Lecture 2

- Part 1 Regression

  - Supervised learning

	1	2	⋯ j ⋯	p	1	2	⋯ j ⋯	d
1								
2								
⋮								
i			x_i^T				y_i^T	
⋮								
n								

Training Data

- Notation :
  - Specific example :

(x_i, y_i)

$$x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{ip} \end{pmatrix} \qquad y_i = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{id} \end{pmatrix}$$

  - generic example

(x, y) drop subscript i

$$x \begin{pmatrix} x_1 \\ \vdots \\ x_j \\ \vdots \\ x_p \end{pmatrix} \qquad y = \begin{pmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_d \end{pmatrix}$$

- Regression (general)
  - Regression (special) : y continuous, e.g. diffusion

    $P(x_t \mid x_{t+\Delta t}, \text{text prompt})$

    $x_t \downarrow$ y: image; $x_{t+\Delta t} \downarrow$ x

  - Classification : y categorical, e.g. chat - GPT

    $P(x_t \mid x_{<t})$

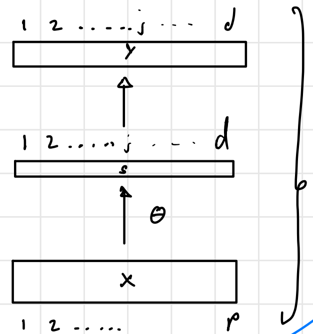
    $x_t \downarrow$ y; $x_{<t} \downarrow$ x

    $\in$ { 50k categories / tokens }

- Gauss Paradigm:
  1. probabilistic formulation
  2. objective function / loss
  3. learning algorithm
  4. experiments, theory

- Unification:

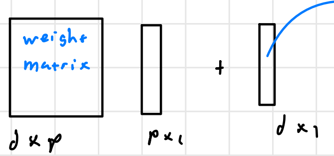
Start with $P(y \mid x)$



$s = f_\theta(x)$

e.g. a linear model

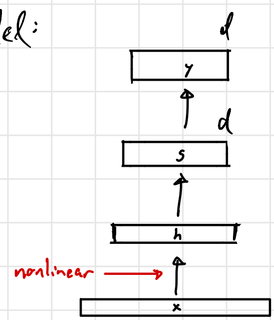
$f_\theta(x) = wx + b$



bias vector

$\theta = (w, b)$

Another model:



hidden variables / nonlinear features

$f_\theta(x) = wh + b$

general formulation

- Regression

1. $p_\theta(y|x) \sim \mathcal{N}(s, \sigma^2 I)$

i.e. $y_j \sim \mathcal{N}(s_j, \sigma)$, $j = 1, \dots, d$

$$p_\theta(y|x) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(y_j - s_j)^2}{2\sigma^2}} \quad \text{by independence}$$

$$= \frac{1}{(\sqrt{2\pi\sigma^2})^d} \exp\left(-\frac{1}{2\sigma^2} |y - s|^2\right).$$

2. Log-Likelihood:

$\ell(\theta) = \log p_\theta(y|x)$ objective for single example

$$LL(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(y_i | x_i)$$

$$= \mathbb{E}_{data}(\log p_\theta(y|x))$$

data =
- data distribution: $p_{data}(x, y)$

$(x_i, y_i) \sim p_{data}(x, y)$, $i = 1, \dots, n$, indep

- empirical data distribution

$$\hat{p}_{data} = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i, y_i} \sim \text{Unif}\{(x_i, y_i), i = 1 \cdots n\}$$

- $\ell(\theta) = \log P_\theta(y \mid x)$

$$= -\frac{1}{2\sigma^2} |y - s|^2 + \text{const}$$

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

- Notice:

$$\max_\theta L(\theta) = \min_\theta \underbrace{\sum_{i=1}^{n} |y_i - f_\theta(x_i)|^2}_{\text{Least - Squares Loss}}$$

- $\text{Loss}(\theta) = \frac{1}{2}\sum_{i=1}^{n} |y_i - s_i|^2$, $s_i = f_\theta(x_i)$, e.g $f_\theta(x_i) = w x_i + b$

- Generic Example:

$$\text{loss}(\theta) = \underbrace{|y - s|^2}_{\text{drop } i} = \frac{1}{2}\sum_{j=1}^{d} (y_j - s_j)^2$$

3. $\min_{\theta}$ Loss $L(\theta)$ by gradient-based method

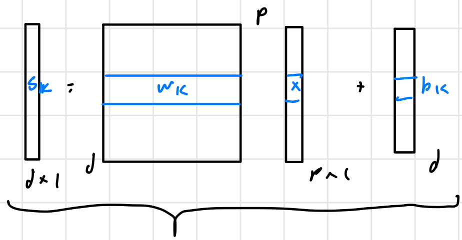
$$\frac{\partial \, \text{loss}L(\theta)}{\partial s_k} = -(y_k - s_k) = \underset{\text{(error)}}{-e_k}$$

$$\frac{\partial l}{\partial s} = \begin{array}{c} \begin{pmatrix} \vdots \\ \frac{\partial l}{\partial s_k} \\ \vdots \end{pmatrix} \end{array} \begin{array}{c} 1 \\ k \\ d \end{array} = \underbrace{\begin{pmatrix} \vdots \\ -e_k \\ \vdots \end{pmatrix}}_{e} \begin{array}{c} 1 \\ k \\ d \end{array} = -\underbrace{\begin{pmatrix} \vdots \\ y_k - s_k \\ \vdots \end{pmatrix}}_{y - s} \begin{array}{c} 1 \\ k \\ d \end{array}$$

$$\frac{\partial l}{\partial s} = -(y - s) = -e$$

- now we want the derivative with respect to θ.

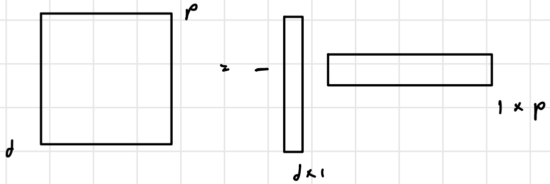
$s = w \, x + b$



$$s_k = w_k x + b_k$$

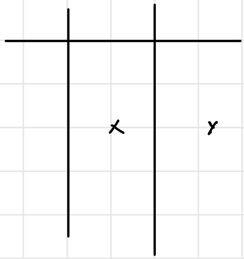
$$\frac{\partial l}{\partial w_k} = \frac{\partial l}{\partial s_k} \frac{\partial s_k}{\partial w_k} \rightarrow \begin{pmatrix} \frac{\partial l}{\partial w_k} \end{pmatrix} \begin{array}{c} 1 \\ k \\ d \end{array} = -\begin{pmatrix} e_k \end{pmatrix} \begin{array}{c} 1 \\ k \\ d \end{array} x^T$$

$$= -e_k x^T$$



$$\frac{\partial l}{\partial b} = -e$$

- Classification :



individual categories

$$Y \in \{1, 2, \ldots, C = 50k\}$$

equivalent notation

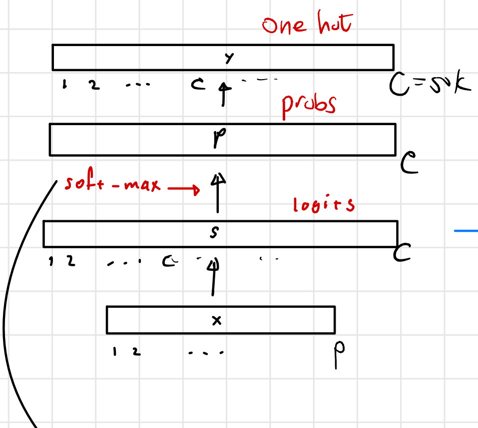
$$Y = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} 1 \\ 2 \\ \vdots \\ \\ c \\ \vdots \\ \\ C = 50k \end{matrix}$$

one - hot vector

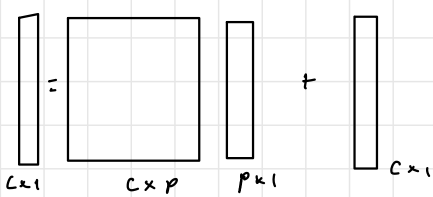
1. $p(y \mid x)$

$$p(y = c \mid x) = p_c \qquad \sum_{i=1}^{C=50k} p_c = 1$$

$$p\left(y = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} 1 \\ \vdots \\ \\ c \\ \\ \vdots \\ C=50k \end{matrix} \;\middle|\; x \right) = p_c \qquad p = \begin{pmatrix} p_1 \\ \vdots \\ p_C \end{pmatrix}$$



$$s = wx + b$$



Softmax :

$$p_k = \frac{e^{s_k}}{\sum_{c=1}^{C=50k} e^{s_c}} = \frac{e^{s_k}}{Z}$$

or

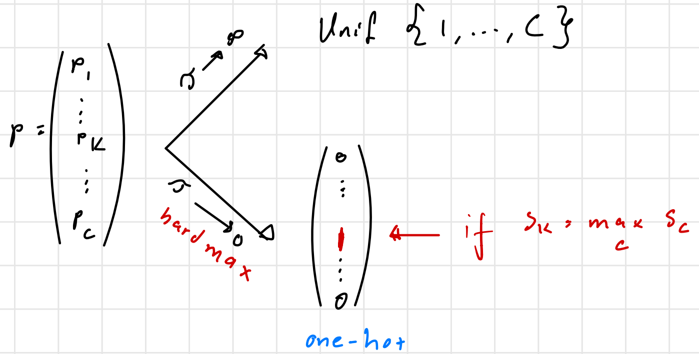
$$s = wh + b$$

$\uparrow$ non linear

x

- Introduce temperature $\tau > 0$

$$p_k = \frac{e^{s_k/\tau}}{\sum_c e^{s_c/\tau}}$$



- note one degree of redundancy: (not an issue if $C >> 2$)

$$p_k = \frac{e^{s_k}}{\sum_c e^{s_c}} = \frac{e^{s_k + \text{const}}}{\sum_c e^{s_c + \text{const}}}$$

2. If $y = k$, $p(y|x) = \frac{e^{s_k}}{Z} = \frac{e^{\langle y, s\rangle}}{Z}$

$$\Downarrow \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \begin{matrix} 1 \\ \vdots \\ k \\ \vdots \\ C \end{matrix} \quad \text{one-hot} \qquad s = \begin{pmatrix} s_1 \\ \vdots \\ s_k \\ \vdots \\ s_C \end{pmatrix} \quad \text{logits}$$

$$\begin{aligned} \ell(\theta) &= \log p(y|x) \\ &= s_k - \log Z \\ &= \langle y, s\rangle - \log Z \end{aligned}$$

3. $$\text{loss}(\theta) = -\log p_\theta(y|x)$$

$$= -(s_K - \log Z)$$

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

$$\frac{\partial\, \text{loss}(\theta)}{\partial s_K} = -\left(y_K - \frac{\partial}{\partial s_K}\log Z\right) \qquad \langle y, s\rangle = \sum_{c=1}^{C} y_c s_c$$

$$= -\left(y_K - \frac{1}{Z}\frac{\partial Z}{\partial s_K}\right), \qquad Z = \sum_{c=1}^{C} e^{s_c}$$

$$= -\left(y_K - \frac{1}{Z} e^{s_K}\right)$$

$$= -(\underbrace{y_K - p_K}_{e_K})$$

$$\frac{\partial\, \text{loss}(\theta)}{\partial s} = -(\underbrace{y - p}_{e})$$

cross-entropy loss $= E_{data}[-\log p_\theta(y|x)]$

cross-entropy of q relative to p $= H(p,q) = -E_p[\log q]$

- Case $C = 2$

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \begin{matrix} + \\ - \end{matrix}, \quad p = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} \begin{matrix} + \\ - \end{matrix}, \quad s = \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} \begin{matrix} + \\ - \end{matrix}$$

$$p_1 = \frac{e^{s_1}}{e^{s_1} + e^{s_2}}, \quad p_2 = \frac{e^{s_2}}{e^{s_1} + e^{s_2}}$$

- Now we fix $s_2 = 0$, so $p_1 = \dfrac{e^{s_1}}{1 + e^{s_1}} = \text{sigmoid}(s_1)$
(due to redundancy)

$$p_2 = 1 - p_1 = \frac{1}{1 + e^{s_1}}$$

$$p(y_1 = 1 \mid x) = p_1 = \frac{e^{s_1}}{1 + e^{s_1}} = \text{sigmoid}(s_1)$$ Special case of softmax

$$p(y \mid x) = \frac{e^{\langle y, s \rangle}}{Z} = \frac{e^{y_1 s_1 + y_2 s_2}}{Z} = \frac{e^{y_1 s_1}}{Z}$$

$$\log p(y_1 \mid x) = y_1 s_1 - \log(1 + e^{s_1})$$

$$y_1 \in \{1, 0\}$$

enough to indicate the two categories