Softmax

Dr. Sibt ul Hussain
Saif Ali Kheraj

August 18, 2019

Let us talk about Softmax. We will be using CIFAR dataset to illustrate softmax and its derivative.

$$\begin{bmatrix} w_{11} & w_{12} & ....w_{13073} \\ w_{21} & w_{22} & ....w_{23073} \\ w_{31} & w_{32} & ....w_{33073} \end{bmatrix} \cdot \begin{bmatrix} x_{11} \\ x_{12} \\ x_{13} \\ ... \\ x_{13073} \end{bmatrix}$$

We will use single example to illustrate and you can extend this concept to multiple examples

Weights: 3 x 3073
X: 3073 x 1

After taking dot product you will get matrix of dimensions 3x1 which represents unnormalised score if we talk in terms of softmax.

$$f = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

Note: You will first have to perform clipping to avoid numerical instability because if you take exponent of a number x it becomes large enough making it numerically unstable.

To normalise it we take anti log which is exponential.

$$\begin{bmatrix} exp(a) \\ exp(b) \\ exp(c) \end{bmatrix} / \sum exp(f)$$

This gives us normalised score. Score is interpreted as log probability for each class and softmax wants probability of correct class to be high. Higher score of correct class will reduce loss . You can check it by plotting log function, Well Log is monotonically increasing function thus we want to maximize the log likelihood or the other way if we take negative of log which then becomes minimization problem. We can do it both way. To maximize likelihood, we use gradient ascent and for minimization problem we use gradient descent.

We will use minimization approach for loss so our Loss function would be:

$$L_i = -log(P(y_i)) \tag{1}$$

Note: You can also derive this formula.
Let us expand this

$$\begin{gathered} L_i = -log((exp(f_y i)/\sum_j exp(f_{yj})) \\ L_i = -f_{yi} + log(\sum_j exp(f_{yj})) \end{gathered} \tag{2}$$

Note: We will now calculate derivate with respect to all Ws. Loss contains 2 parts one is data loss and the other is regularisation loss. Regularisation prevents overfitting and defuses weights. It is good for generalisation. You can add regularization to the loss function (L1 or L2) . Here we will use simple data loss and calculate its derivative.

We will now use above example to calculate derivative so we have 3x1 score Vector. Let us say class 1 is the correct score. We will plugin to the above Loss formula.

$$\begin{split} Loss = -(w_{11}x_{11} + w_{12}x_{12} + ... + w_{13073}x_{13073}) + log(exp(w_{11}x_{11} + w_{12}x_{12} + ... + w_{13073}x_{13073}) + \\ exp(w_{21}x_{11} + w_{22}x_{12} + ...w_{23073}x_{13073}) + exp(w_{31}x_{11} + w_{32}x_{12} + ...w_{33073}x_{13073})) \end{split} \tag{3}$$

This is now in expanded form which now becomes easy to calculate derivative with respect to all Ws.

Since class 1 is the correct class, we will first calculate derivate of w1

$$\begin{gathered} \frac{\partial L_i}{\partial w_{11}} = -x_{11} + (exp(f1)/\sum_j exp(f_{yj}).x_{11} \\ \frac{\partial L_i}{\partial w_{12}} = -x_{12} + (exp(f1)/\sum_j exp(f_{yj}).x_{12} \\ . \\ . \\ . \\ . \\ \frac{\partial L_i}{\partial w_{13073}} = -x_{13073} + (exp(f1)/\sum_j exp(f_{yj}).x_{13073} \end{gathered} \tag{4}$$

Let us now calculate derivative with respect to w2

$$\begin{aligned}
\frac{\partial L_i}{\partial w_{21}} &= (exp(f2)/\sum_j exp(f_{yj}).x_{11} \\
\frac{\partial L_i}{\partial w_{22}} &= (exp(f2)/\sum_j exp(f_{yj}).x_{12} \\
&. \\
&. \\
&. \\
&. \\
\frac{\partial L_i}{\partial w_{23073}} &= (exp(f2)/\sum_j exp(f_{yj}).x_{13073}
\end{aligned} \tag{5}$$

Note: You can now calculate derivative with respect to w3 easily. This is for single example x11, x12......x13073. You can use and extend this concept to Multiple examples. You