Questions and answers - Lukas Heinrich Lecture 1

The following questions were submitted through Google Form. Some may have been answered in the Q&A session already. Nevertheless, we request our lecturers to provide written answers here for the benefit of those who could not attend that session. Thank you!

Slide 62. What is the optimal distance between the validation loss and training loss curves ?

LH: there is no optimal distance, during underfitting the distance is low (but that's also not ideal, since that means your model doesn't have enough capacity. Generally a more should be able to memorize the training data is trained for a very long time

Slide not specified. If you have 3 input variables for the neural network with varying ranges some small some very large. How does the varying scale impact the neural networks ability to learn? Does it focus solely on the input variable with large values?

LH: It's beneficial if all input (and output) features are at a similar order, which is why people do data normalization (e.g. scikit-learns' standard scalar). Adaptive learning rate can mitigate a little but that does not absolve you from normalizing

Slide not specified. To normalize the data for my neural network, is it okay to divide by the maximum value in my dataset to get all the values in a range from 0-1? I work on ATLAS and I'm curious if dividing daughter particles by the mother would cause me to lose valuable information for the neural network

LH: data normalization is important but must be fixed on the training data and then frozen. That means on test data the range might not be any more [0,1]

Slide 41. I assume the Universal Function Approximation refers to one-dimensional functions here? How about the more realistic case of several dimensions?

LH: UFA holds for any R^n->R^m

Slide 51. Why is the "optimal hypothesis set size" slightly before the intersection before where the bias and variance curves intersect? How did you decide what the "optical hypothesis set size" is?

LH: the optimal hyposet size is when the total error is smallest depending on the curvature of the bias and variance curves this may be before or after the crossing point

Slide not specified. How large should your Validation- and Test-dataset be in comparison to the Training-data?

LH: no hard and fast rule, but 80%-10%-10% may be typical

Slide not specified. Presenter mentioned about some book suggesting that Learning has nothing to do with AI/ML. Can you please provide reference to that book?

LH: John Haugeland: Artificial Intelligence - The very Idea

11 Introduction

Box 2
Why Not Start with Learning?

Sometimes it seems that learning is to psychology what energy is to physics or reproduction to biology: not merely a central research topic, but a virtual definition of the domain. Just as physics is the study of energy transformations and biology is the study of self-reproducing organisms, so psychology is the study of systems that learn. If that were so, then the essential goal of AI should be to build systems that learn. In the meantime, such systems might offer a shortcut to artificial adults: systems with the ''raw aptitude'' of a child, for instance, could learn for themselves—from experience, books, and so on—and save AI the trouble of codifying mature common sense. But, in fact, AI more or less ignores learning. Why?

Learning is *acquisition* of knowledge, skills, etc. The issue is typically conceived as: given a system capable of knowing, how can we make it capable of acquiring? Or: starting from a static knower, how can we make an adaptable or educable knower? This tacitly assumes that knowing as such is straightforward and that acquiring or adapting it is the hard part; but that turns out to be false. AI has discovered that knowledge itself is extraordinarily complex and difficult to implement—so much so that even the general structure of a system with common sense is not yet clear. Accordingly, it's far from apparent *what* a learning system needs to acquire; hence the project of acquiring some can't get off the ground.[5]

In other words, Artificial Intelligence must start by trying to understand knowledge (and skills and whatever else is acquired) and then, on that basis, tackle learning. It may even happen that, once the fundamental structures are worked out, acquisition and adaptation will be comparatively easy to include. Certainly the ability to learn is essential to full intelligence; AI cannot succeed without it. But it does not appear that learning is the most basic problem, let alone a shortcut or a natural starting point.

Slide not specified. Is it fair to say that, in general, we can measure a variance (by looking the results from multiple training sets), but can't measure the true bias?

LH: yes that's right, without an exhaustive search you won't get the true bias, but often the "average" many training runs is close to the true best model

SLide titled "First it's very expensive". What I'm getting from this slide is that ML can help in efficiently generating simulations, there seems to be 2 problems: 1. One will need to do lots of simulations first to train the ML model 2. The model is a mimic of the simulator, and it will more or less include some variance. Can we really gain from ML in this simulations process?"

LH: there is no magic "new statistics" to be gained from fast simulation without inductive bias (i.e. constraining the hypothesis set), that's right. But a fast simulation may be "good enough" and that's sometimes a tradeoff worth making

Slide titled "Gradient descent". Why do we want to take bigger steps when the gradient is steepest? A priori, I would think that smaller steps at steeper gradients would be preferred.

LH: cannot find this slide: yes generally there is a term that goes with $1/\sqrt{g^2}$ i.e. you take bigger steps in flatter areas

Summary slide. Can you clarify when we would use both a validation and a test set? Do we only need a test set if we are using the validation set for decision making, instead of just for performance estimation?

LH: test set should be used exclusively once to get an unbiased estimate of the performance of your model. After this you should not go back to flip flop and pick a better model - that's what the valid set is for

Slide 26. What are the examples of exhaustive search and closed form solutions?

Linear Regression can be solved in closed form. Hyperparameter Optimization (optimizing e.g. the number of layers) often means optimizing among a finite set of models