Generalized softmax regression: MLE

13

$$P(\mathbf{y} \mid \mathbf{x}; \boldsymbol{\theta}) = \frac{e^{f(\mathbf{x};\boldsymbol{\theta})}}{\sum_k e^{f^{(k)}(\mathbf{x};\boldsymbol{\theta})}}$$

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} = & \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \; -\mathbb{E}_{\mathbf{x},\mathbf{y}\sim P_D}\left[\log P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta})\right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} - \sum_{i=1}^{K} y_i \log\left[\frac{\exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(i)}\right]}{\sum_j \exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(j)}\right]}\right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} - \left\langle \mathbf{y},\; \log\left[\frac{\exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(i)}\right]}{\sum_j \exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(j)}\right]}\right]\right\rangle \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} - \sum_{i=1}^{K} \delta_{y_i=1} \cdot \log\left[\frac{\exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(i)}\right]}{\sum_j \exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(j)}\right]}\right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \; \sum_{i=1}^{K} \left[\delta_{y_i=1} \log \sum_j \exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(j)}\right] - \delta_{y_i=1} f(\mathbf{x};\boldsymbol{\theta})^{(i)}\right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \; \log \sum_j \exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(j)}\right] - \sum_{i=1}^{K} \delta_{y_i=1} f(\mathbf{x};\boldsymbol{\theta})^{(i)}
\end{aligned}
$$

$$
\begin{aligned}
\nabla_{f^{(k)}} RHS &= \frac{\exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(k)}\right]}{\sum_j \exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(j)}\right]} - \delta_{y_k=1} \\
&= softmax(f(\mathbf{x};\boldsymbol{\theta})^{(k)}) - \delta_{y_k=1}
\end{aligned}
$$

Softmax function:

MLE:

Gradients:

The gradients w.r.t. $f(\cdot)$ get propogated to the parameters, $\boldsymbol{\theta}$, via the chain rule. This gradient is valid for any distribution in the exponential family, not just the softmax.

Generalized softmax regression: MLE

Softmax function: $P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta}) = \dfrac{e^{f(\mathbf{x};\boldsymbol{\theta})}}{\sum_k e^{f^{(k)}(\mathbf{x};\boldsymbol{\theta})}}$

MLE: $$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \; -\mathbb{E}_{\mathbf{x},\mathbf{y}\sim P_D}\left[\log P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta})\right]$$

$$= \underset{\boldsymbol{\theta}}{\text{argmin}} - \sum_{i=1}^{K} y_i \log\left[\frac{\exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(i)}\right]}{\sum_j \exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(j)}\right]}\right]$$

$$= \underset{\boldsymbol{\theta}}{\text{argmin}} - \left\langle \mathbf{y},\; \log\left[\frac{\exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(i)}\right]}{\sum_j \exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(j)}\right]}\right]\right\rangle$$

$$= \underset{\boldsymbol{\theta}}{\text{argmin}} - \sum_{i=1}^{K} \delta_{y_i=1} \cdot \log\left[\frac{\exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(i)}\right]}{\sum_j \exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(j)}\right]}\right]$$

$$= \underset{\boldsymbol{\theta}}{\text{argmin}} \; \sum_{i=1}^{K} \left[\delta_{y_i=1} \log \sum_j \exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(j)}\right] - \delta_{y_i=1} f(\mathbf{x};\boldsymbol{\theta})^{(i)}\right]$$

$$= \underset{\boldsymbol{\theta}}{\text{argmin}} \; \log \sum_j \exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(j)}\right] - \sum_{i=1}^{K} \delta_{y_i=1} f(\mathbf{x};\boldsymbol{\theta})^{(i)}$$

The gradients w.r.t. $f(\cdot)$ get propogated to the parameters, $\boldsymbol{\theta}$, via the chain rule. This gradient is valid for any distribution in the exponential family, not just the softmax.

Gradients: $$\nabla_{f^{(k)}} RHS = \frac{\exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(k)}\right]}{\sum_j \exp\left[f(\mathbf{x};\boldsymbol{\theta})^{(j)}\right]} - \delta_{y_k=1}$$

$$= \mathit{softmax}(f(\mathbf{x};\boldsymbol{\theta})^{(k)}) - \delta_{y_k=1}$$

Gradient Descent for softmax regression

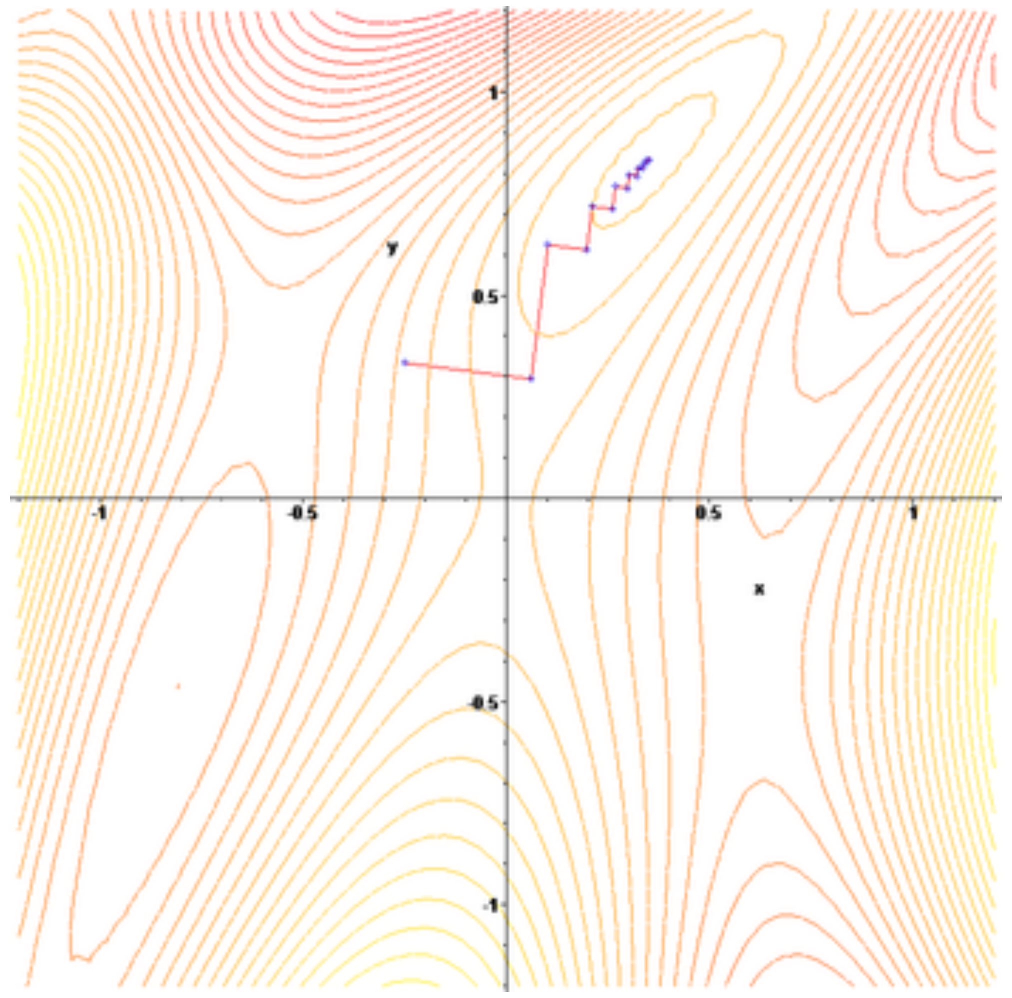
mapping: $f : \mathbf{x} \rightarrow \mathbf{y} = \mathbf{W}\mathbf{x} + \mathbf{b}$

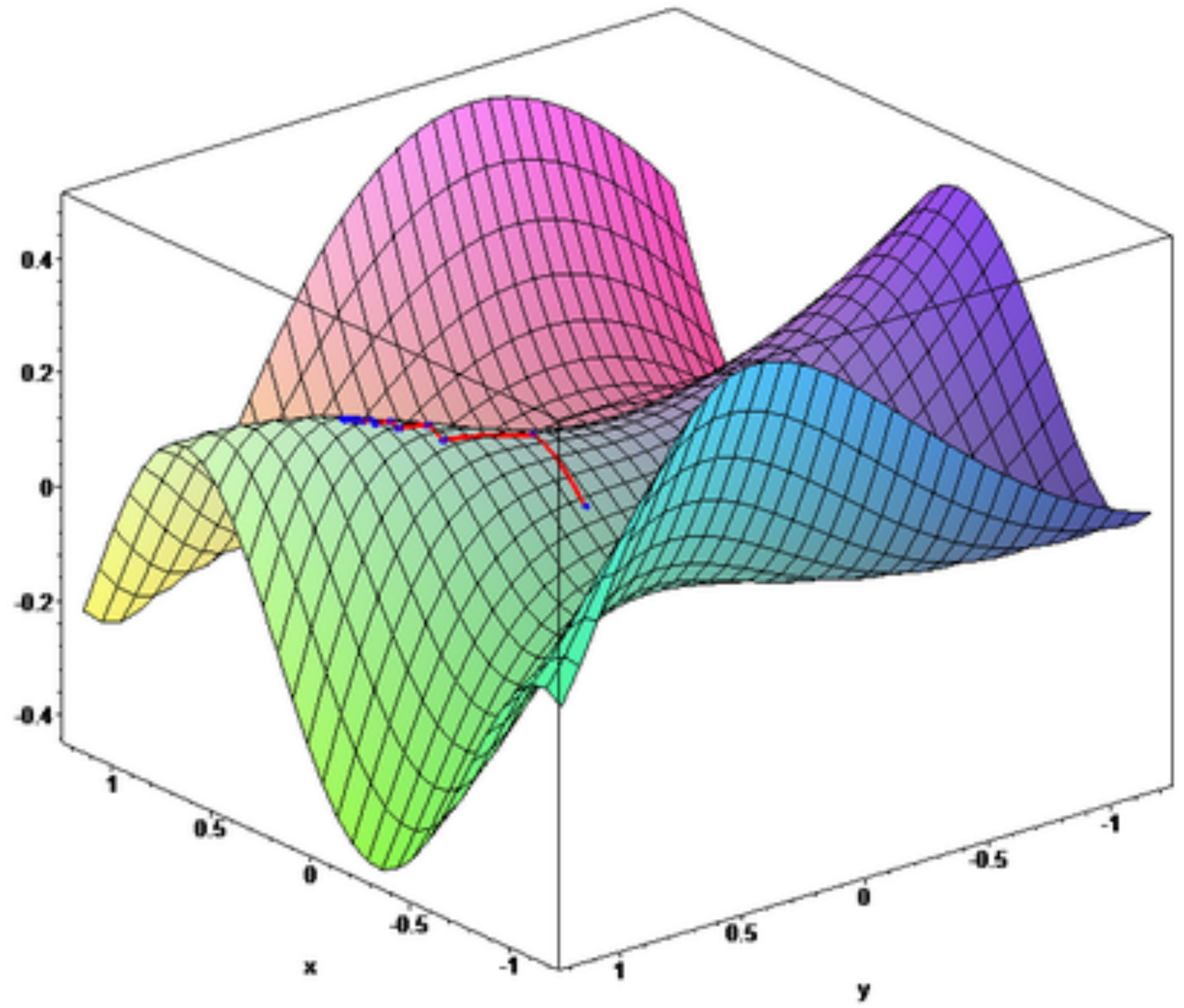
$$\begin{aligned} where \quad & \mathbf{x} \in \mathbb{R}^N \\ & \mathbf{y}, \mathbf{b}, f(\cdot) \in \mathbb{R}^K \\ & \mathbf{x} \in \mathbb{R}^N \\ & \mathbf{W} \in \mathbb{R}^{K \times N} \end{aligned}$$

gradients: $\nabla_{W_{k,i}} \mathbf{L}_{CE} = x_i \left(\frac{\exp[\mathbf{w}_k \mathbf{x} + b_k]}{\sum_j \exp[\mathbf{w}_j \mathbf{x} + b_j]} - \delta_{1,y_k} \right)$

gradient update rule:

$$\begin{aligned} & while\ not\ converged\ do: \ \{ \\ & \qquad \forall k, j \text{ do}: \quad \{ \\ & \qquad\qquad W_{k,j} := W_{k,j} - \eta \nabla_{W_{k,j}} L_{CE} \\ & \qquad \} \\ & \} \end{aligned}$$





- Source: Wikipedia