IAML DL - Study Guide - Week 5

Sambit Paul, Pavlos Andreadis, Nigel Goddard

January 2022

1 Introduction

Week 5 introduces the concepts of Optimisation and Regularisation of machine learning models and why they are essential. Optimisation deals with the *'how'* the machine learning model learns, while Regularisation aims to solve the over-fitting problem.

2 Optimisation and Regularisation

2.1 Optimisation

- Optimisation is the process of using an error or cost metric to find the most optimal solution to a given problem. The process of optimisation can either be arithmetic, or it can be computed stochastically.
- In Machine Learning, we typically use an optimisation process in order to solve the problem of which model parameter assignments make our model best fit the training data (indicated by which parameter assignment minimises our error metric). The problem is known as *model fitting*. Of course there is more nuance here, as we do not typically want the model to perfectly fit our training data (consider Generalisation, and Regularisation below). To read more about the different methods for model fitting, Murphy [2012] Section 8.3 gives a brief introduction to various methods.
- An error manifold is the space of all possible values to your objective function in the optimisation problem (assuming you have a criterion that needs to be minimised, such as a metric of how wrong your predictions with the model are). Given a specific set of samples on which we are comparing, the error manifold will depend on the different possible values our model parameters can take. A brief explanation of error manifolds is provided in this Jupyter Notebook.
- *Cost function*, *loss function*, *error function*, and *objective function* are synonyms, though the first 3 imply that we are dealing with a minimisation problem, while *objective function* is more general. Typically we are

looking to define a cost function which we need to minimise. Refer to Goodfellow et al. [2016] Section 4.3 to understand the basic idea behind convex optimisation.

- To understand and know more about different loss functions in brevity, you can refer to this article. Also, try understanding how each loss function is suited to classification and regression problems. Generally, you might want to introduce further terms to your loss function; this is an engineering problem in and of itself (for example, see Regularisation below).
- Gradient descent has been well explained in Goodfellow et al. [2016] Section 5.2.4 and the key equation behind gradient based optimisation is given in Equation 5.41.
- Learning rate is a hard parameter to tune, and Murphy [2012] Section 8.3.2 on Page 247 aims to explain how it works.
-
  - Batch gradient descent computes the gradient using the whole dataset.
  - Stochastic gradient descent (SGD) computes the gradient using a single sample randomly chosen over each epoch from the dataset.
  - Best of both worlds, is to use a mini-batch (set of samples chosen randomly from the dataset) and perform Batch descent over that mini-batch.

2.2 Regularisation

- The basic idea of regularisation is well articulated in Barber [2012] where he explains it as, "For most purposes, our interest is not just to find the function that best fits the train data but one that will generalise well. To control the complexity of the fitted function we may add an extra regularising term to the train error to penalise rapid changes in the output".
- To go through the method of finding weights when L2 regularisation is used in linear regression, you might want to use Bishop [2006] Pg 144. The equations is $w = (\Phi^T.\Phi + \lambda.I)^{-1}.\Phi^T.\mathbf{y}$. This is *Ridge Regression*.
- A clear explanation of L2 Regularisation (Ridge Regression) is given in Goodfellow et al. [2016] Section 7.1.1.
- Additionally, another important form of regularisation is L1 regularisation or what is known as *Lasso Regression*. It is useful to find sparse matrices of weights. lease use Goodfellow et al. [2016] Section 7.1.2 to read how it works.
- To have a simplified understanding of L1 regularisation, this article is quite useful.

References

David Barber. *Bayesian reasoning and machine learning.* Cambridge University Press, 2012.

Christopher M Bishop. *Pattern recognition and machine learning.* Springer Science+ Business Media, 2006.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning.* MIT press, 2016.

Kevin P Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.