Artificial Intelligence and Machine Learning

Derivation of the Gradient for Logistic Regression

We compute the partial derivative of the binary cross-entropy cost function:

$$\frac{\partial}{\partial \theta_m} \ell^{\text{BCE}}(h_{\boldsymbol{\theta}}(\boldsymbol{x}), y) = \frac{\partial}{\partial \theta_m}\Big(-y \cdot \log(h_{\boldsymbol{\theta}}(\boldsymbol{x})) - (1-y) \cdot \log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}))\Big)$$

$$= -y \cdot \frac{\partial}{\partial \theta_m}\Big(\log(h_{\boldsymbol{\theta}}(\boldsymbol{x}))\Big) - (1-y) \cdot \frac{\partial}{\partial \theta_m}\Big(\log(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}))\Big)$$

[Use the derivative of $\log$ **function:** $(\log x)' = \frac{1}{x}$**]**

$$= \frac{-y}{h_{\boldsymbol{\theta}}(\boldsymbol{x})} \cdot \frac{\partial}{\partial \theta_m}(h_{\boldsymbol{\theta}}(\boldsymbol{x})) - \frac{1-y}{1 - h_{\boldsymbol{\theta}}(\boldsymbol{x})} \cdot \frac{\partial}{\partial \theta_m}(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}))$$

[Factor out the derivative of the model function]

$$= \left(\frac{-y}{h_{\boldsymbol{\theta}}(\boldsymbol{x})} + \frac{1-y}{1 - h_{\boldsymbol{\theta}}(\boldsymbol{x})}\right) \cdot \frac{\partial}{\partial \theta_m} h_{\boldsymbol{\theta}}(\boldsymbol{x})$$

[Find the common denominator]

$$= \frac{-y \cdot (1 - h_{\boldsymbol{\theta}}(\boldsymbol{x})) + (1-y) \cdot h_{\boldsymbol{\theta}}(\boldsymbol{x})}{h_{\boldsymbol{\theta}}(\boldsymbol{x}) \cdot (1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}))} \cdot \frac{\partial}{\partial \theta_m} h_{\boldsymbol{\theta}}(\boldsymbol{x})$$

[Expand the numerator]

$$= \frac{-y + y \cdot h_{\boldsymbol{\theta}}(\boldsymbol{x}) + h_{\boldsymbol{\theta}}(\boldsymbol{x}) - y \cdot h_{\boldsymbol{\theta}}(\boldsymbol{x})}{h_{\boldsymbol{\theta}}(\boldsymbol{x}) \cdot (1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}))} \cdot \frac{\partial}{\partial \theta_m} h_{\boldsymbol{\theta}}(\boldsymbol{x})$$

[Simplify the fraction]

$$= \frac{h_{\boldsymbol{\theta}}(\boldsymbol{x}) - y}{h_{\boldsymbol{\theta}}(\boldsymbol{x}) \cdot (1 - h_{\boldsymbol{\theta}}(\boldsymbol{x}))} \cdot \frac{\partial}{\partial \theta_m} h_{\boldsymbol{\theta}}(\boldsymbol{x})$$

[Use the definition of the model function: $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sigma(\boldsymbol{\theta}^\top \boldsymbol{x})$**]**

$$= \frac{\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}) - y}{\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}) \cdot (1 - \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}))} \cdot \frac{\partial}{\partial \theta_m} \sigma(\boldsymbol{\theta}^\top \boldsymbol{x})$$

[Use the derivative of the sigmoid function and apply the chain rule]

$$= \frac{\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}) - y}{\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}) \cdot (1 - \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}))} \cdot \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}) \cdot (1 - \sigma(\boldsymbol{\theta}^\top \boldsymbol{x})) \cdot \frac{\partial}{\partial \theta_m} \boldsymbol{\theta}^\top \boldsymbol{x}$$

[Cancel redundant terms]

$$= (\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}) - y) \cdot \frac{\partial}{\partial \theta_m} \boldsymbol{\theta}^\top \boldsymbol{x}$$

[Use the derivative of the scalar product]

$$= (\sigma(\boldsymbol{\theta}^\top \boldsymbol{x}) - y) \cdot x_m$$

□