Lecture November 15

Logistic regression

$$0 \leq \mathbb{E}[(y|x)] \leq 1$$

in linear regression

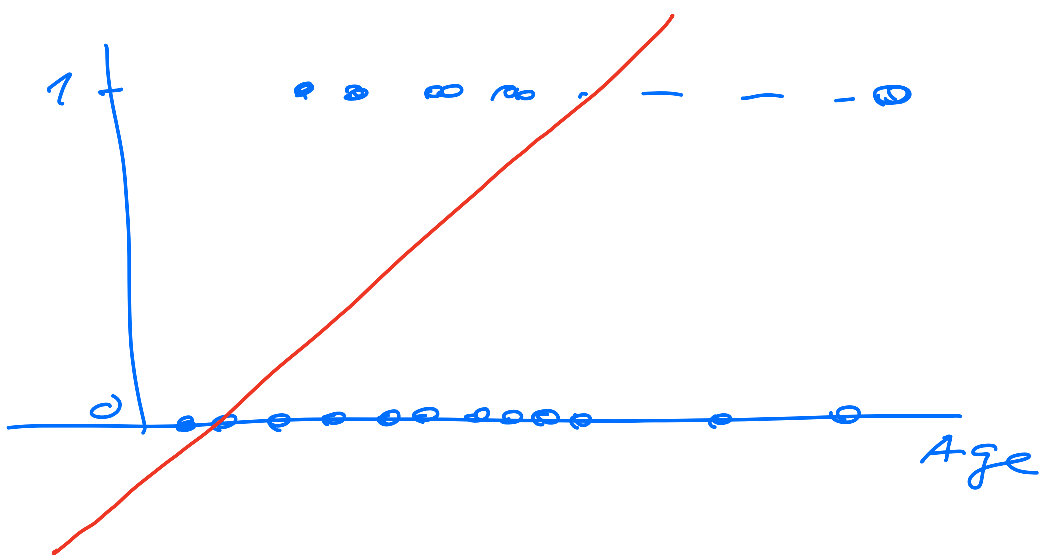
$$\mathbb{E}[y] = X\hat{\beta}$$

with a first-order polynomial

$$y_i = \sum_{j=0}^{p-1} x_{ij}\beta_j$$

$$= \beta_0 + x_i\beta_1$$

CHD



Lin regression

$$y = f(x) + \varepsilon \qquad f(x) \in (-\infty, +\infty)$$
$$\simeq X\beta + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

in logistic regression

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

need a model ?

$$p(x) \in [0, 1]$$

$$\left\{ \mathbb{E}[x] = \begin{array}{l} \int p(x)\, x \, dx \\ \sum_{i \in D} x_i \, p(x_i) \end{array} \right\}$$

$$\left. \begin{array}{l} y_i = p(x_i) + \varepsilon_i \\ y_i = \{0, 1\} \end{array} \right\} \quad \begin{array}{l} y_i \Rightarrow y \\ y = p(x) + \varepsilon \end{array}$$

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

with probability $p(x)$

$$y = 0 = p(x) + \varepsilon \Rightarrow \varepsilon = -p(x)$$

but now with probability $1 - p(x)$

$$p(y=1|x) + p(y=0|x) = p(x) + (1 - p(x))$$

$= \underline{1}$

$p(y_i = 1 \mid x_i) = p(x_i) = p$

$(y = y_i)$

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

$0 \leq p \leq \underline{1}$

which distribution does

ε - follow ?

$\mathbb{E}[\varepsilon] = \sum_i p_i \cdot \varepsilon_i =$

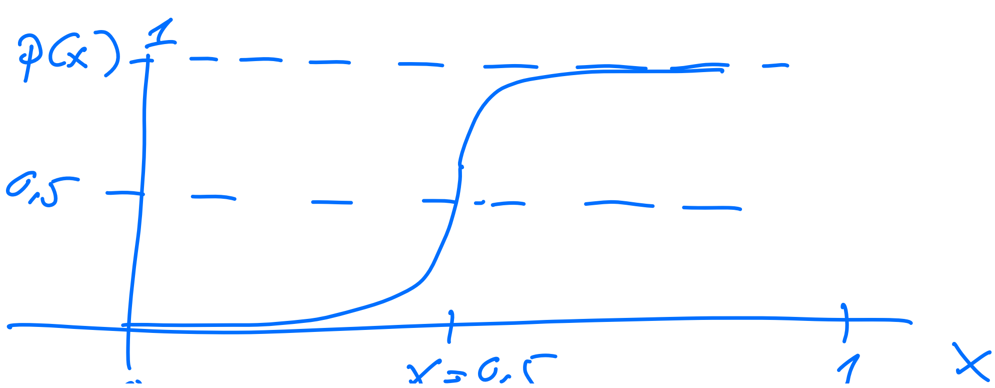
$(1-p)p - p(1-p) = 0$

$\mathbb{E}[\varepsilon] = 0$

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

$= p(1-p)$

P_ε = Binomial distribution



$x \in [0, 1]$

if $p(x) > 0.5$ then output $= 1$

else if $p(x) \leq 0.5$, then output $= 0$

Redefine:

$y_i = 1$ corresponds to

$p(x_i; y_i | \beta)$

$y_i = 0$, corresponds to $1 - p(x_i; y_i | \beta)$

a very popular model is the Logit/sigmoid

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

$$= \frac{e^{t_i}}{1 + e^{t_i}}$$

$t_i = \beta_0 + \beta_1 x_i$

We want to find a total probability $P(D|\beta)$ as probability $P(D|\beta)$

function of β so that with an optimal $\hat{\beta}$, we maximize $P(D|\hat{\beta})$

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

For one single event y_i

$P(x_i|\beta)$ and $1 - P(x_i|\beta)$

$y_i = 1$ $\qquad\qquad$ $y_i = 0$

$$P(x_i/\beta)^{y_i}\left(1 - P(x_i/\beta)\right)^{1-y_i}$$

$$P(D|\beta) = \prod_{i=0}^{n-1} P(x_i/\beta)^{y_i}\left(1 - P(x_i/\beta)\right)^{1-y_i}$$

$$\hat{\beta} = \arg\max_{\beta} P(D|\beta)$$

Maximum likelihood estimator

$$\frac{\partial P(D|\beta)}{\partial \beta} = 0$$

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

maximization problem

or a minimization problem

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

$$= -\sum_i \left\{ y_i \log P_i + (1-y_i)\log(1-P_i) \right\}$$

$$P_i = P(x_i|\beta)$$

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

$$= -\sum_{i=0}^{n-1} \left\{ y_i \left(\beta_0 + \beta_1 x_i - \log\left(1+e^{\beta_0+\beta_1 x_i}\right)\right) - (1-y_i)\log\left(1+e^{\beta_0+\beta_1 x_i}\right) \right\}$$

$$\frac{\partial C}{\partial \beta_0} = 0 = -\sum_i (y_i - P_i)$$

$$\frac{\partial C}{\partial \beta_1} = 0 = -\sum x_i (y_i - p_i)$$

generalize :

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

gradient

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

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

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

$p = \#$ features

For completeness

$$\frac{\partial^2 C}{\partial \beta \partial \beta^T} = X^T W X \in \mathbb{R}^{p \times p}$$

W has only diagonal elements

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

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

linear regression (gradient)

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

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

In logistic regression we have a non-linear equation in the unknown parameters β

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

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

$\Rightarrow$ needs to be solved numerically ?

Newton-Raphson's iterative

$$f(x) \Rightarrow f(x+h) = f(x) + h f'(x) + \frac{h^2}{2} f''(x) + O(h^3)$$

if we skip higher-order terms

$$f(x+h) = f(x) + hf'(x) = 0$$

$$f(x) + hf'(x) = 0$$

$$x \Rightarrow x_0 \qquad x+h \rightarrow x_0 + h = s$$

$$f(s) = f(x_0) + hf'(x_0) = 0$$

$$f(x_0) + (s - x_0) f'(x_0) = 0 \Rightarrow$$

$$s = x_0 - f(x)/f'(x) \Rightarrow$$

$$x^{(n+1)} = x^{(n)} - \frac{f(x^{(n)})}{f'(x^{(n)})}$$

start with guess $x^{(0)}$ and iterate till

$$|x^{(n+1)} - x^{(n)}| \leq \varepsilon \sim 10^{-10}$$

$$\frac{\partial C}{\partial \beta} = 0 \qquad \left(= f(x) \right) = \nabla_\beta C$$

$$\beta^{(n+1)} = \beta^{(n)} - \frac{\nabla_\beta C}{\frac{\partial^2 C}{\partial\beta\partial\beta^T}}$$

$$\frac{\partial^2 C}{\partial\beta\partial\beta^T} = X^T W X = H$$

Hessian matrix

is positive definite; all eigenvalues λ_i of H are larger than zero $\Rightarrow$

C is a concave and we at least a local minimum.

Reminder: Linear-regression and OLS

$$H = \frac{2}{n} X^T X = \frac{\partial^2 C}{\partial\beta\partial\beta^T}$$

$$C = \frac{1}{n}\sum_i (y_i - \tilde{y}_i)^2$$

$$\beta^{(n+1)} = \beta^{(n)} - H^{-1}(\beta^{(n)}) \nabla_\beta C(\beta^{(n)})$$

continue iterations till

$$|\beta^{(n+1)} - \beta^{(n)}| \leq \varepsilon$$

with a given random choice $\beta^{(0)}$

H and $\nabla_\beta C(\beta) \in \mathbb{R}^{p \times p}$

- Many FLOPS
  - we are performing n-iterations.
  - we have to invert $H \in \mathbb{R}^{p \times p}$
  - we need also $\nabla_\beta C(\beta)$

$$\nabla_\beta C(\beta) = -X^T(y - p)$$

$$\Rightarrow \nabla_{\beta_j} C(\beta) = -\sum_{i=0}^{n-1} x_{ji}(y_i - p_i)$$

- $H(\beta^{(n)})$, $\nabla_\beta C(\beta^{(n)})$ need to calculate for every iteration

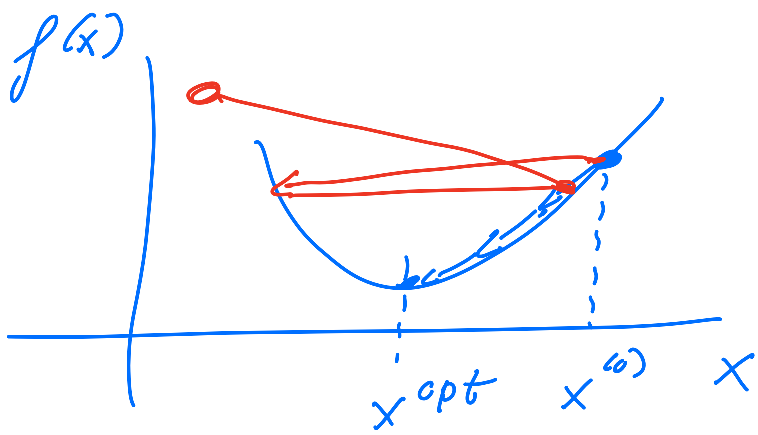
Gradient descent

$$\beta^{(n+1)} = \beta^{(n)} - \gamma \nabla_\beta C(\beta^{(n)})$$

learning rate

$$\gamma < \frac{2}{\lambda_{max}(H)}$$

parameter in ML theory

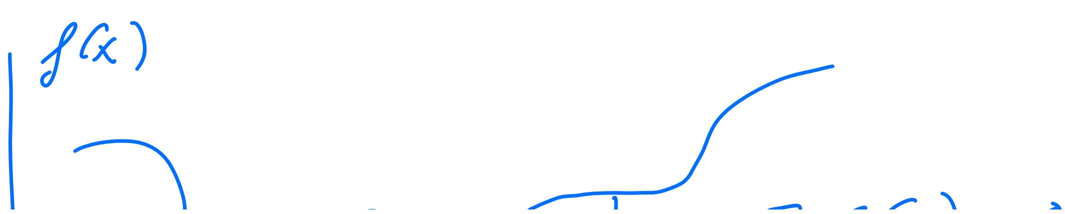


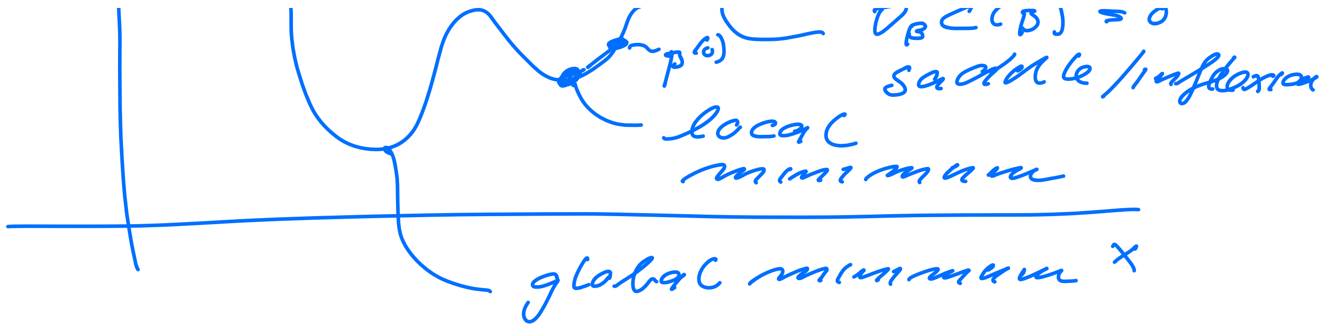
$= \hat{x} = x^{(n+1)}$

$$\gamma \leq \frac{2}{\lambda max}$$

$$\gamma > \frac{2}{\lambda max}$$

OLS : $H = \frac{2}{n} X^T X$





$\nabla_\beta L(\beta) = 0$
saddle/inflexion
$\sim\beta^{(0)}$
local minimum
global minimum
x