Universität des Saarlandes
Prof. Dr. Dietrich Klakow
Lehrstuhl für Signalverarbeitung
NNIA Winter Term 2018/2019

Exercise Sheet 9

Optimization

Deadline: 21.1.2019, 23:59

Exercise 9.1 - Dropout (3 + 0.5 + 1 + 0.5 + 1 = 6 points)

a) Given n random variables $X_i, i = 1, ..., n$ which are identically distributed with positive pairwise correlation ρ and $var(X_i) = \sigma^2$. Show that

$$var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \rho\sigma^2 + \frac{1-\rho}{n}\sigma^2$$

b) Explain why Bootstrap Aggregating (Bagging) can alleviate the effect of overfitting with the help of the formula above.

c) Why can dropout be considered as an approximation to Bagging? Explain in two or three sentences.

d) Do we apply dropout during the inference? Justify your answer.

e) In the lecture, we have been introduced to Dropout, one common way to implement that is inverted Dropout. Explain the idea behind inverted Dropout. [Hint: think about how dropout creates a mask drawn out of a Bernoulli distribution with probability p].

Exercise 9.2 - Stochastic Gradient Descent (0.5 + 1 + 0.5 = 2 points)

In practice, we have 3 common variations for Gradient Descent Method, namely Stochastic Gradient Descent (SGD), Batch Gradient Descent and Mini-Batch Gradient Descent. Stochastic Gradient Descent takes only one data point to update in each step. Batch Gradient Descent takes the whole dataset for updating in each step. Mini-Batch Gradient Descent compromises the two methods above and takes a subset of the whole dataset for updating in each step.

a) Is it necessary to use learning rate decay for batch gradient descent based learning to converge? Please give your reason.

b) Why is it important to use learning rate decay when doing stochastic gradient descent?

c) What is the advantage of using a small batch size instead of the full set of examples in the training data?

Exercise 9.3 - Vanishing and Exploding Gradient (1 + 1 = 2 points)

In the lecture we've discussed some of the challenges in Neural Network Optimization. One among them is the so-called Exploding/Vanishing gradient problem. To be more specific, this problem happens only during the backward pass in training (very deep) Neural Networks.

a) Assume that you have a 100-layer Feed Forward Neural Network with *sigmoid* activation function as non-linearities. Explain why the exploding and vanishing gradient problem occurs only during the backward pass but not the forward pass with formulas.

b) Explain how can we avoid the problem of gradient explosion based on what you learn in the lecture.

Submission instructions

The following instructions are mandatory. If you are not following them, tutors can decide to not correct your exercise.

- You **have to** submit the solutions of this assignment sheet as a team of 2-3 students.
- Hand in a **single** PDF file with your solutions.
- Therefore make sure to write the student ID and the name of each member of your team on your submission.
- Your assignment solution must be uploaded by only **one** of your team members to the course website.
- If you have any trouble with the submission, contact your tutor **before** the deadline.