Generalized softmax regression

12

- Generic softmax regression adopts the underlying model mapping y <- x in logistic regression, but extends it to >2 dimensions

- The softmax function is more broadly useful for other (not necessarily linear) mapping functions, too! It comes in very handy for any function that’s at least once differentiable w.r.t. parameters that are being learned. Here we’ll just say there is some function that’s differentiable with respect to some parameters, . Note, below represents a vector of probabilities, one for each class.

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

Softmax function:

$$f : \mathbf{x} \rightarrow \mathbf{y}$$

θ

Generalized softmax regression

- Generic softmax regression adopts the underlying model mapping y <- x in logistic regression, but extends it to >2 dimensions

- The softmax function is more broadly useful for other (not necessarily linear) mapping functions, too! It comes in very handy for any function that's at least once differentiable w.r.t. parameters that are being learned. Here we'll just say there is some function $f:\mathbf{x}\rightarrow\mathbf{y}$ that's differentiable with respect to some parameters, $\boldsymbol{\theta}$. Note, below represents a vector of probabilities, one for each class.

Softmax function: $$P(\mathbf{y}\,|\,\mathbf{x};\boldsymbol{\theta})=\frac{e^{f(\mathbf{x};\boldsymbol{\theta})}}{\sum_i e^{f^{(i)}(\mathbf{x};\boldsymbol{\theta})}}$$

Generalized softmax regression: MLE

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

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

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

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

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

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

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

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

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