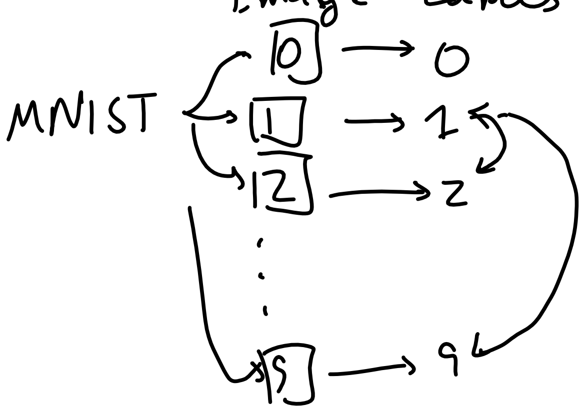
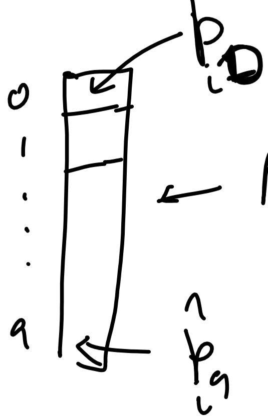
1. Cross Entropy loss : Multi class - classification
2. Weight Decay as Regularizer



label 1 is no more closer to 2 than 9

$\hat{y}_i \cancel{=} MLP(\underline{x}_i)$

$\|\hat{y}_i \cancel{-} y_i\|_2^2$ X



$MLP(x_i)$

$\hat{\underline{p}}_i = MLP(\underline{x}_i)$

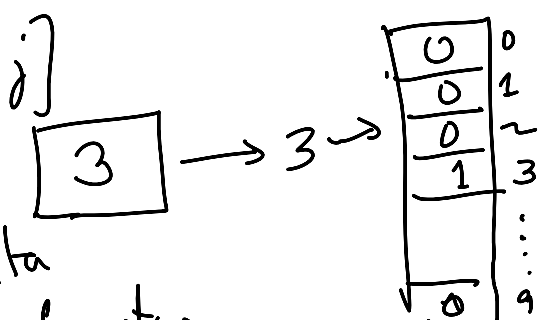
How can we ensure that $\sum_{j=0}^{9} \hat{p}_{ij} = 1$?

$0 \leq \hat{p}_{ij} \leq 1$?

True probabilities $p_{ij} = \mathbb{1}[y_i = j]$

$$p_{ij} = \begin{cases} 1 & \text{if } y_i = j \\ 0 & \text{otherwise} \end{cases}$$

dirac delta
indicator function



Cross entropy loss

$$l(p_{ij}, \hat{p}_{ij}) = \sum_{j=0}^{9} - p_{ij} \log_e \hat{p}_{ij}$$

$$= -\log_e \hat{p}_{i[y_i]}$$

$\hat{p}_{ij} < 1$

$\Rightarrow \log \hat{p}_{ij} < 0$

$\hat{p}_{i[y_i]} = 1$

$\Rightarrow \log \hat{p}_{i[y_i]} = 0$

Entropy

$$H[X] = \sum_{x \in \Omega} - P[X = x] \log_e (P[X = x])$$

$$= \int_{x \in \Omega} - f(x) \log_e f(x) \, dx$$

$$= \mathbb{E}_X [\log_e (P[X])]$$

How can we ensure that $\sum_{j=0}^{9} \hat{p}_{ij} = 1$?

$0 \leq \hat{p}_{ij} \leq 1$?

Softmax layer (temperature)

$$s(\underline{y}) = \frac{\exp(\underline{y})}{\sum_{i=1}^{n} \exp(y_i)}$$

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

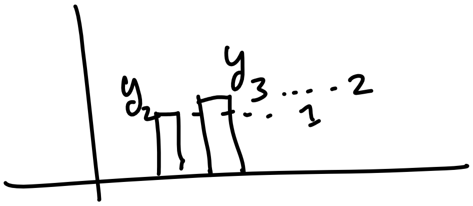
$$= \frac{\exp(\underline{y})}{\mathbf{1}^T \exp(\underline{y})}$$

$-\infty < y_i < +\infty \xrightarrow{+ve} 0 < \exp(y_i) < \infty$

$\downarrow$ normalize

$0 < \exp(y_i) / \sum \exp(y_i) < 1$

Why name softmax?



$\exp(1) < \exp(2)$

$1 < 2$

$e^1 < e^2$ ← 2.71 times

$$\underbrace{\sum \frac{\exp(y_i)}{\sum \exp(y_i)}}_{\hat{p}_i} = 1$$

$y_2 = 10$
$y_3 = 20$

$e^{10} < e^{20}$ ← e^{10} times

softmax exaggatenates differences so much that the highest item goes close to 1 while others go close to zero.

Increasing temperature

$$S(y_i) = \frac{\exp(T y_i)}{\sum \exp(T y_i)}$$

$$\underbrace{\text{loss} \quad \text{cross entropy} \leftarrow \text{softmax}}\leftarrow \text{MLP}$$

Cross entropy with softmax $= \sum_{j=0}^{9} p_{ij} \log\left[\frac{\exp(\hat{y}_{ij})}{\sum_{k=0}^{9} \exp(\hat{y}_{ik})}\right] = \sum_{j=0}^{9} p_{ij}\, \hat{y}_{ij} - \sum_{j=0}^{9} p_{ij} \log\left(\sum_{k=0}^{9} \exp(\hat{y}_{ik})\right)$

$$= \sum_{j=0}^{9} p_{ij}\, \hat{y}_{ij} - \log\left(\sum_{k=0}^{9} \exp(\hat{y}_{ik})\right) \underbrace{\sum_{j=0}^{9} p_{ij}}_{1}$$

minimize $\sum$ loss
Dataset

loss
- → least square loss for regression
- → Hinge loss for two class classification
- → Cross entropy loss for multi-class classification

for m classes

m-1

$$\text{cross entropy loss with softmax}(p_{ij}, \hat{y}_{ij}) = \sum_{j=0}^{9} p_{ij}\, \hat{y}_{ij} - \log\left(\sum_{k=0}^{9} \exp(\hat{y}_{ik})\right)$$

logits (log+bits)

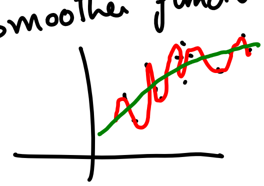
$\log(\hat{p}_{ij})$

Weight Decay

$= L_2$-regularization

Prob. Prior
Occam razor
smoother functions

$$\underset{w}{\text{minimize}} \underbrace{\overbrace{\sum_{\text{Data}} l(y_i, \hat{y}_i)}^{\text{Data loss}} + \overbrace{\lambda \|w\|_2^2}}$$



$1 > (1 - 2\alpha^t \lambda) > 0$

weight decay parameter ≈ 0.1

$$\underline{w}_{t+1} = \underline{w}_t - \alpha^t \sum_B \nabla_{\underline{w}} l(y_i, \hat{y}_i) \underbrace{- 2\alpha^t \lambda \underline{w}_t}_{\text{weight decay}}$$

$$\underline{w}_{t+1} = \underbrace{(1 - 2\alpha^t \lambda)}\underline{w}_t - \alpha^t \sum_B \nabla_w l(y_i, \hat{y}_i)$$

$$\frac{\partial}{\partial \underline{w}} \underline{w}^T I \underline{w} = 2\underbrace{\underline{w}^T}_{1\times n} \underbrace{\underline{I}}_{n\times n} = 2\underline{w}^T_{1\times n}$$

$\|w\|^2 = \underline{w}^T \underline{w}$

$= \underline{w}^T I \underline{w}$

$$\frac{\partial}{\partial w} \underline{w}^T I \underline{w} = 2\underline{w}^T \underline{I}$$

$= 2\underline{w}^T$

$$\frac{\partial}{\partial \underline{w}} \underline{w}^T A \underline{w} = \underline{w}^T (A + A^T)$$

$= 2\underline{w}^T A$