Universität des Saarlandes
Prof. Dr. Dietrich Klakow
Lehrstuhl für Signalverarbeitung
NNTI Winter Term 2020/2021

Exercise Sheet 7

Regularization for Deep Learning

Deadline: 12.01.2021, 23:59

Exercise 7.1 - L_1 and L_2 Regularization (0.25+0.25+0.25+0.25+1 + 1 points)

In the lecture, you were introduced to different regularization methods such as L_1 and L_2 Regularization. Please answer the following questions.

a) Sketch the contour plot of L_1 and L_2 norms.

b) Why would one prefer to use L_1 regularization instead of L_2 regularization?

c) Is it possible to use a new regularization method that makes use of both of L_1 and L_2? Motivate your answer and if yes what it would be beneficial for?

d) Regularization is defined as any modification that we make to a learning algorithm that is intended to reduce its generalization error as well as its training error. Is this true? Motivate your answer.

e) Why doesn't L_2 -Regularization set variables to 0? Consider the case with L_2 -regularized least squares problem with 1 feature. Note that you should give mathematical justification.

f) Why does L_1 -Regularization set variables to 0? Consider the case with L_1 -regularized least squares problem with 1 feature. Note that you should give mathematical justification.

Exercise 7.2 - Multi-Task Learning and Data Augmentation (1.5 +1 points)

Multi-Task Learning and Data Augmentation are two important concepts in Deep Learning.

a) Read the following paper on Multi-Task Learning. Explicitly state **3 problems** and **3 solutions/methods** that the author pose related to information transfer in Multi-Task Learning. Your answer should not exceed more than 6-7 sentences. Reading the introduction (go simply through conceptually without focusing too much on the exact details) should be more than enough to give a clear view of the topic.

b) In the previous lecture you were shown a Pytorch sample code on the MNIST dataset. We have attached it again named as "multilayer_perceptron.py" file. Using that code please do the following data augmentation techniques.

- Add random noise (or any distribution that you think might be useful). Comment on the amount of noise and how it affects the results.
- Shift the images in all four directions by the given. Comment on the amount of shift that there is an improvement in accuracy (if there is an improvement).
- Pick another augmentation technique on your own and comment on how changing it can improve the results.
-

Please copy the code from the .py file given, edit it, and upload the solution as a Jupiter Notebook (It will not be graded if you simply give a .py file because it will give us difficulty in grading).

Exercise 7.3 - Early Stopping (1+1+1 points)

Consider a quadratic error function of the form

$$E = E_0 + \frac{1}{2}(w - w^*)^T H(w - w^*)$$

where w^* represents the minimum, and the Hessian matrix H is positive definite and constant. Suppose the initial weight vector $w^{(0)}$ is chosen to be at the origin and is updated using simple gradient descent

$$w^\tau = w^{\tau-1} - \epsilon \nabla E$$

where τ denotes the step number, and ϵ is the learning rate (which is assumed to be small). After τ steps, the components of the weight vector parallel to the eigenvectors of H can be written

$$w_j^\tau = \{1 - (1 - \epsilon\lambda_j)^\tau\} w_j^*$$

where $w_j = w^T v_j$ and v_j and λ_j are the eigenvectors and eigenvalues, respectively, of H.

a) Show that as $\tau \to \infty$, this gives $w_\tau \to w^*$ as expected, provided $|1 - \epsilon\lambda_j| < 1$.

b) Now suppose that training is halted after a finite number of τ steps. Show that the components of the weight vector parallel to the eigenvectors of the Hessian satisfy

$$w_j^\tau \simeq w_j^* \text{ when } \lambda_j \gg (\epsilon\tau)^{-1}$$

$$|w_j^\tau| \ll |w_j^*| \text{ when } \lambda_j \ll (\epsilon\tau)^{-1}$$

c) Show that $(\epsilon\tau)^{-1}$ is analogous to the regularization parameter α in L_2 Parameter Regularization.

Exercise 7.4 - Universal Approximation Theorem (0.5+0.5+0.5 points)

a) Please take a look at the Universal Approximation Theorem paper. Please concisely cite the main contribution/idea of the paper.

b) Indeed there is a previous work related to the representation capacity of neural networks. Why is this result weaker than the paper in Part a? What is the main idea of the paper?

c) State a reason and explain why, in practice, you would use deeper networks.

Submission instructions

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

- Make sure to write the Microsoft Teams user name, 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 'Assignments' tab of the tutorial team (in **Microsoft Teams**).
- Upload the solution as a zip file which contains the solution to the theoretical and programming part.
- If you have any trouble with the submission, contact your tutor **before** the deadline.