* Probabilist interpretation of linear Regression

MLE : For IID data from some distribution $p(x|\theta_0)$ the MLE minimizes the KL divergence (KULLBACK - LEIBLER Divergence)

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{m} p(x^{(i)}|\theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{m} \log p(x^{(i)}|\theta)$$

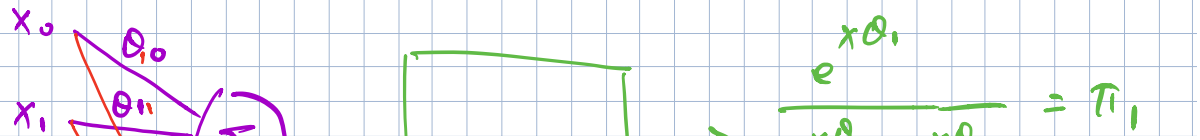
$$= \arg\max_{\theta} \frac{1}{m} \sum_{i=1}^{m} \log p(x^{(i)}|\theta) - \frac{1}{m} \sum_{i=1}^{m} \log p(x^{(i)}|\theta_0)$$

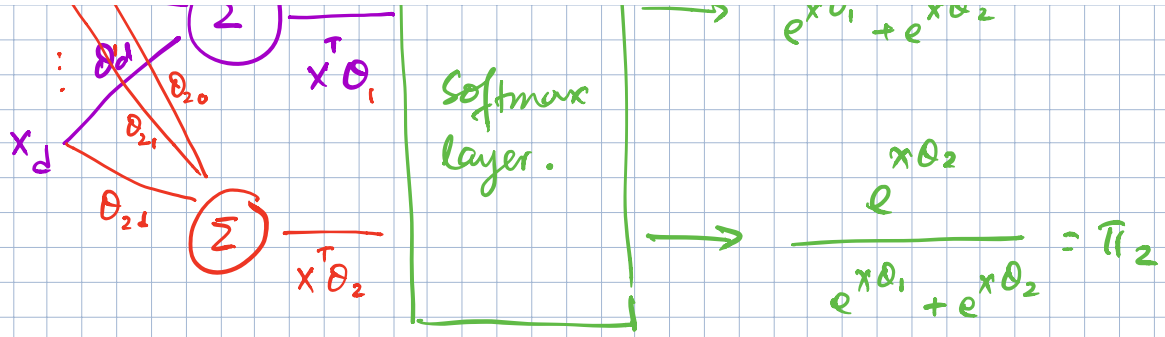
$$= \arg\max_{\theta} \frac{1}{m} \sum_{i=1}^{m} \log p(x^{(i)}|\theta) \Big/ p(x^{(i)};\theta_0)$$

$$= \arg\min_{\theta} \frac{1}{m} \sum_{i=1}^{m} \log p(x^{(i)}|\theta_0) \Big/ p(x^{(i)}|\theta)$$

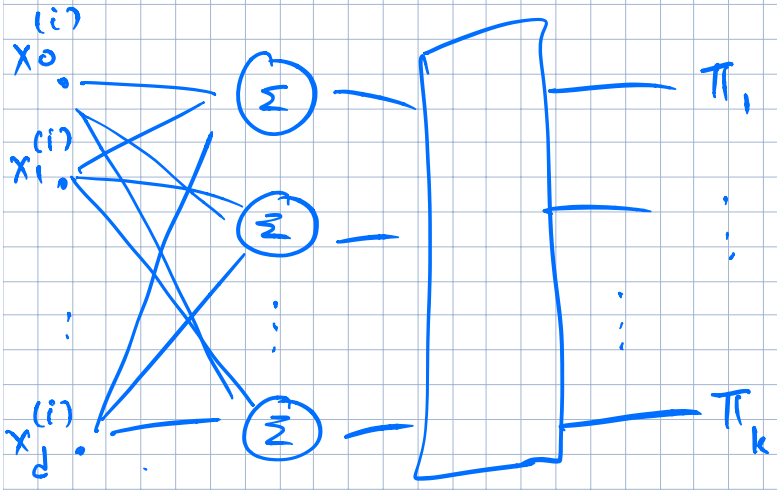
$$\xrightarrow{m\to\infty} \arg\min_{\theta} \int \log \frac{p(x|\theta_0)}{p(x|\theta)} p(x|\theta_0)\, dx$$

Softmax Formulation





$\theta_1 = \begin{matrix}\theta_{10}\\ \theta_{11}\\ \vdots \\ \theta_{1d}\end{matrix}$ $\theta_2 = \begin{matrix}\theta_{20}\\ \theta_{21}\\ \vdots \\ \theta_{2d}\end{matrix}$ $\pi_1 + \pi_2 = 1$



Likelihood

$$P(y \mid x; \theta) = \prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)}; \theta)$$

$$= \prod_{i=1}^{n} \pi_{i1}^{\mathbb{1}_0(y^{(i)})} \pi_{i2}^{\mathbb{1}_1(y^{(i)})}$$

$\pi_{i1} = P(y=0 \mid x; \theta)$

$\pi_{i2} = P(y=1 \mid x; \theta)$

$\mathbb{1}_0(y^{(i)}) = 1 \quad$ if $\quad y^{(i)} = 0$

$\mathbb{1}_1(y^{(i)}) = 1 \quad$ if $\quad y^{(i)} = 1$

$$\pi_{i1} = \frac{e^{x^T\theta_1}}{e^{x^T\theta_1} + e^{x^T\theta_2}}$$

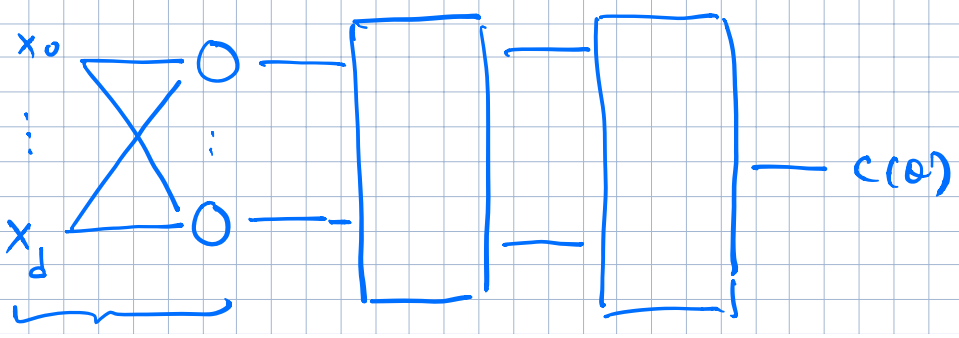
$$\pi_{i2} = \frac{e^{x^T\theta_2}}{e^{x^T\theta_1} + e^{x^T\theta_2}}$$

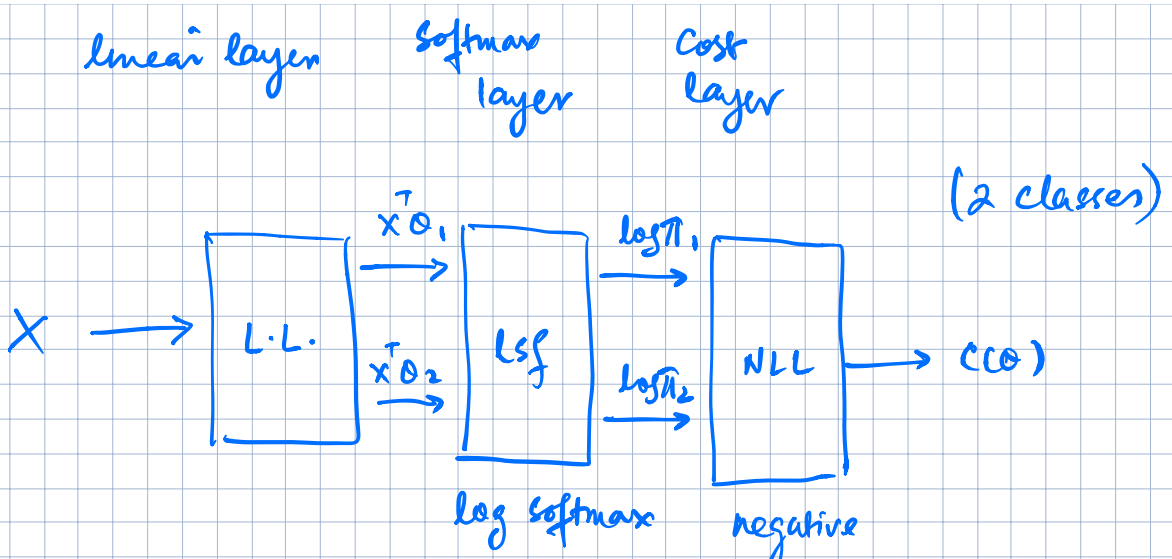
I am interested in MLE.

$$C(\theta) = -\log \text{likelihood}$$

$$= -\sum_{i=1}^{n} \left(\mathbb{1}_0(y^{(i)}) \log \pi_{i1} + \underbrace{\mathbb{1}_1(y^{(i)})}_{\sum_{i=1}^{k}} \log \pi_{i2} \right)$$

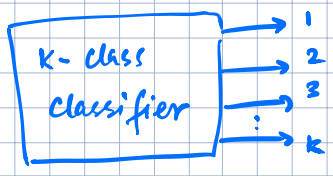
Layered view of Softmax





Cross-Entropy

Problem: How do we compare two vectors?



$\hat{y}$ (predicted)

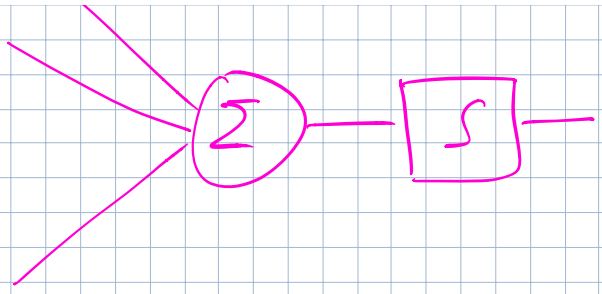
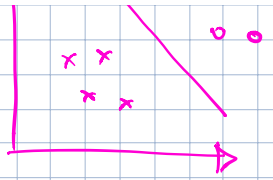
x_1	x_2	labels	
3	7	3	0 0 1 0 0 0 0
18	47	1	1 0 0 0 0 0 0
2	4	4	0 0 0 1 0 0 0
42	1	7	0 0 0 0 0 0 1

y (targets)

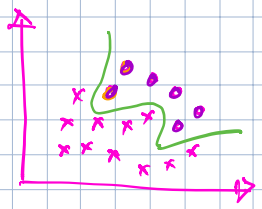
Compare $\hat{y}$ and y using cross-entropy.

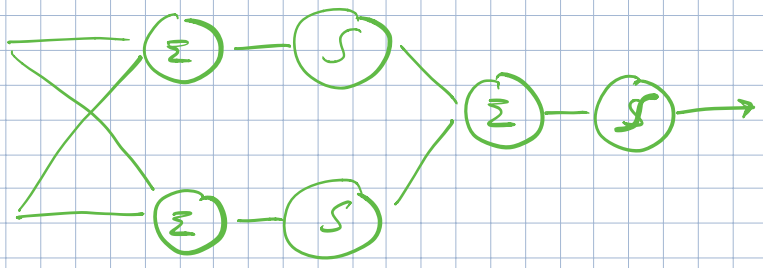
$$D(\hat{y}, y) = -\sum_{i=1}^{7} y_i \log \hat{y}_i$$

$$D(\hat{y}, y) \neq D(y, \hat{y})$$

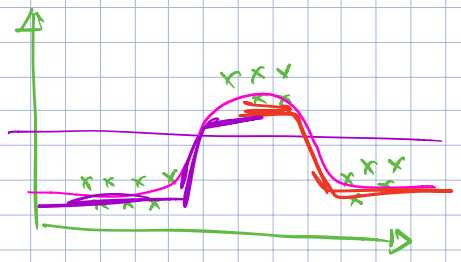



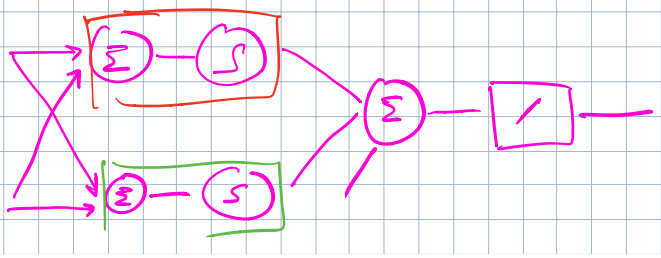
How do I classify the following:





The more neurons the more complex (wobbly) the discriminant function.

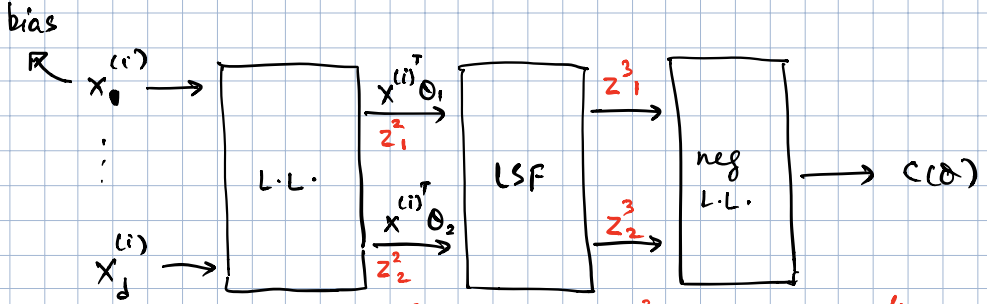


Backpropagation

What information does a layer need to know?

- How to compute output from its input
- How to propagate derivatives backwards
- How to compute derivatives w.r.t. learning parameters.



$z^1 = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$

$z_1^2 = x^T\theta_1$

$z_2^2 = x^T\theta_2$

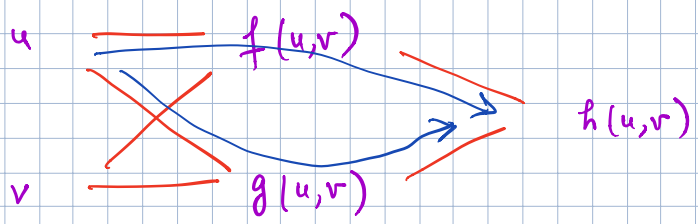
$z_1^3 = \log \frac{e^{z_1^2}}{e^{z_1^2} + e^{z_2^2}}$, $z_2^3 = \log \frac{e^{z_2^2}}{e^{z_1^2} + e^{z_2^2}}$

$z^4 = -\sum_{i=1}^{n} \mathbb{1}_0(y_i) z_1^3 + \mathbb{1}_1(y_i) z_2^3$

$C(\theta) = z^4\left[z_1^3\{z_1^2(\theta_1, z^1), z_2^2(\theta_2, z^1)\}, z_2^3\{z_1^2(\theta_1, z^1), z_2^2(\theta_2, z^1)\}\right]$

$\frac{\partial C(\theta)}{\partial \theta} = ?$

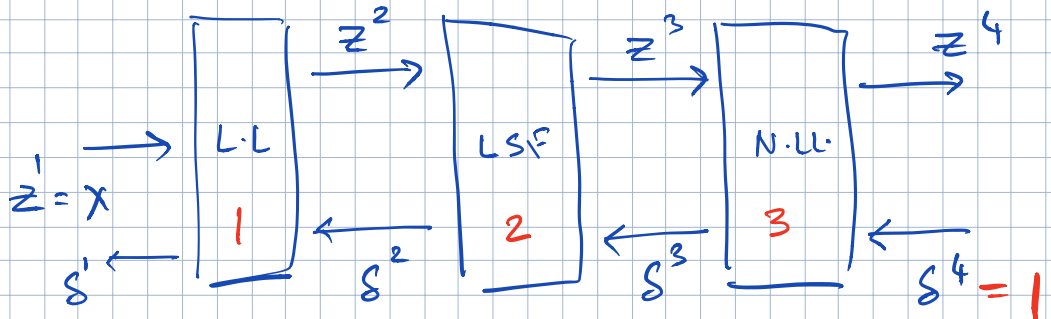
Chain Rule:

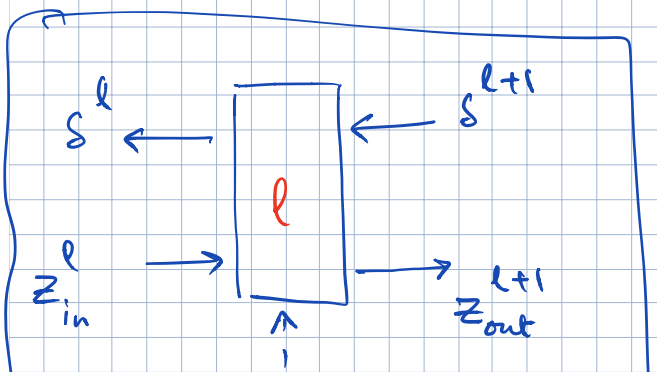


$$h(u,v) = \{ f(u,v),\ g(u,v) \}$$

$$\frac{\partial h}{\partial u} = \frac{\partial h}{\partial f}\frac{\partial f}{\partial u} + \frac{\partial h}{\partial g}\frac{\partial g}{\partial u}$$

$$\frac{\partial z^4}{\partial \theta_1} = \frac{\partial z^4}{\partial z_1^3}\frac{\partial z_1^3}{\partial \theta_1} + \frac{\partial z^4}{\partial z_2^3}\frac{\partial z_2^3}{\partial \theta_1}$$





$$\frac{\partial C}{\partial \theta^l}$$

$$\delta_i^l = \frac{\partial C}{\partial z_i^l} = \sum_j \frac{\partial C}{\partial z_j^{l+1}} \frac{\partial z_j^{l+1}}{\partial z_i^l}$$

$$= \sum_j \delta_j^{l+1} \frac{\partial z_j^{l+1}}{\partial z_i^l}$$

$$\delta_i^l = \sum_j \delta_j^{l+1} \frac{\partial z_j^{l+1}}{\partial z_i^l}$$

$$z^{l+1} = f^l(z^l)$$

$$\frac{\partial C}{\partial \theta^l} = \sum_j \frac{\partial C}{\partial z_j^{l+1}} \frac{\partial z_j^{l+1}}{\partial \theta}$$

↑ ?

$$= \sum_j \delta_j^{l+1} \frac{\partial z_j^{l+1}}{\partial \theta}$$

$$\frac{1}{1+e^{z'}}$$

$$= \sigma(z)(1-\sigma(z))$$