Deep Learning – Activation, Optimization

Summer 2023
Prof. Dr. Jochen Schmidt

Technische Hochschule Rosenheim

1) Activation Functions

Which of the following would you consider to be valid activation functions to train a neural net?

(1) $g(x) = 0.7x + 1$
(2) $g(x) = -\min(1, x)$
(3) $g(x) = \begin{cases} \min(x, 0.1x) & if\ x \geq 0 \\ \min(x, 0.1x) & if\ x < 0 \end{cases}$
(4) $g(x) = \begin{cases} \max(x, 0.1x) & if\ x \geq 0 \\ \min(x, 0.1x) & if\ x < 0 \end{cases}$

2) Loss-Functions

Many supermarket customers use the yellow creamy spot on the outside of a watermelon to evaluate its level of sweetness. To help customers who are not aware of this fact, you decide to build an image classifier to predict whether a watermelon is sweet (label=1) or not (label=0).

a) How many neurons and which activation function would be appropriate for the output layer?

b) You have built your own labeled dataset, chosen a neural network architecture, and are thinking about using the mean squared error (MSE) loss to optimize model parameters. Give one reason why MSE might not be a good choice for your loss function.

c) You decide to use the binary cross-entropy (BCE) loss to optimize your network. Write down the formula for this loss $L(y, \hat{y})$ (for a single sample) in terms of the label y and prediction $\hat{y}$.

d) You want to sanity-check your implementation of the BCE loss. What value does the loss take for a prediction of 0.9 on a negative (label = 0) example?

e) Consider the extreme cases for BCE for a single sample:
(1) You get a prediction of 0 and the actual class label is 0
(2) You get a prediction of 1 and the actual class label is 1
(3) You get a prediction of 1 and the actual class label is 0
(4) You get a prediction of 0 and the actual class label is 1
What values do you get for the loss?

f) Now we feed a mini batch of three samples through the network. The actual outputs are 0.1, 0.2, 0.8; the ground truth class labels are 1, 0, 0. What is the total loss averaged over all three samples?

g) You decide to use a single unified neural network to predict both the level of sweetness and the weight of a watermelon given an image.
(1) How many neurons and which activation function would be appropriate for the output layer?
(2) Propose a new loss function to train the unified model.

3) Activation Functions

You come across a nonlinear function that passes –1 if its input is negative, else evaluates to +1, i.e.

$$g(x) = \begin{cases} +1 & if\ x \geq 0 \\ -1 & if\ x < 0 \end{cases}$$

A friend recommends you use this non-linearity in your neural network with the Adam optimizer. Would you follow their advice?

Acknowledgements: Exercises based on course CS230 “Deep Learning”, Andrew Ng et al., Stanford University.