Why Softmax?

Motivation Need a function. scores $\in [-\infty, \infty] \rightarrow$ probs $\in [0, 1]$

$\Rightarrow$ Softmax is continuous, differentiable, normalizes to 1

$\boxed{\text{But still.. why exponential ??}}$ Goal: Find $\frac{\partial L}{\partial s}$.

$p_k = \text{softmax}\left(\begin{bmatrix} s_1 \\ s_2 \\ \vdots \end{bmatrix}\right) = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \end{bmatrix}$

L (negative log likelihood): $L(y, p) = -\sum_{k=1}^{K} y_k \log p_k$

- If correct class is C: $L = -\log p_c$
- For each component i: $L_i = -y_i \log p_i \Rightarrow \frac{\partial L}{\partial p_i} = -\frac{y_i}{p_i}$ (equal to 0 except correct class C)
- Vector form: $\frac{\partial L}{\partial p} = -y \oslash p$ [$\oslash$ = element-wise div]

To find $\frac{\partial L}{\partial s} = \frac{\partial L}{\partial p} \cdot \left(\frac{\partial p}{\partial s}\right)^T$ = Jacobian!

- For vector-valued function $p = \begin{bmatrix} \sigma(s_1) \\ \sigma(s_2) \\ \vdots \end{bmatrix}$:

$$J = \begin{bmatrix} \frac{\partial p_1}{\partial s_1} & \frac{\partial p_2}{\partial s_2} & \cdots \\ \frac{\partial p_1}{\partial s_1} & \ddots & \\ \vdots & & \end{bmatrix} \Rightarrow \left[C \times C \text{ matrix, } J_{ij} = \frac{\partial p_i}{\partial s_j} \right]$$

how much does probability of class i change if we change the score of class j?

Finding the Jacobian of softmax.

- Case $i = j$: $p_i = \frac{e^{s_i}}{\sum_k e^{s_k}}$

$$\Rightarrow \frac{\partial p_i}{\partial s_i} = \frac{(\sum_k e^{s_k})(e^{s_i}) - (e^{s_i})(e^{s_i})}{(\sum_k e^{s_k})^2} = p_i - p_i^2 = p_i(1 - p_i)$$

- Case $i \neq j$:

$$\Rightarrow \frac{\partial p_i}{\partial s_j} = \frac{(\sum_k e^{s_k})(0) - (e^{s_i})(e^{s_j})}{(\sum_k e^{s_k})^2} = -p_i p_j$$

$$\begin{bmatrix} p_1 - p_1 p_1 & -p_1 p_2 & \cdots \\ -p_2 p_1 & p_2 - p_2 p_2 & \cdots \\ \vdots & & \ddots \end{bmatrix}$$

- In matrix form: $\frac{\partial p}{\partial s} = \text{diag}(p) - pp^T$

$$\text{Computing } \frac{\partial L}{\partial s} = \Big(\overbrace{\frac{\partial p}{\partial s}}^{\text{symmetric so } J^T = J}\Big)^T \frac{\partial L}{\partial p}$$

$$\begin{bmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_1p_2 & p_2(1-p_2) & -p_2p_3 \\ -p_1p_3 & -p_2p_3 & p_3(1-p_3) \end{bmatrix} \begin{bmatrix} -y_1/p_1 \\ -y_2/p_2 \\ -y_3/p_3 \end{bmatrix}$$

$$= -\frac{y_1}{p_1}\begin{bmatrix} p_1(1-p_1) \\ -p_1p_2 \\ -p_1p_3 \end{bmatrix} - \frac{y_2}{p_2}\begin{bmatrix} -p_1p_2 \\ p_2(1-p_2) \\ -p_2p_3 \end{bmatrix} - \frac{y_3}{p_3}\begin{bmatrix} -p_1p_3 \\ -p_2p_3 \\ p_3(1-p_3) \end{bmatrix}$$

$$= \begin{bmatrix} -y_1(1-p_1) + y_2p_1 + y_3p_1 \\ y_1p_2 - y_2(1-p_2) + y_3p_2 \\ y_1p_3 + y_2p_3 - y_3(1-p_3) \end{bmatrix}$$

$$= \begin{bmatrix} y_1p_1 + y_2p_1 + y_3p_1 - y_1 \\ y_2p_2 + y_1p_2 + y_3p_2 - y_2 \\ y_3p_3 + y_1p_3 + y_2p_3 - y_3 \end{bmatrix} \begin{matrix} \rightarrow p_1(y_1+y_2+y_3) - y_1 = p_1 - y_1 \\ \rightarrow p_2(y_1+y_2+y_3) - y_2 = p_2 - y_2 \\ \rightarrow p_3(y_1+y_2+y_3) - y_3 = p_3 - y_3 \end{matrix}$$

$$\therefore \boxed{\frac{\partial L}{\partial s} = p - y}$$