Lecture January 22

Linear Regression

$$y_i = X_{i*}\beta + \varepsilon_i$$

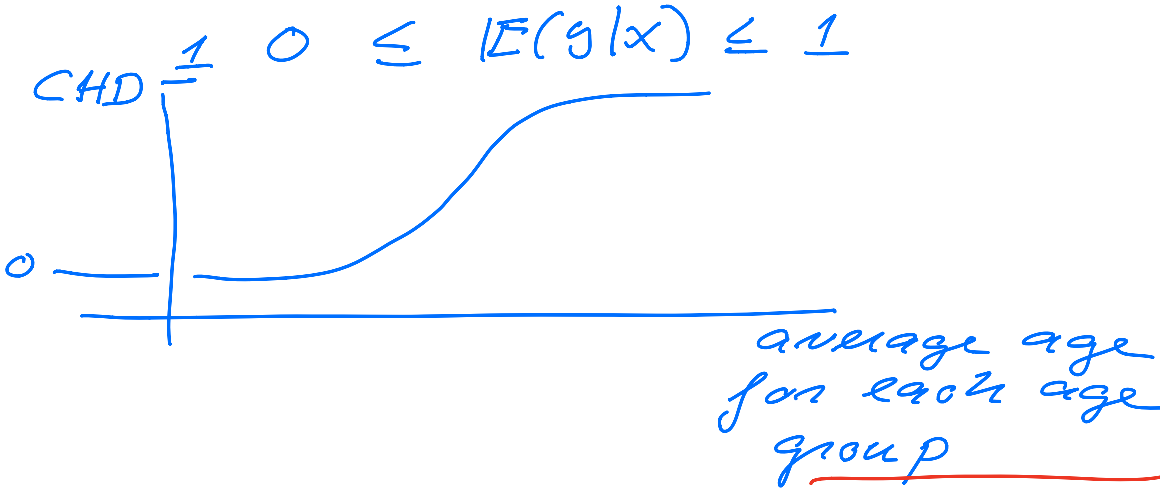
$$\mathbb{E}[y_i] = X_{i*}\beta = \beta_0 + \beta_1 x_i$$
(simplest choice)

$$= \mathbb{E}(y \mid x\beta)$$

This is a continuous function.

Binary case

$$0 \le \mathbb{E}(y \mid x) \le 1$$



$$p(x) = P(y \mid x) = \mathbb{E}(y \mid x) = \boxed{\frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}}$$

Linear Regression

$$y = \boxed{X\beta} + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

Logistic regression

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

y takes two values $0, 1$

$y = 1 \Rightarrow \quad \varepsilon = 1 - p(x)$ with $p(x)$

$y = 0 \Rightarrow \quad \varepsilon = -p(x)$ with probability $1-p(x)$

$$\sum_{i=1}^{2} p(x_i) = p(x) + (1-p(x)) = 1$$

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

$$\left(\mathbb{E}[x] = \sum p(x_i) x_i \right)$$

$\mathrm{var}[\varepsilon^2] = p(1-p)$, corresponds to the binomial distribution.

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

$y_i = 0 \quad ; \quad 1 - p(y_i | x_i \beta)$

$$p(y_i \mid x_i \beta) = p(x_i)$$

probability:

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

$$C(\beta) = \prod_{i=0}^{n-1} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

$$= \sum_{i=0}^{n-1} \log \left[p(x_i)^{y_i}(1 - p(x_i))^{1-y_i} \right]$$

$$\beta^{opt} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left[-C(\beta) \right]$$

$$-C(\beta) = -\sum_{i=0}^{n-1} \left[y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i)) \right]$$

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

$$-\frac{\partial C(\beta)}{\partial \beta_0} = 0 = \sum_{i=0}^{n-1} \left(y_i - p(x_i) \right) \quad x_i = 1$$

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

$\Rightarrow$ (in matrix-vector form)

$$\boxed{X^T(y-p)} \qquad p, y \in \mathbb{R}^n \quad X \in \mathbb{R}^{n \times p}$$

$$p = \{(p(x_0), p(x_1) \ldots p(x_{n-1}))\}$$

$$X^T(y-p) \in \mathbb{R}^p$$

$$\beta \in \mathbb{R}^p \quad (p=2, \beta_0, \beta_1)$$

$$X^T(y-p) = 0$$

is a non-linear equation in β.

2nd-derivative of $C(\beta)$ wrt to β

$$\frac{\partial^2 C(\beta)}{\partial \beta \partial \beta^T} = \underbrace{X^T W X}_{= \text{Hession}} = H$$

$$\left(W = p(y_i | x_i \beta)(1 - p(y_i | x_i \beta)) \right.$$

only diagonal elements

$W \in \mathbb{R}^{n \times n}$

$H \in \mathbb{R}^{p \times p}$.

H is a positive-definite matrix.

Newton-Raphson for Root searching $\underline{(x^T(y-p)=0)}$

$$f(x) = 0$$

$$x^{(k+1)} = x^{(k)} - \left[\frac{f(x)}{f'(x)}\right]_{x=x^{(k)}}$$

$$\|x^{(k+1)} - x^{(k)}\|_2 \leq \delta \approx 10^{-10}$$

$$\beta^{(k+1)} = \beta^{(k)} - \left(H^{-1} \frac{\partial C}{\partial \beta}\right)_{\beta=\beta^{(k)}}$$

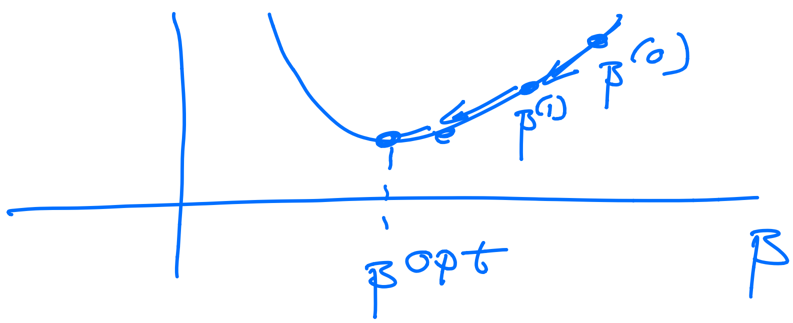
p large

$$\boxed{\beta^{(k+1)} = \beta^{(k)} - \gamma_k \left[\frac{\partial C}{\partial \beta}\right]_{\beta=\beta^{(k)}}}$$

new hyperparameter

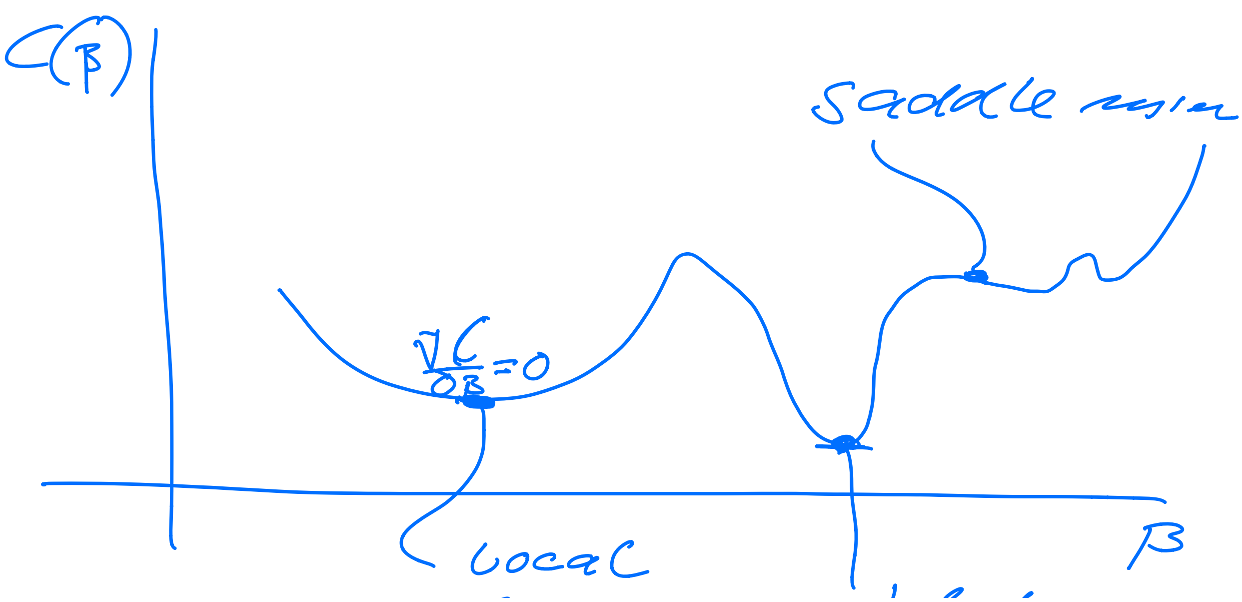
= Learning rate,

$$\gamma_k \Rightarrow \gamma = \{10^{-5}, 10^{-4}, 10^{-3}, \dots\}$$



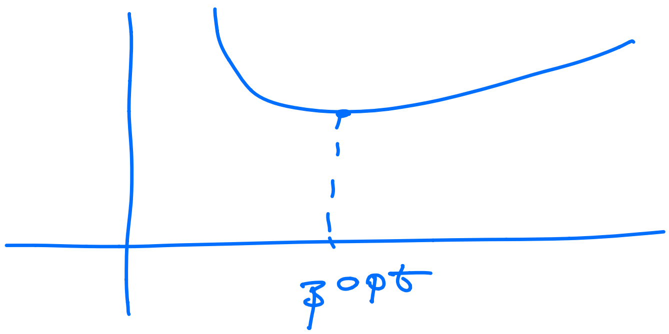
accuracy score

$$\text{score} = \frac{\sum_{i=0}^{n-1} I(y_i = \tilde{y}_i)}{\# \text{ events}}$$



min

global min

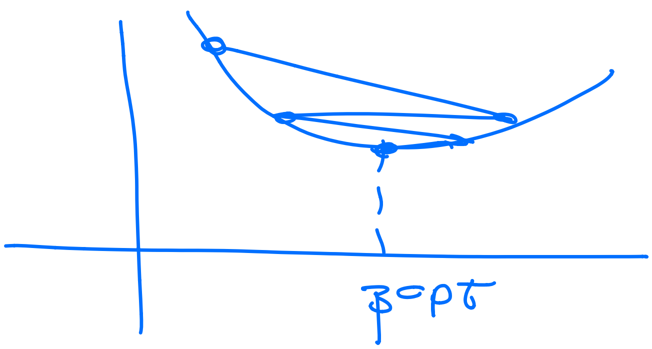


$\gamma < \boxed{\frac{2}{\lambda_{max}}}$

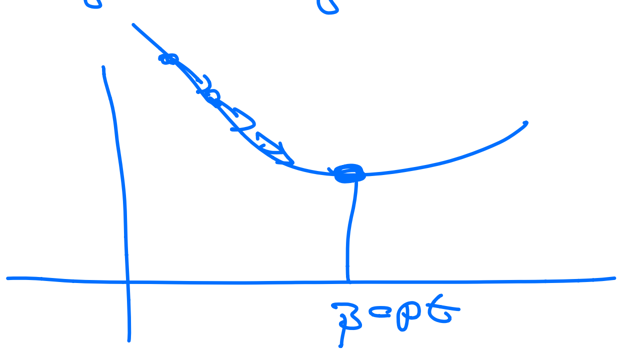
largest value of Hessian matrix

if $\underline{\gamma^{opt} < \gamma < 2\gamma^{opt}}$

$$\gamma^{opt} = \frac{1}{\lambda_{max}}$$



$\gamma < \gamma^{opt}$



opt

if $\gamma > \gamma^{[illegible]}$, no convergence