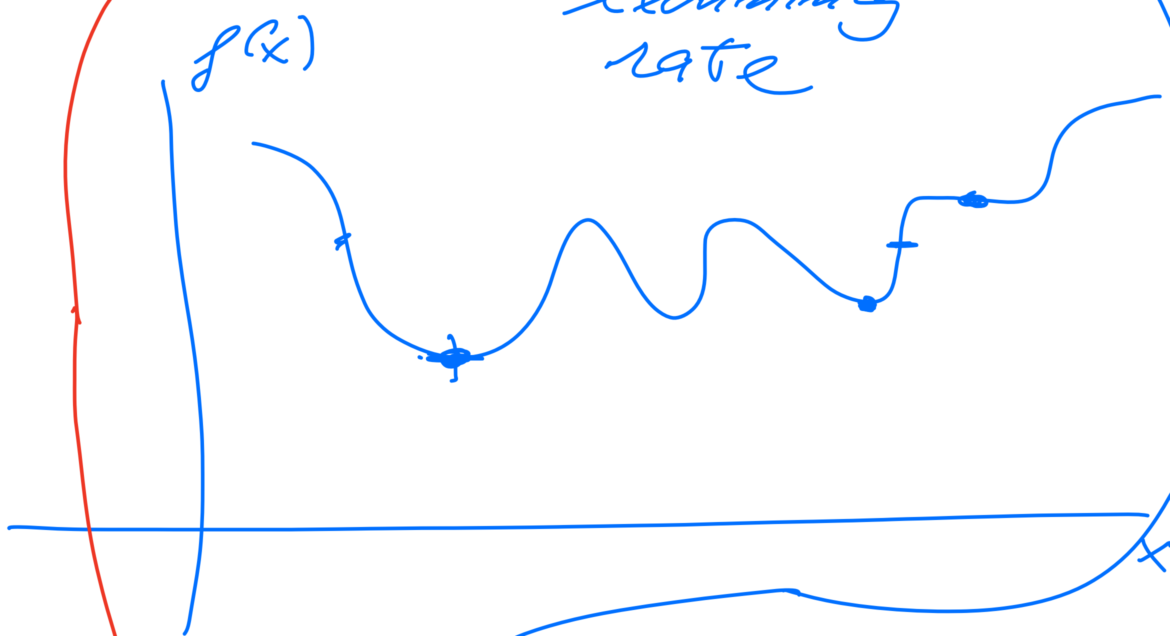
$$\beta^{(k+1)} = \beta^{(k)} - \underbrace{H^{-1}(\beta^{(k)})}\, g(\beta^{(k)})$$

$$H = \begin{bmatrix} \frac{\partial^2 C}{\partial \beta_0 \partial \beta_0} & \frac{\partial^2 C}{\partial \beta_0 \partial \beta_1} & \cdots & \frac{\partial^2 C}{\partial \beta_0 \partial \beta_{p-1}} \\ \vdots & & & \\ \frac{\partial^2 C}{\partial \beta_{p-1} \partial \beta_0} & \cdots & & \end{bmatrix}$$

in "all" algorithms

$$\beta^{(k+1)} = \beta^{(k)} - \gamma^{(k)} g(\beta^{(k)})$$

learning rate




$g(\beta^{(k)}) \to$ stochastic gradient descent

(efficient evaluations of gradients)

↳ momentum (memory)

GD
---
momentum SGD

Learning rate magic?

- fixed $\gamma^{(t)} = \gamma$

  $\gamma \in [10^{-5}, 10^{-4}, \ldots 10^{-1}]$

- schedulers for $\gamma^{(t)}$

  $\gamma^{(t)} \sim \gamma_0 \exp(-kT)$

- adaptive learning rates

  - ADAM
  - RootMeansquared propagation

(RMSprop)

— ADAgrad

Taylor expansion of

$C(\beta)$ around $\hat{\beta} - \beta^{(k)}$

$\hat{\beta} \Rightarrow \beta$

$$C(\beta) = C(\beta^{(k)})$$

$$+ \underbrace{g^T(\beta^{(k)})}_{g^T_{(k)}}(\beta - \beta^{(k)})$$

$$+ \frac{1}{2}(\beta - \beta^{(k)})^T H(\beta^{(k)})(\beta - \beta^{(k)})$$

$$+ O\left((\beta - \beta^{(k)})^3\right)$$

$$b = \beta - \beta^{(k)}$$

approximate to second derivative

$$C(\beta) \simeq C(\beta^{(k)}) + g_{(k)}^T b + \frac{1}{2} b^T H(\beta^{(k)}) b$$

$$\frac{\partial C(\beta)}{\partial \beta} = \frac{\partial C}{\partial b} = 0 \Rightarrow$$

$$\beta = \beta^{(k)} - H^{-1}(\beta^{(k)}) g_{(k)}$$

$$f(x) = c + \frac{1}{2} x^T A x + g^T x$$

$$\frac{\partial f}{\partial x} = 0 \Rightarrow Ax = -g$$

$$x = -A^{-1} g$$