IAML - Study Guide - Week 1

Sambit Paul, Pavlos Andreadis, Nigel Goddard

January 2022

1 Overview

Week 1 introduces the course and focuses on refreshing prerequisite knowledge. Learn provides a number of prerecorded lectures on Maths and Probability, while we also point to some useful readings below. Lab 0 is a good refresher on Python. We advise that you attempt Lab 1, though some points in it will become clearer in week 2 when the lab is run.

You will also be learning how to *pre-process* (i.e. 'prepare') your data for using them to train Machine Learning models.

Introductory Week 1 class meetings are to be held on (all UK time and on Blackboard Learn), see the Course Schedule on Learn for the times, which will be set after the poll is run. No preparation is needed for this session.

We will to provide weekly study guides, such as this one, which you can refer to while studying. It will provide references to portions of the reading list which elucidate specific topics and areas very well.

2 Introduction to Machine Learning

The Introductions subsection in Learn's "Course Materials" section includes an overview of the course, relevant applications, and an aside on bias. There is a reference here to a **"W&F" book which you can ignore** (there are some references to this in the slides, but it is not part of the reading list).

For an introduction to the concept of Machine Learning, we would recommend Chapter 1 of Géron [2017]. This might also be a good chance to get a quick look at the content of different books in the reading list, and contrast their approach to the subject.

3 Mathematical Preliminaries

The course assumes that you have prerequisite knowledge in Linear Algebra, Calculus, and Probability. These are not examinable on their own, but an

understanding of the relevant concepts is required.

3.1 Linear Algebra

Methods like Linear/Logistic Regression and SVMs use a lot of linear algebra (logarithms, vectors, matrices, etc) while processes like PCA and ICA (for dimensionality reduction) also use eigen values, ranks etc. which again falls under linear algebra.

For a thorough explanation of major linear algebra topics that we will be using, please skim through Chapters 1 through 7 from Strang [2016]. Students who just need a quick refresher can read Chapter 2 from Goodfellow et al. [2016].

3.2 Probability

Other than that, machine learning also relies heavily on probability. These includes theories and applications of conditional probability, Bayes' rule. Other than that, different kinds of distributions are key in building generative models and also different methods of model fitting. Few of the distributions that you need to be aware of are *Uniform Distribution*, *Gaussian Distribution*, *Bernoulli Distribution* and *Binomial Distribution*.

For students who need a deeper explanation of concepts, you can read through Chapter 3 of Goodfellow et al. [2016]. Students who just need a quick refresher may read Chapter 1 from Barber [2012].

3.3 Calculus

Calculus comes into play when we start moving into the realms of gradient-based optimisation. Unlike the problems which have a closed form solution, most real world problems are not such. For them, gradient based optimisation is usually the means to a solution.

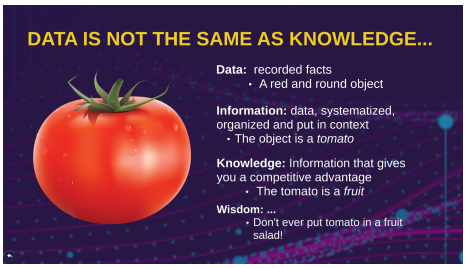
To get a refresher on the calculus required for this course, you can refer to Appendix 2 (Multivariate Calculus) of Barber [2012].

4 Dealing with data

- All machine learning tasks require using data, which the machine can "learn" representations of. This data can be of 3 types:
  - **Categorical**: Data which can be grouped into specific categories. Example: music genre.
  - **Ordinal**: The data has natural, ordered categories and the distances between the categories is not known. Example: job title.
  - **Numerical**: The data that is measurable. Example: height, weight.

This article provides a succinct explanation of the different data types that you will usually see in ML applications.

- Data Normalisation is a key aspect in data preparation. Normalisation involves converting data under a specific feature into a common scale using the mean and the standard deviation. This article presents data normalisation and how it improves ML models very articulately.
- Often in datasets used in the real world (uncurated), certain values might be missed. These need to be dealt with before using the data on any machine learning models. Witten et al. [2011] Section 2.4 provides a good insight into dealing with missing data.
- Data science is the analysis of data such that we can understand the viability of the data and then transform and use it to guide decisions. The raw data that is collected if often not as valuable. Converting the data to information makes it more useful (often, labelling is a means of converting data to information). Using this information, we can start machine learning to generate knowledge out of it.



- Machine Learning can be broken down into 4 basic types:
  1. **Supervised**: Supervised learning describes a class of problem that involves using a model to learn a mapping between input examples and the target variable. The objective is to generate a generalised representation of the input-target mapping.
  2. **Unsupervised**: Unsupervised learning describes a class of problems that involves using a model to describe or extract relationships in data. This is usually done based on the attributes of the data itself and does not require labels.
  3. **Semi-supervised**: Semi-supervised learning is supervised learning where the training data contains very few labeled examples and a large number of unlabeled examples. Usually unsupervised methods are employed to group data and then learning happens on clustered data using the labelled data.
  4. **Reinforcement Learning**: Reinforcement learning describes a class of problems where an agent operates in an environment and must learn to operate using feedback. The learning happens based on re-

wards and punishment for actions it takes in certain scenarios.

- There are two types of modeling approaches used in machine learning: *Generative* approach and *Discriminative* approach. Barber [2012] Section 13.2.3 has a very clear description of the two approaches.

References

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

Aurélien Géron. *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.", 2017.

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

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

Ian H Witten, Eibe Frank, and Mark A Hall. Data mining: Practical machine learning tools and techniques, 2011.