CS224d Assignment-1 Solutions

Raghav Chalapathy
Deep Learning for Natural Language Processing

July 1, 2016

1 SoftMax Function

In a multi-class classification problem, using support vector machines (SVM) produces classification score or confidence score for each of the class. Sometimes it is desirable to convert these confidence scores into probabilities which could be achieved using softmax functions.

In this section let us understand some of the important properties of softmax function and provide some proofs for the same

1.1 Properties of Softmax function

1.1.1 Property 1:

Prove that softmax is invariant to constant offsets in the input, that is, for any input vector x and any constant c, $softmax(x) = softmax(x+c)$ where $(x+c)$ means adding the constant c to every dimension of x.

Remember that $softmax(x)_i = \frac{e^{x_i}}{\Sigma_{j=1}^{N} e^{x_j}}$
for $j = 1, ..., N.$, where N represents the number of dimensions in the score set x_i.

Solution:

In practice, when computing softmax probabilities, in order to obtain numerical stability $c = -max(x_i)$ is choosen, where x_i is the score set. This constant c is subtracted from each of the elements in x_i.

$$softmax(x) = softmax(x + c)$$

Proof:

Given a N-dimensional vector x (x_j is the j-th component of x, and 1 ¡= j ¡= N):

$$softmax(x + c)_i = \frac{e^{x_i + c}}{\Sigma_{j=1}^{N} e^{x_j + c}} = \frac{e^{c} \bullet e^{x_i}}{\Sigma_{j=1}^{N} e^{c} \bullet e^{x_j}} = \frac{e^{c} \bullet e^{x_i}}{e^{c} \bullet \Sigma_{j=1}^{N} e^{x_j}} = \frac{e^{x_i}}{\Sigma_{j=1}^{N} e^{x_j}} = softmax(x)_i$$

Since we have shown that for all components j $softmax(x + c)_j = softmax(x)_j$, then it follows that $softmax(x + c) = softmax(x)$.

1.1.2 Property 2:

Let x_i be the score set and $S(x)_i$ denote the Softmax function for each element in x_i .

If we mulitply each element in set x_i by 10 then $S(x \bullet 10)_i$ the probabilities get close to either 1 **or** 0

If we divide each element in set x_i by 10 then $S(x \div 10)_i$ the probabilities get close to the *uniform distribution*