Problem Set 2 *

Tasha Pais†

Problem 1

Given:

$$p(y|x) = \frac{exp(w_y^T x)}{\sum_{y'=1}^{L} exp(w_{y'}^T x)}$$

To show that the log-odds between two labels y and y' is modeled by a linear function, we can consider:

$$\log\left(\frac{p(y|x)}{p(y'|x)}\right)$$

Substitute the given probabilities:

$$\log\left(\frac{\frac{exp(w_y^T x)}{\sum_{y''=1}^{L} exp(w_{y''}^T x)}}{\frac{exp(w_{y'}^T x)}{\sum_{y''=1}^{L} exp(w_{y''}^T x)}}\right)$$

Now, since the denominators are the same, they cancel out:

$$\log\left(\frac{exp(w_y^T x)}{exp(w_{y'}^T x)}\right)$$

Using properties of logarithms:

$$w_y^T x - w_{y'}^T x$$

This is clearly a linear function with respect to x.

Therefore, the log-odds between any two labels y and y' in the softmax model is modeled by a linear function.

Problem 2

Softmax Model for L=2:

$$p(y = 1|x) = \frac{\exp\left(w_1^T x\right)}{\exp(w_1^T x) + \exp(w_2^T x)}$$
$$p(y = 2|x) = \frac{\exp\left(w_2^T x\right)}{\exp(w_1^T x) + \exp(w_2^T x)}$$

*Due: `October 18, 2023`, Student(s) worked with: `Collaborators`

†NetID: `tdp74`, Email: `tdp74@rutgers.edu`

Logistic Regression Model:

$$p(y=1|x) = \frac{1}{1+\exp(-w^T x)}$$
$$p(y=2|x) = 1 - p(y=1|x)$$

Equating the probability of the first label from both models:

$$\frac{\exp\big(w_1^T x\big)}{\exp(w_1^T x) + \exp(w_2^T x)} = \frac{1}{1+\exp(-w^T x)} \tag{1}$$

From the logistic regression, we have:

$$\exp\big(-w^T x\big) = \frac{1-p(y=1|x)}{p(y=1|x)} \tag{2}$$

Using the expression for $p(y=1|x)$ from the softmax model:

$$\exp\big(-w^T x\big) = \frac{\exp\big(w_2^T x\big)}{\exp(w_1^T x)} \tag{3}$$

Taking logarithm on both sides:

$$-w^T x = w_2^T x - w_1^T x$$
$$w^T x = w_1^T x - w_2^T x$$

From this, we express w as:

$$w = w_1 - w_2 \tag{4}$$

Problem 3

Recall the softmax function for class j:

$$p_W(j|x) = \frac{\exp\big(w_j^T x\big)}{\sum_{y'=1}^{L} \exp\big(w_{y'}^T x\big)} \tag{5}$$

Adding an arbitrary vector c to each w_i, we get:

$$p_{W'}(j|x) = \frac{\exp\big((w_j + c)^T x\big)}{\sum_{y'=1}^{L} \exp((w_{y'} + c)^T x)} \tag{6}$$

Expanding, this is:

$$p_{W'}(j|x) = \frac{\exp\big(w_j^T x + c^T x\big)}{\sum_{y'=1}^{L} \exp\big(w_{y'}^T x + c^T x\big)} \tag{7}$$

Using properties of exponentials:

$$p_{W'}(j|x) = \frac{\exp\big(w_j^T x\big) \cdot \exp\big(c^T x\big)}{\sum_{y'=1}^{L} \exp\big(w_{y'}^T x\big) \cdot \exp(c^T x)} \tag{8}$$

As the term $\exp\left(c^T x\right)$ is common, it can be canceled out, yielding:

$$p_{W'}(j|x) = \frac{\exp\left(w_j^T x\right)}{\sum_{y'=1}^{L} \exp\left(w_{y'}^T x\right)} \tag{9}$$

Now, if we set $v_i = w_i - w_L$, and since w_L becomes the zero vector, we have:

$$v_i = w_i \tag{10}$$

For the class L, corresponding to w_L:

$$p_V(L|x) = \frac{\exp(0)}{\sum_{y'=1}^{L-1} \exp\left(v_{y'}^T x\right) + \exp(0)} \tag{11}$$

Inspecting closely, the normalization term in the denominator remains unchanged, implying that the probability distribution over the classes remains unchanged. We can represent the original parameters as:

$$\mathbf{v}_i = \mathbf{w}_i - \mathbf{w}_L \quad \text{for} \quad i = 1, \ldots, L-1 \tag{12}$$

and use a zero vector for the L-th class.

This demonstrates that we can represent the original softmax model using only $L-1$ nonzero parameter vectors instead of L. Hence, the softmax model is overparameterized.

Problem 4

Given the loss function:

$$J(W) = -\frac{1}{N} \sum_{i=1}^{N} \log p_W(y_i|x_i) + \lambda \sum_{j=1}^{d} \sum_{l=1}^{L} W_{j,l}^2$$

For a single data point:

$$J_i(W) = -\log p_W(y_i|x_i) + \lambda \sum_{j=1}^{d} \sum_{l=1}^{L} W_{j,l}^2$$

Differentiating the first term with respect to W:

$$\frac{\partial(-\log p_W(y_i|x_i))}{\partial W_{j,l}} = -\frac{1}{p_W(y_i|x_i)} \frac{\partial p_W(y_i|x_i)}{\partial W_{j,l}}$$

Given the softmax definition:

$$p_W(l|x_i) = \frac{\exp\left(W_l^T x_i\right)}{\sum_{k=1}^{L} \exp(W_k^T x_i)}$$

Differentiating with respect to $W_{j,l}$ and handling both cases:

$$\frac{\partial p_W(l|x_i)}{\partial W_{j,l}} = x_{i,j}(I(l = y_i) - p_W(l|x_i))$$

Where $I(\cdot)$ is the indicator function.

The gradient of the regularization term is:

$$\frac{\partial}{\partial W_{j,l}} \left(\lambda \sum_{j=1}^{d} \sum_{l=1}^{L} W_{j,l}^2 \right) = 2\lambda W_{j,l}$$

Combining both parts for the entire dataset, we get the gradient in matrix form:

$$\nabla J(W) = -X^T(G - P) + 2\lambda W$$

Where:

- X is the data matrix of size $N \times d$.
- G is the gold label matrix of size $N \times L$.
- P is the model probability matrix of size $N \times L$.