Apply Loss Functions

Neural networks are versatile models that can be used to do both regression and classification. The difference between those two types of networks will be the final activation function and the output vector.

Recall that a neural network with $l-1$ hidden layers has the following generic form:

$$\mathbf{h}(\mathbf{x}) = \mathbf{W}_l\sigma(\mathbf{W}_{l-1}\sigma(\cdots\mathbf{W}_2\sigma(\mathbf{W}_1\mathbf{x})\cdots))$$

Regression:

If we want to do a one dimensional regression (i.e., our label y is a real scalar), then our neural network must have a one dimensional output. This can be trivially done by setting the dimension of $\mathbf{W}_l$ to $1 \times d_{l-1}$, where d_{l-1} is the dimension of the output of the last hidden layer. With the size of the output set, all we need to do now is decide a loss function. Because we are performing regression, we will use the more popular loss function, squared loss:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}\ell(\mathbf{h}(\mathbf{x}_i), y_i) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{h}(\mathbf{x}_i) - y_i)^2$$

In this case, note that the outputs $\mathbf{h}(\mathbf{x}_i)$ are one-dimensional (i.e., scalars). With this loss function, we can calculate the gradient with respect to all weights $\mathbf{W}_1, \ldots, \mathbf{W}_{l-1}, \mathbf{W}_l$ and do (stochastic) gradient descent to learn those weights, which will be elaborated upon soon. Of course nothing is stopping us from regressing higher dimensional labels.

Classification:

Classification problems output belief vectors (also known as logits). Every element in the belief vector corresponds to the model's belief that the input example belongs to a certain class. For example, imagine we have a k-class classification problem with labels $\{1, \ldots, k\}$. Our model will output a k dimensional belief vector. If the i^{th} element has the highest value, then the model believes that that input belongs to class i.

To output a k-dimensional belief vector, we set the dimension of $\mathbf{W}_l$ to $k \times d_{l-1}$, where d_{l-1} is the dimension of the output of the last hidden layer. However, we are not done yet. The output of a neural network is a real vector, which is not easy to work with, as each element could be any value on different scales. As such, we usually apply the **softmax** transition function (or softmax layer).

The softmax layer is a k-dimensional vector $\mathbf{softmax}(\mathbf{h}(\mathbf{x}))$ whose components are defined as follows:

$$[\mathbf{softmax}(\mathbf{h}(\mathbf{x}))]_i = \frac{\exp([\mathbf{h}(\mathbf{x})]_i)}{\sum_{j=1}^{k}\exp([\mathbf{h}(\mathbf{x})]_j)} \quad \text{for all } i \in \{1, \ldots, k\}$$

Note: $\exp(u)$ is the natural exponential function e^u.

All the softmax is doing is that it first exponentiates each dimension $[\mathbf{h}(\mathbf{x})]_i \mapsto \exp([\mathbf{h}(\mathbf{x})]_i)$ and then normalizes these values across classes. The exponentiation has two effects:

1. all values become positive;
2. whatever output is largest will be highly dominant.

The normalization then turns all outputs into well-defined probabilities.

Note: Convince yourself that the sum of all entries of the softmax output is 1 and each entry is in the range of 0 to 1.

The softmax layer essentially produces a vector of probabilities where the i-th dimension represents the probability of $\mathbf{x}$ belonging to class i, i.e., $[\mathbf{softmax}(\mathbf{h}(\mathbf{x}))]_i = P(y = i|\mathbf{x})$ for any $i \in \{1, \ldots, k\}$. For a single example $(\mathbf{x}_i, y_i)$, we interpret $P(y_i|\mathbf{x}_j) = [\mathbf{softmax}(\mathbf{h}(\mathbf{x}_i))]_{y_i}$ as the probability that the neural network predicts $\mathbf{x}_i$ has the (true) class label y_i.

With this probabilistic representation, we can use our MLE principle — minimizing the negative log-likelihood — to identify an appropriate loss function for classification ($\boldsymbol{\theta}$ represents the parameters of the neural network model that must be adjusted to find the max or min):

☆ **Key Points**

Neural networks can be used for regression and classification.

The output of a regression network is a real number or vector of real numbers.

The output of the neural network for classification is a normalized probability vector.

$$\overbrace{\arg\max_{\boldsymbol{\theta}} \underbrace{\prod_{i=1}^{n} P(y_i|\mathbf{x}_i)}_{\text{Likelihood}}}^{\text{MLE Principle}} = \arg\max_{\boldsymbol{\theta}} \underbrace{\sum_{i=1}^{n} \log(P(y_i|\mathbf{x}_i))}_{\text{Log-Likelihood}}$$

$$= \arg\min_{\boldsymbol{\theta}} \underbrace{-\sum_{i=1}^{n} \log(P(y_i|\mathbf{x}_i))}_{\text{Negative Log-Likelihood}}$$

$$= \arg\min_{\boldsymbol{\theta}} \underbrace{\sum_{i=1}^{n} -\log\Big([\text{softmax}(\mathbf{h}(\mathbf{x}_i))]_{y_i}\Big)}_{\text{softmax models probability that } \mathbf{x}_i \text{ has label } y_i}$$

$$= \arg\min_{\boldsymbol{\theta}} \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{h}(\mathbf{x}_i), y_i)}_{\text{include } \frac{1}{n} \text{ to interpret as average loss}}$$

Applying the MLE principle leads to an (average per example) loss function of the following form:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{h}(\mathbf{x}_i), y_i)$$

Here, the individual loss for a single example $(\mathbf{x}_i, y_i)$ is defined by

$$\ell(\mathbf{h}(\mathbf{x}_i), y_i) = -\log\Big([\text{softmax}(\mathbf{h}(\mathbf{x}_i))]_{y_i}\Big) = -\log\left(\frac{\exp\big([\mathbf{h}(\mathbf{x}_i)]_{y_i}\big)}{\sum_{j=1}^{k}\exp([\mathbf{h}(\mathbf{x}_i)]_j)}\right)$$

So the (average per example) loss for a batch of examples is

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{\exp\big([\mathbf{h}(\mathbf{x}_i)]_{y_i}\big)}{\sum_{j=1}^{k}\exp([\mathbf{h}(\mathbf{x}_i)]_j)}\right)$$