Neural Network Derivatives

Justin Domke
June 7, 2008

This note gives the derivatives needed to fit “logistic” neural networks for classification by maximum likelihood. Here, these are thought of as doing logistic regression on features derived by the learning process. This is essentially the interpretation of neural nets given in The Elements of Statistical Learning by T. Hastie, R. Tibshirani, and J. H. Friedman.

I wrote this note because I needed these formulas for my research, and I couldn’t find them in this form. The difference from the normal neural net presentation is A) the output is a “logistic layer” B) learning is by maximum likelihood, and C) the derivatives are given in a “direct” matrix format. The matrix form, by avoiding indexes, can make things much easier to implement.

For the impatient, see the “cheatsheet” at the end. In these notes, ⊙ denotes the element-wise product.

1 Regular old logistic regression

First, let’s review regular logistic regression. This will be done in a somewhat clumsy way, to make it easier to generalize. Given input x, this gives a probability distribution over some discrete y by

\[ p(y|x) = \frac{1}{Z(x)} \exp(l(y|x)) = \frac{1}{Z(x)} \exp(w^{(y)^T x}) \]

\[ Z(x) = \sum_y \exp(w^{(y)^T x}). \]

So, x is projected linearly onto the vectors w^{(y)}, then exponentiated and normalized to give a probability distribution. A very standard, classical method. Usually, these are fit by maximum likelihood by maximizing (over w^{(y)}) the criterion

\[ \sum_{(x,y)} \log p(y|x) = \sum_{(x,y)} L(y, x) = \sum_{(x,y)} l(y|x) - \log \sum_y \exp l(y|x). \]

For some arbitrary parameter \( \theta \), the derivatives are given by
\[
\frac{\partial}{\partial \theta} L(\hat{y}, \hat{x}) = \frac{\partial}{\partial \theta} l(\hat{y}\hat{x}) - \frac{1}{Z(x)} \sum_y \exp l(y|\hat{x}) \frac{\partial}{\partial \theta} l(y|\hat{x}).
\]

So all we finally need is \( \frac{\partial}{\partial \theta} l(y|x) \). This is easy.

\[
\frac{\partial}{\partial w(y')} l(y|x) = \begin{cases} 
    x, & y' = y \\
    0, & y' \neq y 
\end{cases}
\]

The derivative of the probabilities \( \frac{\partial}{\partial \theta} p(y|x) \) is not needed for learning, but can be useful in other circumstances. It is given by

\[
\frac{\partial}{\partial \theta} p(y|x) = \frac{\partial}{\partial \theta} \frac{1}{Z(x)} \exp l(y|x)
\]

\[
= \frac{1}{Z(x)} \exp l(y|x) \frac{\partial}{\partial \theta} l(y|x) + \exp l(y|x) \frac{1}{Z(x)^2} \frac{\partial}{\partial \theta} Z(x)
\]

\[
= p(y|x) \frac{\partial}{\partial \theta} l(y|x) - \exp l(y|x) \frac{1}{Z(x)} \frac{\partial}{\partial \theta} l(y'|x)
\]

\[
= p(y|x) \frac{\partial}{\partial \theta} l(y|x) - p(y|x) \sum_{y'} \frac{1}{Z(x)} \exp l(y'|x) \frac{\partial}{\partial \theta} l(y'|x)
\]

\[
= p(y|x) \frac{\partial}{\partial \theta} l(y|x) - p(y|x) \sum_{y'} p(y'|x) \frac{\partial}{\partial \theta} l(y'|x)
\]

\[
= p(y|x) \frac{\partial}{\partial \theta} l(y|x) - p(y|x) \frac{\partial}{\partial \theta} l(y|x).
\]

Notice, however, this last step uses the particular form of the derivative of \( l \), so cannot be used later on.

## 2 One hidden layer

Now, we add a “hidden layer”. The change is to the form of \( l(y|x) \).

\[
l(y|x) = w^{(y)T} \sigma(Vx)
\]

Thus, \( x \) is linearly transformed, then passed through a sigmoid unit. So \( \sigma(Vx) \) should be thought of as the “derived” features for logistic regression.

Now, many of the formulas from above apply unchanged.

\[
p(y|x) = \frac{1}{Z(x)} \exp l(y|x)
\]
\[
\frac{\partial}{\partial \theta} L(\hat{y}, \hat{x}) = \frac{\partial}{\partial \theta} l(\hat{y}|\hat{x}) - \sum_y p(y|\hat{x}) \frac{\partial}{\partial \theta} l(y|\hat{x})
\]
\[
\frac{\partial}{\partial \theta} p(y|x) = p(y|x) \frac{\partial}{\partial \theta} l(y|x) - \sum_{y'} p(y'|x) \frac{\partial}{\partial \theta} l(y'|x)
\]

So all we need to do is derive a new formula for \(\frac{\partial}{\partial \theta} l(y|x)\) where \(\theta\) could be either a parameter of \(w\) or of \(V\). For \(w\), the derivatives are immediate from above.

\[
\frac{\partial}{\partial w(y')} l(y|x) = \begin{cases} 
\sigma(Vx), & y' = y \\ 
0, & y' \neq y 
\end{cases}
\]

The derivatives with respect to \(V\) are not too bad either.

\[
\frac{\partial}{\partial V_{ij}} l(y|x) = \frac{\partial}{\partial V_{ij}} w(y)^T \sigma(Vx) = w_i(y) \sigma'(V_{-i}x) x_j
\]

\[
\frac{\partial}{\partial V} l(y|x) = w(y) \odot \sigma'(Vx)x^T
\]

3 Two hidden layers

Now, there are two linear transformations of the data, hopefully giving the flexibility of more nonlinear features.

\[
l(y|x) = w(y)^T \sigma(V\sigma(Ux))
\]

The derivatives for \(w\) and \(V\) are direct from above.

\[
\frac{\partial}{\partial w(y')} l(y|x) = \begin{cases} 
\sigma(V\sigma(Ux)), & y' = y \\ 
0, & y' \neq y 
\end{cases}
\]

\[
\frac{\partial}{\partial V} l(y|x) = w(y) \odot \sigma'(V\sigma(Ux))\sigma(Ux)^T
\]

For \(U\), we still need to do some work. Again, I will use indices to derive the derivative, then go back to a matrix form. This derivation is unbelievably long.
\[
\frac{\partial}{\partial U_{ij}} l(y|x) = \sum_k \frac{\partial}{\partial U_{ij}} w_k^{(y)} \sigma(V_k \sigma(U_k)) = \sum_k w_k^{(y)} \sigma'(V_k \sigma(x)) \frac{\partial}{\partial U_{ij}} V_k \sigma(U_k)
\]

And so, at last,

\[
\frac{\partial}{\partial U} l(y|x) = V^T \left[ w^{(y)} \odot \sigma'(V \sigma(U)) \right] \odot \sigma'(Ux)^T.
\]
4 Cheatsheet

First, there are several formulas that apply to all the models. The only thing needed to make them specific is the form for $l$, and the derivatives of it.

\[ p(y|x) = \frac{1}{Z(x)} \exp(l(y|x)), \quad Z(x) = \sum_y \exp l(y|x) \]

\[ \frac{\partial}{\partial \theta} p(y|x) = p(y|x) \frac{\partial}{\partial \theta} l(y|x) - p(y|x) \sum_{y'} p(y'|x) \frac{\partial}{\partial \theta} l(y'|x) \]

\[ L(\hat{y}, \hat{x}) = l(\hat{y}|\hat{x}) - \log Z(\hat{x}) \]

\[ \frac{\partial}{\partial \theta} L(\hat{y}, \hat{x}) = \frac{\partial}{\partial \theta} l(\hat{y}|\hat{x}) - \sum_y p(y|\hat{x}) \frac{\partial}{\partial \theta} l(y|\hat{x}) \]

4.1 Regular old logistic regression

\[ l(y|x) = w^{(y)^T} x \]

\[ \frac{\partial}{\partial w^{(y')}} l(y|x) = \begin{cases} x, & y' = y \\ 0, & y' \neq y \end{cases} \]

4.2 One hidden layer

\[ l(y|x) = w^{(y)^T} \sigma(Vx) \]

\[ \frac{\partial}{\partial w^{(y')}} l(y|x) = \begin{cases} \sigma(Vx), & y' = y \\ 0, & y' \neq y \end{cases} \]

\[ \frac{\partial}{\partial V} l(y|x) = w^{(y)} \odot \sigma'(Vx)x^T \]

4.3 Two hidden layers

\[ l(y|x) = w^{(y)^T} \sigma(V\sigma(Ux)) \]

\[ \frac{\partial}{\partial w^{(y')}} l(y|x) = \begin{cases} \sigma(V\sigma(Ux)), & y' = y \\ 0, & y' \neq y \end{cases} \]

\[ \frac{\partial}{\partial V} l(y|x) = w^{(y)} \odot \sigma'(V\sigma(Ux))\sigma(Ux)^T \]

\[ \frac{\partial}{\partial U} l(y|x) = V^T [w^{(y)} \odot \sigma'(V\sigma(Ux))] \odot \sigma'(Ux)x^T. \]