CS 224N: Assignment 1

October 22, 2018

1 Softmax (10 points)

(a) (5 points) Prove that softmax is invariant to constant offsets in the input, that is, for any input vector x and any constant c, Applying the law of total probability we have

$$softmax(\mathbf{x}) = softmax(\mathbf{x} + c)$$

where $x + c$ means adding the constant c to every dimension of x. Remember that

$$softmax(\mathbf{x})_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

Note: In practice, we make use of this property and choose c = maxi xi when computing softmax probabilities for numerical stability (i.e., subtracting its maximum element from all elements of x).

Solution:

$$softmax(\mathbf{x} + c)_i = \frac{e^{x_i + c}}{\sum_j e^{x_j + c}} = \frac{e^c \cdot e^{x_i}}{e^c \cdot \sum_j e^{x_j}} = \frac{e^{x_i}}{\sum_j e^{x_j}} = softmax(\mathbf{x})_i$$

2 Neural Network basics (30 points)

(a) (3 points) Derive the gradients of the sigmoid function and show that it can be rewritten as a function of the function value (i.e., in some expression where only $\sigma(x)$, but not x, is present). Assume that the input x is a scalar for this question. Recall, the sigmoid function is

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Solution:

$$\frac{\partial \sigma(x)}{\partial x} = \frac{e^{-x}}{(1 + e^{-x})^2} = \frac{1}{1 + e^{-x}} \cdot \frac{e^{-x}}{1 + e^{-x}} = \sigma(x) \cdot (1 - \sigma(x))$$

(b)(3 points) Derive the gradient with regard to the inputs of a softmax function when cross entropy loss is used for evaluation, i.e., find the gradients with respect to the softmax input vector θ, when the prediction is made by $\hat{y} = softmax(\theta)$. Remember the cross entropy function is

$$CE(y, \hat{y}) = -\sum_i y_i log(\hat{y}_i)$$

where y is the one-hot label vector, and ŷ is the predicted probability vector for all classes. (Hint: you might want to consider the fact many elements of y are zeros, and assume that only the k-th dimension of y is one.)

Solution:

$$\frac{\partial CE(y, \hat{y})}{\partial \theta} = \frac{\partial CE(y, \hat{y})}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \theta}$$

Derivative of the fiorst part is

$$\frac{\partial CE(y, \hat{y})}{\partial \hat{y}} = \frac{\partial(-y^T log\hat{y})}{\partial \hat{y}} = \frac{-y^T}{\hat{y}}$$

Derivative of the second part is

$$\frac{\partial \hat{y}}{\partial \theta} = \frac{\partial(\frac{e^\theta}{\sum_j e^\theta})}{\partial \theta} = \frac{e^\theta \cdot (\sum_j e^\theta - e^\theta)}{(\sum_j e^\theta)^2} = \hat{y} \cdot (1 - \hat{y})$$

Finally

$$\frac{\partial CE(y, \hat{y})}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \theta} = \frac{-y^T}{\hat{y}} \cdot \hat{y} \cdot (1 - \hat{y}) = \hat{y} \cdot (1 - \hat{y})$$