Machine Learning
Problem Set 2

Ozaner Hansha

October 15, 2020

Question 1

Part a: Show that the softmax model corresponds to modeling the log-odds between any two classes $c_i, c_j \in \{1, \cdots, C\}$ by a linear function.

Solution: Note the following chain of equalities:

$$\begin{aligned}
\log \frac{\hat{p}(y = c_i \mid \mathbf{x}; \mathbf{W})}{\hat{p}(y = c_j \mid \mathbf{x}; \mathbf{W})} &= \log \frac{\text{softmax}(\mathbf{w}_i \cdot \mathbf{x})}{\text{softmax}(\mathbf{w}_j \cdot \mathbf{x})} \\
&= \log \frac{\exp(\mathbf{w}_i \cdot \mathbf{x})}{\exp(\mathbf{w}_j \cdot \mathbf{x})} \\
&= \log \exp(\mathbf{w}_i \cdot \mathbf{x} - \mathbf{w}_j \cdot \mathbf{x}) \\
&= \mathbf{w}_i \cdot \mathbf{x} - \mathbf{w}_j \cdot \mathbf{x} \\
&= (\mathbf{w}_i - \mathbf{w}_j) \cdot \mathbf{x}
\end{aligned}$$

Letting $\mathbf{v}_{ij} = \mathbf{w}_i - \mathbf{w}_j$, we have that the log-odds of any two classes is modeled by the following linear function:

$$\log \frac{\hat{p}(y = c_i \mid \mathbf{x}; \mathbf{W})}{\hat{p}(y = c_j \mid \mathbf{x}; \mathbf{W})} = \mathbf{v}_{ij} \cdot \mathbf{x}$$

Part b: Show that in the binary case ($C = 2$), for any two D-dimensional parameter vectors w_1 and w_2 in the softmax model, there exists a single D-dimensional parameter vector $\mathbf{v}$ such that:

$$\text{softmax}(\mathbf{w}_1 \cdot \mathbf{x}) = \frac{\exp(\mathbf{w}_1 \cdot \mathbf{x})}{\exp(\mathbf{w}_1 \cdot \mathbf{x}) + \exp(\mathbf{w}_2 \cdot \mathbf{v})} = \sigma(\mathbf{v} \cdot \mathbf{x})$$

Solution: Consider the following chain of equalities:

$$\begin{aligned}
\text{softmax}(\mathbf{w}_1 \cdot \mathbf{x}) &= \frac{\exp(\mathbf{w}_1 \cdot \mathbf{x})}{\exp(\mathbf{w}_1 \cdot \mathbf{x}) + \exp(\mathbf{w}_2 \cdot \mathbf{v})} \\
&= \left(\frac{\exp(\mathbf{w}_1 \cdot \mathbf{x}) + \exp(\mathbf{w}_2 \cdot \mathbf{v})}{\exp(\mathbf{w}_1 \cdot \mathbf{x})} \right)^{-1} \\
&= \left(1 + \frac{\exp(\mathbf{w}_2 \cdot \mathbf{v})}{\exp(\mathbf{w}_1 \cdot \mathbf{x})} \right)^{-1} \\
&= (1 + \exp(\mathbf{w}_2 \cdot \mathbf{v} - \mathbf{w}_1 \cdot \mathbf{x}))^{-1} \\
&= (1 + \exp((\mathbf{w}_2 - \mathbf{w}_1) \cdot \mathbf{x}))^{-1} \\
&= (1 + \exp(-(\mathbf{w}_1 - \mathbf{w}_2) \cdot \mathbf{x}))^{-1} \\
&= \sigma((\mathbf{w}_1 - \mathbf{w}_2) \cdot \mathbf{x})
\end{aligned}$$

Again, letting $\mathbf{v}_{ij} = \mathbf{w}_i - \mathbf{w}_j$, we have that the binary case of softmax is equivalent to the logistic model:

$$\text{softmax}(\mathbf{w}_1 \cdot \mathbf{x}) = \sigma(\mathbf{v}_{12} \cdot \mathbf{x})$$

Question 2

Part a: Show that the softmax model is over-parameterized by showing that for any weight matrix $\mathbf{W}$ there is another $\mathbf{W}'$ that produces the same probabilities, i.e. $p(y \mid \mathbf{x}; \mathbf{W}) = p(y \mid \mathbf{x}; \mathbf{W}')$.

Solution: First note that for an arbitrary class c_i, its predicted probability given some input $\mathbf{x}$ and weight $\mathbf{W}$ is:

$$\hat{p}(y = c_i \mid \mathbf{x}; \mathbf{W}) = \text{softmax}(\mathbf{w}_i \cdot \mathbf{x}) = \frac{\exp(\mathbf{w}_i \cdot \mathbf{x})}{\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x})}$$

Now consider another weight matrix $\mathbf{W}'$ whose rows are given by $\mathbf{w}'_j = \mathbf{w}_j + \mathbf{u}$, for any arbitrary vector $\mathbf{u} \in \mathbb{R}^D$. The predicted probability of class c_i with this is given by:

$$\begin{aligned}
\hat{p}(y = c_i \mid \mathbf{x}; \mathbf{W}') &= \text{softmax}(\mathbf{w}'_i \cdot \mathbf{x}) \\
&= \frac{\exp(\mathbf{w}'_i \cdot \mathbf{x})}{\sum_{j=1}^{C} \exp(\mathbf{w}'_j \cdot \mathbf{x})} \\
&= \frac{\exp((\mathbf{w}_i + \mathbf{u}) \cdot \mathbf{x})}{\sum_{j=1}^{C} \exp((\mathbf{w}_j + \mathbf{u}) \cdot \mathbf{x})} \\
&= \frac{\exp(\mathbf{w}_i \cdot \mathbf{x} + \mathbf{u} \cdot \mathbf{x})}{\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x} + \mathbf{u} \cdot \mathbf{x})} \\
&= \frac{\exp(\mathbf{u} \cdot \mathbf{x}) \exp(\mathbf{w}_i \cdot \mathbf{x})}{\exp(\mathbf{u} \cdot \mathbf{x}) \sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x})} \\
&= \frac{\exp(\mathbf{w}_i \cdot \mathbf{x})}{\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x})} \\
&= \hat{p}(y = c_i \mid \mathbf{x}; \mathbf{W})
\end{aligned}$$

And so for any matrix $\mathbf{W}$ and any choice of vector $\mathbf{u} \in \mathbb{R}^D$, there exists another matrix $\mathbf{W}'$ that provides the same probabilites in the context of the softmax model.

Part b: Explain how this over-paramaterization implies that we only need $C - 1$ vector paramaters $\mathbf{w}_i$ for the softmax model rather than C

Solution: Recall that any matrix $\mathbf{W} \in \mathbb{R}^{C \times D}$ is part of a larger equivalence class of matrices that produce the same predictions with softmax:

$$[\mathbf{W}] = \{\mathbf{W} + \mathbf{1}_D \otimes \mathbf{u} \mid \mathbf{u} \in \mathbb{R}^C\}$$

where $\mathbf{1}_D$ *is a* D *dimensional vector of 1s and* $\otimes$ *is the outer product. The above characterization of* $[\mathbf{W}]$ *is equivalent to the one we used in part a.*

As such, the space of these equivalence classes is isomorphic to the quotient product $\mathbb{R}^{CD}/\mathbb{R}^C$. But note that this quotient product itself is isomorphic to:

$$\mathbb{R}^{CD}/\mathbb{R}^C \cong \mathbb{R}^{(C-1)D} \cong \mathbb{R}^{(C-1) \times D}$$

Or to be more direct, the space of weight matrices that are unique under softmax has a dimension of $(C - 1) \times D$ meaning we only need $C - 1$, D-dimensional vectors to parameterize our model.

Question 3

Part a: Give the L_2-regularized log-loss of the softmax model, for a single training example $(\mathbf{x}, y)$.

Solution: If $y = c_i$, then the log-loss of this regularized softmax model is given by:

$$\begin{aligned}
L((\mathbf{x}, y), \mathbf{W}) &= -\log \hat{p}(y = c_i \mid \mathbf{x}; \mathbf{W}) + \lambda \|\mathbf{W}\|^2 \\
&= -\log \operatorname{softmax}(\mathbf{w}_i \cdot \mathbf{x}) + \lambda \|\mathbf{W}\|^2 \\
&= -\log \frac{\exp(\mathbf{w}_i \cdot \mathbf{x})}{\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x})} + \lambda \|\mathbf{W}\|^2 \\
&= -\mathbf{w}_i \cdot \mathbf{x} + \log \left(\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x}) \right) + \lambda \|\mathbf{W}\|^2
\end{aligned}$$

Part b: Give the gradients of the loss from part a, with respect to each weight vector $\mathbf{w}_j$.

Solution: For the case of $\mathbf{w}_i$, i.e. $y = c_i$, we have:

$$\begin{aligned}
\nabla_{\mathbf{w}_i} L((\mathbf{x}, y), \mathbf{W}) &= \nabla_{\mathbf{w}_i} \left(-\mathbf{w}_i \cdot \mathbf{x} + \log \left(\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x}) \right) + \lambda \|\mathbf{W}\|^2 \right) \\
&= -\nabla_{\mathbf{w}_i} \mathbf{w}_i \cdot \mathbf{x} + \nabla_{\mathbf{w}_i} \log \left(\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x}) \right) + \lambda \nabla_{\mathbf{w}_i} \|\mathbf{W}\|^2 \\
&= -\mathbf{x} + \frac{\nabla_{\mathbf{w}_i} \left(\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x}) \right)}{\left(\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x}) \right)} + \lambda \nabla_{\mathbf{w}_i} \|\mathbf{W}\|^2 && \text{(chain rule)} \\
&= -\mathbf{x} + \frac{\nabla_{\mathbf{w}_i} \exp(\mathbf{w}_i \cdot \mathbf{x})}{\left(\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x}) \right)} + \lambda \nabla_{\mathbf{w}_i} \|\mathbf{W}\|^2 \\
&= -\mathbf{x} + \frac{\exp(\mathbf{w}_i \cdot \mathbf{x})\mathbf{x}}{\left(\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x}) \right)} + \lambda \nabla_{\mathbf{w}_i} \|\mathbf{W}\|^2 \\
&= -\mathbf{x} + \frac{\exp(\mathbf{w}_i \cdot \mathbf{x})\mathbf{x}}{\left(\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x}) \right)} + \lambda \nabla_{\mathbf{w}_i} \sum_{j=1}^{C} \sum_{k=1}^{D} W_{jk}^2 \\
&= -\mathbf{x} + \frac{\exp(\mathbf{w}_i \cdot \mathbf{x})\mathbf{x}}{\left(\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x}) \right)} + \lambda \nabla_{\mathbf{w}_i} \sum_{k=1}^{D} W_{ik}^2 \\
&= -\mathbf{x} + \frac{\exp(\mathbf{w}_i \cdot \mathbf{x})\mathbf{x}}{\left(\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x}) \right)} + 2\lambda \mathbf{w}_i
\end{aligned}$$

And in the case of $\mathbf{w}_j$ where $j \neq i$, the gradient is given by:

$$
\begin{aligned}
\nabla_{\mathbf{w}_j} L((\mathbf{x}, y), \mathbf{W}) &= \nabla_{\mathbf{w}_j} \left(-\mathbf{w}_i \cdot \mathbf{x} + \log \left(\sum_{k=1}^{C} \exp(\mathbf{w}_k \cdot \mathbf{x}) \right) + \lambda \|\mathbf{W}\|^2 \right) \\
&= -\nabla_{\mathbf{w}_j} \mathbf{w}_i \cdot \mathbf{x} + \nabla_{\mathbf{w}_j} \log \left(\sum_{k=1}^{C} \exp(\mathbf{w}_k \cdot \mathbf{x}) \right) + \lambda \nabla_{\mathbf{w}_j} \|\mathbf{W}\|^2 \\
&= \nabla_{\mathbf{w}_j} \log \left(\sum_{k=1}^{C} \exp(\mathbf{w}_k \cdot \mathbf{x}) \right) + \lambda \nabla_{\mathbf{w}_j} \|\mathbf{W}\|^2 \\
&= \frac{\nabla_{\mathbf{w}_i} \left(\sum_{k=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x}) \right)}{\left(\sum_{k=1}^{C} \exp(\mathbf{w}_k \cdot \mathbf{x}) \right)} + \lambda \nabla_{\mathbf{w}_j} \|\mathbf{W}\|^2 && \text{(chain rule)} \\
&= \frac{\nabla_{\mathbf{w}_j} \exp(\mathbf{w}_j \cdot \mathbf{x})}{\left(\sum_{k=1}^{C} \exp(\mathbf{w}_k \cdot \mathbf{x}) \right)} + \lambda \nabla_{\mathbf{w}_j} \|\mathbf{W}\|^2 \\
&= \frac{\exp(\mathbf{w}_j \cdot \mathbf{x}) \mathbf{x}}{\left(\sum_{k=1}^{C} \exp(\mathbf{w}_k \cdot \mathbf{x}) \right)} + \lambda \nabla_{\mathbf{w}_j} \sum_{k=1}^{C} \sum_{l=1}^{D} W_{kl}^2 \\
&= \frac{\exp(\mathbf{w}_j \cdot \mathbf{x}) \mathbf{x}}{\left(\sum_{k=1}^{C} \exp(\mathbf{w}_k \cdot \mathbf{x}) \right)} + \lambda \nabla_{\mathbf{w}_j} \sum_{l=1}^{D} W_{jl}^2 \\
&= \frac{\exp(\mathbf{w}_j \cdot \mathbf{x}) \mathbf{x}}{\left(\sum_{k=1}^{C} \exp(\mathbf{w}_k \cdot \mathbf{x}) \right)} + 2\lambda \mathbf{w}_j
\end{aligned}
$$

Part c: Give the update equations for stochastic gradient descent for the softmax model, with learning rate η.

Solution: The update equation for the weight vector $\mathbf{w}_i$, where $y = c_i$ is given by:

$$
\begin{aligned}
\mathbf{w}_i^{(t+1)} &= \mathbf{w}_i^{(t)} - \eta \nabla_{\mathbf{w}_i} L((\mathbf{x}, y), \mathbf{W}^{(t)}) \\
&= \mathbf{w}_i^{(t)} - \eta \left(-\mathbf{x} + \frac{\exp(\mathbf{w}_i \cdot \mathbf{x}) \mathbf{x}}{\left(\sum_{j=1}^{C} \exp(\mathbf{w}_j \cdot \mathbf{x}) \right)} + 2\lambda \mathbf{w}_i \right)
\end{aligned}
$$

While the update equation for the weight vector $\mathbf{w}_j$, where $j \neq i$ is given by:

$$
\begin{aligned}
\mathbf{w}_j^{(t+1)} &= \mathbf{w}_j^{(t)} - \eta \nabla_{\mathbf{w}_j} L((\mathbf{x}, y), \mathbf{W}^{(t)}) \\
&= \mathbf{w}_j^{(t)} - \eta \left(\frac{\exp(\mathbf{w}_j \cdot \mathbf{x}) \mathbf{x}}{\left(\sum_{k=1}^{C} \exp(\mathbf{w}_k \cdot \mathbf{x}) \right)} + 2\lambda \mathbf{w}_j \right)
\end{aligned}
$$