IAML - Study Guide - Week 4

Sambit Paul, Pavlos Andreadis, Nigel Goddard

January 2022

1 Introduction

Week 4 introduces the concepts of *regression* in the context of machine learning. Regression is a statistical measure to compute the strength of relationship between the observed (independent) variable and the resultant (dependent) variable. Following that, we will start with *Linear Regression* which is used to model linear parametric models to determine relation in the form $y = mx + c$. Finally, we will study about *Logistic Regression* which is used for classification of linearly separable data. Simple version of Logistic Regression is used for classification between 2 categories, however it can be extended to support multi-category classification.

A brief refresher on Bayesian concepts can be found in this article.

2 Linear Regression

- For understanding the mathematics behind linear regression, use Goodfellow et al. [2016] Pg 106. It uses the formula $\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{estimate} - y_{actual})_i^2$ and then takes the differential over the weights to find the minima and solves for the weights.
- 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.
- It might also be worthwhile looking over linear algebra methods to solve matrix equations using LU Decomposition Strang [2016] Pg 97 and Cholesky Decomposition Deisenroth et al. [2019] Pg 114 and understanding when and why either might be used. For a quick introduction to Cholesky Decomposition, you can use this article. An insight into the use of Cholesky Decomposition for numerically stable computations to compute the weights for MLE can be found in Murphy [2012] Pg 229.
- Good resource for understanding basis functions is provided in the MLPR notes for Week 1B Section 2.2
- What is **One Hot Encoding**?

For categorical data, numerical stamps like 1, 2, 3... can be used but the numbers have an innate relationship between them (2 is two times 1, 3 is three times as 1). Algorithms may be able to harness this relationship and learn latent properties which were not meant to be learnt.

Hence, we go for one-hot encoding where the categories are represented as booleans (yes/no). So, categories will be represented as:

- 1 : 100
- 2 : 010
- 3 : 001

So on, and so forth.

- Why use **Radial Basis Functions**?
  - For $X.W$, we get a linear response along the preferred direction $\vec{W}$.
  - For RBF $[exp(-|x-\mu|^2)]$, we get local units and can use use the distance measure to compute degree of match. Used for locally weighted systems (LOESS).

  Read more about how RBFs can be used in computational networks, use this article.

  You can delve deeper into RBFs using this article Buhmann [2010].

3 Logistic Regression

- Classification problems can be resolved in three approaches as noted below:
  - Generative Models which models the distribution of the outputs and inputs.
  - Discriminative Models which models the posterior probability directly.
  - Using discriminant functions to map input directly to class.

  Further details can be found in Bishop [2006] Section 1.5.4.

- Logistic Regression is an extension of Linear Regression to allow for classification. We need to convert the continuous output of the Linear parametric model to discrete values representing classes.

- One of the key elements of classification is the ability to convert results from linear parametric model [Section 2] to discrete values. Function which do this are called Discriminant Functions. To read more about them, refer to Bishop [2006] Pg 181.

- A brief introduction to what a decision boundary means in the mathematical sense and how logistic regression is related to linear regression is provided in the book Barber [2012] Pg 363.
- A good resource to understand the math behind the cost function (Log Likelihood) is provided in the MLPR notes for Week 6B Sections 2 and 3.
- How to fit **Logistic Regression** model?

  There are multiple methods to fit logistic regression models, the most popular ones being:

  - *Maximum Likelihood Estimation*
  - *Gradient Descent*

  Both these methods and many others are discussed in details which you can read about in Murphy [2012] Pg 249.
- To read about multi-class classification, you can refer to Barber [2012] Section 17.4.4 (Pg 372).

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.

M. Buhmann. Radial basis function. *Scholarpedia*, 5(5):9837, 2010. doi: 10.4249/scholarpedia.9837. revision #137035.

Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. *Mathematics for Machine Learning.* Cambridge University Press, 2019.

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

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

Gilbert Strang. *Introduction to linear algebra.* Cambridge Press, Wellesley, MA, 2016. ISBN 978-09802327-7-6.