CS 181 Spring 2020
Softmax

Consider a K-class classification problem.

Let $\{\mathbf{w}_k\}_{k=1}^K$ be defined such that for some data point $\mathbf{x}$, $z_k = \mathbf{w}_k^\top \mathbf{x}$ can be interpretted as a score for $\mathbf{x}$ belonging to class k.

Multi-class Logistic Regression (LR) with a trained set of weights assigns $\mathbf{x}$ the class k for which it has the highest such score.

The **softmax transformation** takes as input a vector, and outputs a transformed vector of the same size.

$$\text{softmax}(\mathbf{z})_k = \frac{\exp(z_k)}{\sum_{\ell=1}^K \exp(z_\ell)}, \text{ for all } k$$

LR uses the softmax over a vector of K scores $\mathbf{z} = [\mathbf{w}_1^\top \mathbf{x}, \ldots, \mathbf{w}_K^\top \mathbf{x}]$ so that it can be normalized and intepretted as a vector of *probabilities*.

$$p(\mathbf{y} = C_k \,|\, \mathbf{x}; \{\mathbf{w}_\ell\}_{\ell=1}^K) = \text{softmax}([\mathbf{w}_1^\top \mathbf{x} \ldots \mathbf{w}_K^\top \mathbf{x}]^\top)_k = \frac{\exp(\mathbf{w}_k^\top \mathbf{x})}{\sum_{\ell=1}^K \exp(\mathbf{w}_\ell^\top \mathbf{x})}\,.$$

where C_k is a *one-hot* vector with a 1 in coordinate k and 0s elsewhere.

Assuming data $D = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, the negative log-likelihood can be written as:

$$\mathcal{L}(\{\mathbf{w}_\ell\}) = -\sum_{i=1}^N \ln p(\mathbf{y}_i \,|\, \mathbf{x}_i; \{\mathbf{w}_\ell\})$$

The softmax is an important function and you will see it again in other models, such as neural networks. In this problem, we aim to gain intuitions into the properties of softmax and multiclass logistic regression.

1. The output of the softmax is a vector with non-negative components that are at most 1.

   Reason The j^{th} component of the softmax function $\text{softmax}(\mathbf{z})$ is:

   $$\text{softmax}(\mathbf{z})_j = \frac{\exp(z_j)}{\sum_i \exp(z_i)}.$$

   As $\exp(x) > 0$ for all $x \in \mathbb{R}$, we have $\exp(z_j) > 0$ and $\sum_i \exp(z_i) > 0$. Thus the output of the softmax function is a vector with non-negative components. Since $\exp(z_j)$ appears in both the numerator and the denominator (as the $i = j$ term in the sum), the denominator must be at least as large as the numerator, and so the components are at most 1.

2. The output of the softmax defines a distribution, so the components sum to 1.

   Reason Summing over the components:

$$\sum_j \text{softmax}(\mathbf{z})_j = \sum_j \frac{\exp(z_j)}{\sum_i \exp(z_i)} = \frac{\sum_j \exp(z_j)}{\sum_i \exp(z_i)} = 1.$$

3. Softmax preserves order. This means that if elements $z_k < z_\ell$ in $\mathbf{z}$, then $\text{softmax}(\mathbf{z})_k < \text{softmax}(\mathbf{z})_\ell$ for any k, ℓ.

   Reason If $z_j \geq z_k$, then $\exp(z_j) \geq \exp(z_k)$ as the exponential is a monotonically increasing function. Dividing by the positive constant $\sum_i \exp(z_i)$, this inequality implies that:

$$\text{softmax}(\mathbf{z})_j = \frac{\exp(z_j)}{\sum_i \exp(z_i)} \geq \frac{\exp(z_k)}{\sum_i \exp(z_i)} = \text{softmax}(\mathbf{z})_k,$$

   which shows that the softmax function preserves the order of the elements of $\mathbf{z}$.

4.

$$\frac{\partial \text{softmax}(\mathbf{z})_k}{\partial z_j} = \text{softmax}(\mathbf{z})_k (I_{kj} - \text{softmax}(\mathbf{z})_j) \text{ for any } k, j$$

   where indicator $I_{kj} = 1$ if $k = j$ and $I_{kj} = 0$ otherwise.

   Reason If $j \neq k$, then:

$$\begin{aligned}\frac{\partial \text{softmax}(\mathbf{z})_k}{\partial z_j} &= \frac{\partial}{\partial z_j} \frac{\exp(z_k)}{\sum_i \exp(z_i)} = -\frac{\exp(z_k)}{\left(\sum_i \exp(z_i)\right)^2} \exp(z_j) \\ &= -\frac{\exp(z_k)}{\sum_i \exp(z_i)} \frac{\exp(z_j)}{\sum_i \exp(z_i)} = -\text{softmax}(\mathbf{z})_k \text{softmax}(\mathbf{z})_j.\end{aligned}$$

   If $j = k$ then:

$$\begin{aligned}\frac{\partial \text{softmax}(\mathbf{z})_k}{\partial z_j} &= \frac{\partial}{\partial z_j} \frac{\exp(z_k)}{\sum_i \exp(z_i)} = \frac{\exp(z_k)}{\sum_i \exp(z_i)} - \frac{\exp(z_j)^2}{\left(\sum_i \exp(z_i)\right)^2} \\ &= \left(1 - \frac{\exp(z_k)}{\sum_i \exp(z_i)}\right) \frac{\exp(z_k)}{\sum_i \exp(z_i)} = \text{softmax}(\mathbf{z})_k (1 - \text{softmax}(\mathbf{z})_j).\end{aligned}$$

   Putting these results together: $\boxed{\frac{\partial \text{softmax}(\mathbf{z})_k}{\partial z_j} = \text{softmax}(\mathbf{z})_k (I_{kj} - \text{softmax}(\mathbf{z})_j)}$

5. Using (4), show that:

$$\frac{\partial}{\partial \mathbf{w}_j}\mathcal{L}(\{\mathbf{w}_\ell\}) = \sum_{i=1}^{N}\Big[p(\mathbf{y}_i = C_j|\mathbf{x}_i;\{\mathbf{w}_\ell\}) - y_{ij}\Big]\mathbf{x}_i$$

Solution Write the negative log-likelihood.

$$\mathcal{L}(\{\mathbf{w}_\ell\}) = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\ln p(\mathbf{y} = C_k|\mathbf{x}_i;\{\mathbf{w}_\ell\})$$

$$\frac{\partial}{\partial \mathbf{w}_j}\mathcal{L}(\{\mathbf{w}_\ell\}) = -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\frac{\partial}{\partial \mathbf{w}_j}\ln p(\mathbf{y} = C_k|\mathbf{x}_i;\{\mathbf{w}_\ell\})$$

Using Derivative of log + chain rule

$$= -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\left(\frac{1}{p(\mathbf{y} = C_k|\mathbf{x}_i;\{\mathbf{w}_\ell\})}\right)\frac{\partial}{\partial \mathbf{w}_j}p(\mathbf{y} = C_k|\mathbf{x}_i;\{\mathbf{w}_\ell\})$$

Rewrite using chain rule

$$= -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\left(\frac{1}{p(\mathbf{y} = C_k|\mathbf{x}_i;\{\mathbf{w}_\ell\})}\right)\frac{\partial}{\partial z_j}p(\mathbf{y} = C_k|\mathbf{x}_i;\{\mathbf{w}_\ell\})\frac{\partial}{\partial \mathbf{w}_j}z_j$$

The derivative at the end is just the derivative of a dot product:

$$= -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\left(\frac{1}{p(\mathbf{y} = C_k|\mathbf{x}_i;\{\mathbf{w}_\ell\})}\right)\frac{\partial}{\partial z_j}p(\mathbf{y} = C_k|\mathbf{x}_i;\{\mathbf{w}_\ell\})\mathbf{x}$$

Use the derivative of the softmax found in (d)

$$= -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\left(\frac{1}{p(\mathbf{y} = C_k|\mathbf{x}_i;\{\mathbf{w}_\ell\})}\right)\Big(p(\mathbf{y} = C_k|\mathbf{x}_i;\{\mathbf{w}_\ell\})\Big)\Big(I_{kj} - p(\mathbf{y} = C_j|\mathbf{x}_i;\{\mathbf{w}_\ell\})\Big)\mathbf{x}_i$$

Notice that two terms are conveniently reciprocals and simplify

$$= -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\Big(I_{kj} - p(\mathbf{y} = C_j|\mathbf{x}_i;\{\mathbf{w}_\ell\})\Big)\mathbf{x}_i$$

Foil the terms

$$= -\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}I_{kj}\mathbf{x}_i + \sum_{i=1}^{N} p(\mathbf{y} = C_j|\mathbf{x}_i;\{\mathbf{w}_\ell\})\mathbf{x}_i\Big(\sum_{k=1}^{C} y_{ik}\Big)$$

The I_{kj} in the first sum collapses the sum over k to the term where $j = k$. As the $\mathbf{y}_i$ are *one-hot*, we have that $\sum_{k=1}^{K} y_{ik} = 1$. Using these facts:

$$\frac{\partial}{\partial \mathbf{w}_j} \mathcal{L}(\{\mathbf{w}_\ell\}) = -\sum_{i=1}^{N} y_{ij} \mathbf{x}_i + \sum_{i=1}^{N} p(\mathbf{y} = C_j | \mathbf{x}_i; \{\mathbf{w}_\ell\}) \mathbf{x}_i$$

$$\boxed{= \sum_{i=1}^{N} (p(\mathbf{y} = C_j | \mathbf{x}_i; \{\mathbf{w}_\ell\}) - y_{ij}) \mathbf{x}_i}$$