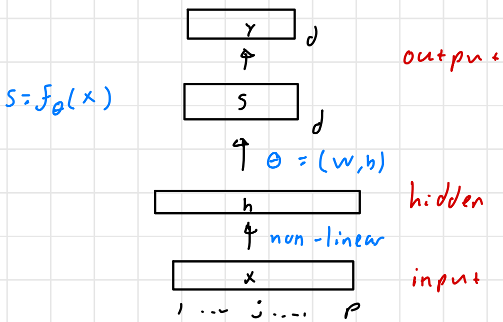
Lecture 3

- Regression:

i	x_i^T	y_i^T

$$P(y \mid x) = \begin{cases} \overbrace{\frac{1}{(\sqrt{2\pi\sigma^2})^d} \exp\left(-\frac{1}{2\sigma^2}|y-s|^2\right)}^{\sim \mathcal{N}(s,\, \sigma^2 I_d)} & \text{Continuous} \to \text{regression} \\ \underbrace{\frac{e^{\langle y, s\rangle}}{Z}}_{\sim \text{Multinomial}(p = \text{softmax}(s))} & \text{discrete / categorical} \to \text{classification} \end{cases}$$

- general model:



$$s = wh + b$$

$$\underbrace{\boxed{s}}_{d} = \boxed{w}\,\boxed{h} + \boxed{b}$$

- log-likelihood $(\theta) = \log p_\theta(y|x)$

- Regression:

$$\ell(\theta) = -\frac{1}{2\sigma^2} |y - s|^2$$

$$L(\theta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} |y_i - s_i|^2$$

- Classification

$$\ell(\theta) = \langle y, s \rangle - \log Z(s)$$

$$L(\theta) = \sum_{i=1}^{n} \left[\langle y_i, s_i \rangle - \log Z(s_i) \right]$$

$$y = \begin{array}{cc} \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} & \begin{array}{c} 1 \\ \vdots \\ \vdots \\ k \\ \vdots \\ \vdots \\ c \end{array} \end{array} \qquad s = \begin{pmatrix} s_1 \\ \vdots \\ s_k \\ \vdots \\ s_c \end{pmatrix} \qquad p = \begin{pmatrix} p_1 \\ \vdots \\ p_k \\ \vdots \\ p_c \end{pmatrix} = \mathrm{softmax}(s)$$

So $\langle y, s \rangle = s_k$

$$p_k = \frac{e^{s_k}}{\sum_{c=1}^{C} e^{s_c}} = p(y = k | x)$$

Least Squares

$$\mathrm{Loss}(\theta) = -\text{log-likelihood}(\theta) = \begin{cases} \text{Regression: } \frac{1}{2}|y - s|^2 \\ \text{Classification: } -\log p(y|x) \end{cases}$$

$$= -\left(\langle y, s \rangle - \log Z(s) \right)$$

Cross-entropy

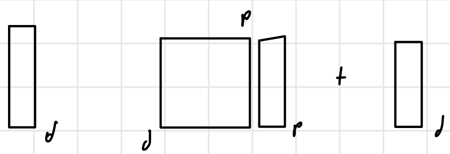
- Specialize to $d = 1$.

  - Regression:

	Father height	Mother height	Sons height
i	x_{i1}	x_{i2}	y_i

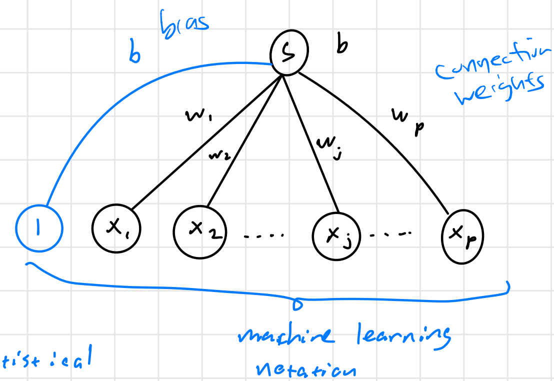
- Recall linear model:

$$s = Wx + b$$



$d = 1$

$$s = \langle w, x \rangle + b$$



$$s = x^T \beta + \beta_0$$ statistical notation

$$= \left(x_1 \; \cdots \; x_j \; \cdots \; x_p \right) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix} + \beta_0$$

intercept

coefficients

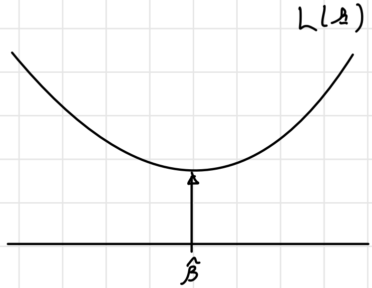
$$= \beta_0 + \sum_{i=1}^{p} x_i \beta_i$$

* Can make β_0 ambiguous:

$$s = x^T \beta$$

- Solving Least Squares Problem:

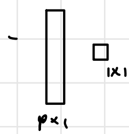
$$\text{Loss}(\beta) = \frac{1}{2}\sum_{i=1}^{n}(y_i - s_i)^2$$
$$= \frac{1}{2}\sum_{i=1}^{n} e_i^2$$
$$= \frac{1}{2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$



$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial L}{\partial e_i}\frac{\partial e_i}{\partial s_i}\frac{\partial s_i}{\partial \beta_k}, \qquad \text{note} \quad s_i = \sum_{j=1}^{p} x_{ij}\beta_j$$
$$= -\sum_{i=1}^{n} e_i x_{ik}$$

- Package:

$$\frac{\partial L}{\partial \beta} = \begin{array}{c} \\ \left(\frac{\partial L}{\partial \beta_k}\right) \\ \\ \end{array}\begin{array}{c} 1 \\ \vdots \\ k \\ \vdots \\ p \end{array} = -\sum_{i=1}^{n}\begin{array}{c} \\ \left(e_i x_{ik}\right) \\ \\ \end{array}\begin{array}{c} 1 \\ \vdots \\ k \\ \vdots \\ p \end{array} = -\sum_{i=1}^{n} \underset{p\times 1}{x_i}\ \underset{1\times 1}{e_i}$$

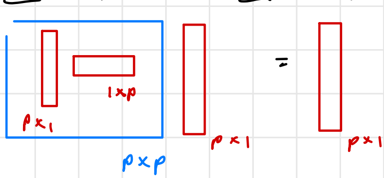


$$\frac{\partial L}{\partial \beta} = 0 \implies \sum_{i=1}^{n} x_i e_i = 0$$

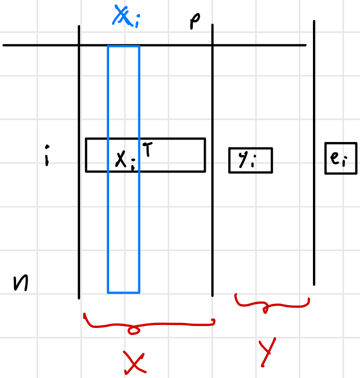
$$\sum_{i=1}^{n} x_i(y_i - x_i^T\beta) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i x_i^T\beta = 0$$

$$\sum x_i x_i^T\beta = \sum x_i y_i$$



$$\implies \hat{\beta} = \left(\sum_{i=1}^{n} x_i x_i^T\right)^{-1}\left(\sum_{i=1}^{n} x_i y_i\right)$$



$$\sum_{i=1}^{n} x_i x_i^T = \underbrace{(x_1 \cdots x_i \cdots x_n)}_{X^T} \overset{X}{\begin{pmatrix} x_1^T \\ \vdots \\ x_i^T \\ \vdots \\ x_n^T \end{pmatrix}}$$

meta-rule: Pretend each symbol is a number (scalar)

$$\sum x_i y_i = \underbrace{(x_1, \dots, x_i, \dots, x_n)}_{X^T} \overset{Y}{\begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}}$$

So: $\hat{\beta} = (X^T X)^{-1} X^T Y$

- $\text{Loss}(\beta) = \frac{1}{2} \sum_{i=1}^{n} e_i^2$

$$= \frac{1}{2} |e|^2$$

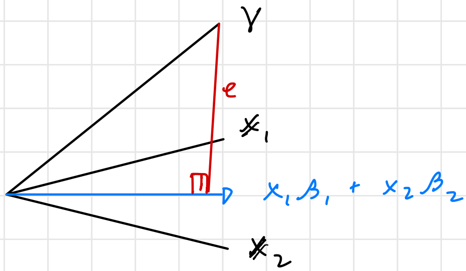
$$e = \begin{pmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} y_i - (x_{i1}\beta_1 + \dots + x_{ip}\beta_p \end{pmatrix}$$

$$= \frac{1}{2} \left| Y - X_1 \beta_1 + \dots + X_j \beta_j + \dots + X_p \beta_p \right|^2$$

$$= \frac{1}{2} \left| Y - (X_1 \dots X_j \dots X_p) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix} \right|^2$$

$$= \frac{1}{2} | Y - X\beta |^2$$

- Geometric View:



$$\frac{\partial L}{\partial \beta_k} = -\sum_{i=1}^{n} e_i x_{ik} = -\langle e, \mathbf{x}_k \rangle = 0 \quad \text{, so} \quad e \perp \mathbf{x}_k$$

$$\mathbf{x}_k^T e = 0$$

So
$$\begin{array}{c} 1 \\ \vdots \\ k \\ \vdots \\ p \end{array} \begin{pmatrix} \\ \\ \mathbf{x}_k^T e \\ \\ \\ \end{pmatrix} = 0$$

$$\begin{pmatrix} \\ \mathbf{x}_k^T \\ \\ \end{pmatrix} e = 0$$

$$X^T e = 0$$

So $X^T(Y - X\beta) = 0$

$X^T Y - X^T X \beta = 0$, Thus $\hat{\beta} = (X^T X)^{-1} X^T Y$

- Suppose $p(y_i | x_i) \sim \mathcal{N}(s_i, \sigma_i^2)$, different variance per observation. (precision)

$$p_\theta(y_i | x_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{1}{2\sigma_i^2}(y_i - s_i)^2\right)$$

$$\text{Loss}(\beta) = \frac{1}{2}\sum_{i=1}^{n} \frac{1}{\sigma_i^2}(y_i - x_i^T\beta)^2 = \frac{1}{2}\sum_{i=1}^{n} w_i (y_i - x_i^T\beta)^2$$

$w_i = \frac{1}{\sigma_i^2}$ weight $\Rightarrow$ weighted least squares

$$\frac{\partial L}{\partial \beta} = -\sum_{i=1}^{n} w_i x_i \underbrace{(y_i - x_i^T\beta)}_{e_i} = 0$$

$$\hat{\beta} = \left(\sum_{i=1}^{n} w_i x_i x_i^T\right)^{-1}\left(\sum_{i=1}^{n} w_i x_i y_i\right)$$

or

$$\underbrace{\begin{aligned}\tilde{x}_i &= \sqrt{w_i}\, x_i \\ \tilde{y}_i &= \sqrt{w_i}\, y_i\end{aligned}}$$

$\Downarrow$ translate to Least squares

$$\frac{1}{2}\sum(\tilde{y}_i - \tilde{x}_i^T\beta)^2$$

- Logistic Regression:

	x_i^T	$y_i \in \{0, 1\}$
	height, weight	gender

$$p(y|x) = \frac{e^{ys}}{Z} \begin{cases} p = \dfrac{e^s}{Z} = \dfrac{e^s}{1+e^s} & y=1 \\ 1-p = \dfrac{1}{Z} = \dfrac{1}{1+e^s} & y=0 \end{cases}$$

Thus $Z = 1 + e^s$

So $$p = p(y=1 \mid s) = \frac{e^s}{1+e^s} = \text{sigmoid}(s) = \sigma(s)$$

$$1 - p = \frac{1}{1+e^s}$$

$$\frac{p}{1-p} = e^s$$

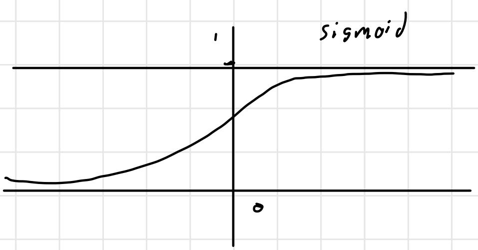
$$\Rightarrow s = \log\left(\frac{1}{1-p}\right) = \text{logit}(p)$$

- objective function:

$$\text{log-likelihood}(\beta) = ys - \log(1 + e^s)$$

$$\text{Log-likli}(\beta) = \sum_{i=1}^{n} y_i s_i - \log(1 + e^{s_i})$$

$$s_i = x_i^T \beta$$



Note derivatives of the tails of sigmoid approach 0.
Derivative is largest near 0

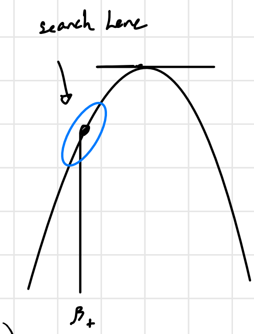
$$\sigma'(s) = \frac{d}{ds}\left(\frac{e^s}{1+e^s}\right)$$

$$= \frac{d}{ds}\left(1 - \frac{1}{1+e^s}\right)$$

$$= \frac{e^s}{(1+e^s)^2} = \frac{e^s}{1+e^s} \cdot \frac{1}{1+e^s}$$

$$\sigma'(s) = p\,(1-p)$$

- Iterated Reweighted Least Squares.



$$\beta_0 \to \beta_1 \to \cdots \to \beta_t \to \beta_{t+1} \to \cdots$$

$$\ell(s_i) = y_i s_i - \log(1+e^{s_i})$$

current $\beta_t \longrightarrow \hat{s}_i = x_i^T \beta_t \longrightarrow \hat{p}_i = \sigma(\hat{s}_i)$

$$\hat{w}_i = \hat{p}_i(1-\hat{p}_i)$$

$$\beta = \beta_t + \Delta\beta$$

$$s_i = x_i^T\beta = x_i^T(\beta_i + \Delta\beta)$$

$$= \underbrace{x_i^T\beta_t}_{\hat{s}_i} + \underbrace{x_i^T\Delta\beta}_{\Delta s_i}$$

$$\ell(s_i) = \ell(\hat{s}_i + \Delta s_i)$$

$$\approx \ell(\hat{s}_i) + \ell'(\hat{s}_i)\Delta s_i + \frac{1}{2}\ell''(\hat{s}_i)\Delta s_i^2$$

$$= \text{const} + \underline{\hat{e}_i}\,\Delta s_i - \frac{1}{2}\underline{\hat{w}_i}\,\Delta s_i^2$$

$$\underline{\ell'(s_i)} = y_i - \frac{e^{s_i}}{1+e^{s_i}} = y_i - p_i = e_i$$

$\sigma(s_i)$

$$\underline{\ell''(s_i)} = -p_i(1-p_i) = -w_i$$

$$= -\frac{1}{2}\hat{w}_i\left(\Delta s_i^2 - 2\frac{\hat{e}_i}{\hat{w}_i}\Delta s_i\right) + \text{const}$$

$$= -\frac{1}{2}\hat{w}_i\left(\underbrace{\Delta s_i}_{x_i^T\Delta\beta} - \underset{\hat{y}_i}{\boxed{\frac{\hat{e}_i}{\hat{w}_i}}}\right)^2 + \text{const}$$

$$\text{Loss}(\Delta\beta) = \frac{1}{2}\sum_{i=1}^{n} w_i\left(\hat{y}_i - x_i^T\Delta\beta\right)^2$$

Therefore $\Delta\beta_t = \left(\sum_{i=1}^{n}\hat{w}_i x_i x_i^T\right)^{-1}\left(\sum_{i=1}^{n}\hat{w}_i\hat{y}_i\right)$

Thus $\beta_{t+1} = \beta_t + \Delta\beta_t$

until $|\Delta\beta_t| < \varepsilon$