AI6102 Assignment 1

Sourabh Vyas

16/09/2025

Question 1 (10 marks)

Consider a multi-class classification problem of C classes. Based on the parametric forms of the conditional probabilities of each class introduced on the 39th Page ("Extension to Multiple Classes") of the lecture notes of L4, derive the learning procedure of regularized logistic regression for multi-class classification problems.

Hint: define a loss function by borrowing an idea from binary classification, and derive the gradient descent rules to update $\{\mathbf{w}^{(c)}\}$ for $c = 1, \ldots, C-1$.

Answer

1. Conditional Probabilities We treat class 0 as the baseline and model the others via a "softmax" over negative scores. Then for $c = 0, \ldots, C-1$:

$$P(y = 0 \mid \mathbf{x}) = \frac{1}{1 + \sum_{k=1}^{C-1} \exp\big(-\mathbf{w}^{(k)\top}\mathbf{x}\big)}, \quad P(y = c \mid \mathbf{x}) = \frac{\exp\big(-\mathbf{w}^{(c)\top}\mathbf{x}\big)}{1 + \sum_{k=1}^{C-1} \exp\big(-\mathbf{w}^{(k)\top}\mathbf{x}\big)}.$$

This ensures all C probabilities sum to 1.

2. Multiclass Negative Log Likelihood We assume each example (x_i, y_i) is drawn independently, so the likelihood of all labels given all inputs is

$$L(\{\mathbf{w}^{(c)}\}) = \prod_{i=1}^{N} P(y_i \mid \mathbf{x}_i)$$

Taking the logarithm turns the product into a sum:

$$\ell(\{\mathbf{w}^{(c)}\}) = \sum_{i=1}^{N} \log P(y_i \mid \mathbf{x}_i)$$

Two cases for each i:
If $y = c \mid c = 1, \ldots, C-1$:

$$P\big(y_i = c \mid \mathbf{x}_i\big) = \frac{\exp\big(-\mathbf{w}^{(c)\top}\mathbf{x}_i\big)}{D_i}, \quad D_i = 1 + \sum_{k=1}^{C-1} \exp\big(-\mathbf{w}^{(k)\top}\mathbf{x}_i\big)$$

Taking its log

$$\log P(y_i = c \mid \mathbf{x}_i) = -\,\mathbf{w}^{(c)\top}\mathbf{x}_i - \log D_i$$

If $y_i = 0$:

$$P(y_i = 0 \mid \mathbf{x}_i) = \frac{1}{D_i}, \quad \log P(y_i = 0 \mid \mathbf{x}_i) = -\log D_i$$

Introduce $\mathbb{I}(y_i = c)$ which is 1 when $y_i = c$. Then for each i

$$-\log P(y_i \mid \mathbf{x}_i) = \log D_i + \sum_{c=1}^{C-1} \mathbb{I}(y_i = c)\, \mathbf{w}^{(c)\top} \mathbf{x}_i$$

Summing the negative log-probabilities over $i = 1, \ldots, N$ yields the total unregularized loss:

$$\mathcal{L} = \sum_{i=1}^{N} \Big[\log\big(1 + \sum_{k=1}^{C-1} e^{-\mathbf{w}^{(k)\top} \mathbf{x}_i}\big) + \sum_{c=1}^{C-1} \mathbb{I}(y_i = c)\, \mathbf{w}^{(c)\top} \mathbf{x}_i \Big]$$

3. Regularized Negative Log-Likelihood The unregularized negative log-likelihood is

$$\mathcal{L} = \sum_{i=1}^{N} \Big[\log\big(1 + \sum_{k=1}^{C-1} e^{-\mathbf{w}^{(k)\top} \mathbf{x}_i}\big) + \sum_{c=1}^{C-1} \mathbb{I}(y_i = c)\, \mathbf{w}^{(c)\top} \mathbf{x}_i \Big]$$

Add an ℓ_2 penalty:

$$\mathcal{L}_{\text{reg}} = \mathcal{L} \; + \; \frac{\lambda}{2} \sum_{c=1}^{C-1} \big\| \mathbf{w}^{(c)} \big\|^2.$$

4. Gradient Computation Regularized negative log-likelihood for class c:

$$\mathcal{L}_{\text{reg}} = \sum_{i=1}^{N} \Big[\log\big(1 + \sum_{k=1}^{C-1} e^{-\mathbf{w}^{(k)\top} \mathbf{x}_i}\big) + \sum_{k=1}^{C-1} \mathbb{I}(y_i = k)\, \mathbf{w}^{(k)\top} \mathbf{x}_i \Big] + \frac{\lambda}{2} \sum_{k=1}^{C-1} \|\mathbf{w}^{(k)}\|^2$$

Consider the loss for a single example i:

$$\ell_i = \log D_i + \sum_{k=1}^{C-1} \mathbb{I}(y_i = k)\, \mathbf{w}^{(k)\top} \mathbf{x}_i, \quad D_i = 1 + \sum_{k=1}^{C-1} e^{-\mathbf{w}^{(k)\top} \mathbf{x}_i}$$

Taking thr gradient w.r.t $\mathbf{w}^{(c)}$ splits into two: Derivative of $\log D_i$:

$$\frac{\partial}{\partial \mathbf{w}^{(c)}} \log D_i = \frac{1}{D_i} \frac{\partial}{\partial \mathbf{w}^{(c)}} \big(1 + \sum_{k} e^{-\mathbf{w}^{(k)\top} \mathbf{x}_i}\big) = -\frac{e^{-\mathbf{w}^{(c)\top} \mathbf{x}_i}}{D_i}\, \mathbf{x}_i = -P(y = c \mid \mathbf{x}_i)\, \mathbf{x}_i$$

Derivative of indicator term:

$$\frac{\partial}{\partial \mathbf{w}^{(c)}} \big[\mathbb{I}(y_i = c)\, \mathbf{w}^{(c)\top} \mathbf{x}_i \big] = \mathbb{I}(y_i = c)\, \mathbf{x}_i$$

Combining these two for an i:

$$\frac{\partial \ell_i}{\partial \mathbf{w}^{(c)}} = \big(\mathbb{I}(y_i = c) - P(y = c \mid \mathbf{x}_i) \big)\, \mathbf{x}_i$$

Summing the per-example gradients over $i = 1, \ldots, N$ yields

$$\nabla_{\mathbf{w}^{(c)}} \mathcal{L} = \sum_{i=1}^{N} \big(\mathbb{I}(y_i = c) - P(y = c \mid \mathbf{x}_i) \big)\, \mathbf{x}_i$$

Gradient for regularizer

$$\frac{\partial (\frac{\lambda}{2} \big\| \mathbf{w}^{(c)} \big\|^2)}{\partial \mathbf{w}^{(c)}} = \lambda \mathbf{w}^{(c)}$$

Adding the ℓ_2 Regularizer

$$\nabla_{\mathbf{w}^{(c)}} \mathcal{L}_{\text{reg}} = \sum_{i=1}^{N} \big(\mathbb{I}(y_i = c) - P(y = c \mid \mathbf{x}_i) \big)\, \mathbf{x}_i + \lambda\, \mathbf{w}^{(c)}$$

5. Gradient Descent Update With learning rate $\eta > 0$, update each class-weight:

$$\mathbf{w}^{(c)} \leftarrow \mathbf{w}^{(c)} - \eta \Big[\nabla_{\mathbf{w}^{(c)}} \mathcal{L}_{\text{reg}} \Big]$$

5. Prediction Rule For a new input $\mathbf{x}^*$,

$$y^* = \arg \max_{c \in \{0,\ldots,C-1\}} P(y = c \mid \mathbf{x}^*).$$