W3WI DS304.1 Applied Machine Learning Fundamentals

Derivation of the Gradient for Logistic Regression

$$\frac{\partial}{\partial \theta_j}\mathcal{J}(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_j}\Big(-y \cdot \log\big(h_{\boldsymbol{\theta}}(\boldsymbol{x})\big) - (1-y) \cdot \log\big(1 - h_{\boldsymbol{\theta}}(\boldsymbol{x})\big)\Big)$$

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

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

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

[Factor out the derivative of the model function]

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

[Find the common denominator]

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

[Expand the numerator]

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

[Simplify the fraction]

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

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

$$= \frac{g(\boldsymbol{\theta}^\intercal \boldsymbol{x}) - y}{g(\boldsymbol{\theta}^\intercal \boldsymbol{x}) \cdot \big(1 - g(\boldsymbol{\theta}^\intercal \boldsymbol{x})\big)} \cdot \frac{\partial}{\partial \theta_j} g(\boldsymbol{\theta}^\intercal \boldsymbol{x})$$

[Use the derivative of the sigmoid function (see equation ??) and apply the chain rule]

$$= \frac{g(\boldsymbol{\theta}^\intercal \boldsymbol{x}) - y}{g(\boldsymbol{\theta}^\intercal \boldsymbol{x}) \cdot \big(1 - g(\boldsymbol{\theta}^\intercal \boldsymbol{x})\big)} \cdot g(\boldsymbol{\theta}^\intercal \boldsymbol{x}) \cdot \big(1 - g(\boldsymbol{\theta}^\intercal \boldsymbol{x})\big) \cdot \frac{\partial}{\partial \theta_j} \boldsymbol{\theta}^\intercal \boldsymbol{x}$$

[Cancel redundant terms]

$$= \big(g(\boldsymbol{\theta}^\intercal \boldsymbol{x}) - y\big) \cdot \frac{\partial}{\partial \theta_j} \boldsymbol{\theta}^\intercal \boldsymbol{x}$$

[Use the derivative of the scalar product]

$$= \big(g(\boldsymbol{\theta}^\intercal \boldsymbol{x}) - y\big) \cdot x_j$$