IAML DL - Study Guide - Week 03

Sambit Paul, Pavlos Andreadis

January 2020

1 Introduction

Week 3 of the course introduces our first discriminative classification method called *Decision Trees*. Decision Trees can be used for both classification and regression problems. The objective of these binary trees are to categorise data based on features of the data. This requires some understanding of entropy and probability which you can refresh using this video.

The next section that would be covered this week is the *generalisation* of models over varied data and the *evaluation* of models. This involves using techniques to improve the overall reliability of our models and assessing the performance.

2 Decision Trees

- The basic working of decision trees can be understood using these two videos. They provide an intuitive basis for setting up the foundations for dealing with classification and regression problems.
  - Decision Tree for Classification
  - Decision Tree for Regression
- For an introduction to the evolution of the different tree algorithms, you can use this article. One of the first ones to be used for decision making is called **ID3** (Iterative Dichotomiser 3) followed by different iterations called **C4.5** and **C5.0**. Other than this, we also use **CaRT** (Classification and Regression Trees) algorithm for decision trees.
- A key aspect of decision trees is the determination at each split, which features can be used for the split. This can be measured using *the purity of the split*.

  To measure the purity of the split, we need to compute the entropy of the

split based on various features using the formula:

$$E = \sum_{i=1}^{n} p_i \times log_2(p_i)$$

where,
p_i is the probability of getting category i
n is the number of categories

- Over-fitting is a problem in machine learning where the model learns to recognise known data very well, but for unseen data, is very inaccurate. This often happens in case of very granular level of splitting (cases in which each leaf node holds only one example). The various ways of dealing with over-fitting:

  1. Stop splitting when relative entropy change between parent and child node is statistically insignificant
  2. Using a validation set
  3. Tree pruning

- **Random Forests**:
  Random forests form a family of methods that consist of building an ensemble (or forest) of decision trees grown from a randomized variant of the tree induction algorithm. Tree induction algorithms state how the splits happen at nodes and are driven by hyper-parameters like loss function, splitting criterion etc. Decision trees are ideal candidates for ensemble methods since they usually have low bias and high variance, making them very likely to benefit from the averaging process Louppe [2014].

- Reading List:
  Bishop [2006] pp. 663 - 666
  Witten et al. [2011] pp. 70 - 71, 105 - 113, see index at pp. 606 for a list of relevant topics (topic is more complex than it looks)
  Hastie et al. [2009] pp. 305 - 317

3 Generalisation & Evaluation

- Generalisation refers to the reliability of the model's outcomes. Simply put, the more generalised a model is, the more it can correctly predict unseen data.
  To understand the concept of generalisability, the bias-variance trade-off is a key concept that needs to be clarified. You can refer to this article for a quick refresher. Hastie et al. [2009] Chapter 7, Sections 7.2 and 7.3 provides a good explanation of the bias-variance relationship.

- **Over-fitting** & **Under-fitting**: When a model is not well generalised, it is implicitly understood that the model is either over-fitted to the training data, or under-fitted to the problem. A visual representation is provided in Figure 1
  - Over-fitting: Fits very well to the training data, but performs poorly on unseen data. Usually happens when the model has high variance and low bias.
  - Under-fitting: Cannot fit well to the training data itself. Usually happens when the model has high bias and low variance.

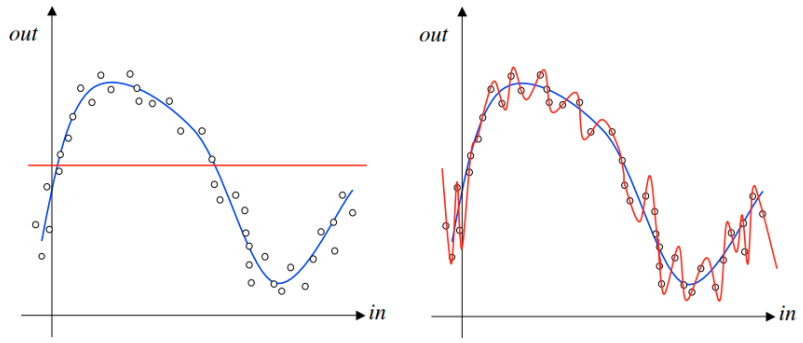

Figure 1: The left plot shows (in red) how an under-fitted model would perform, and the right plot shows (in red) how an over-fitted model would perform. The line in blue is the representation of the function.

- The dataset used for machine learning problems is split for 3 purposed:
  - Training set: Used to construct the model.
  - Validation set: Used to determine the hyperparameters of the model.
  - Test set: Used to estimate the generalisability of the model.

  Please refer to Hastie et al. [2009] Chapter 7, Page 222 to read about dataset splitting. In case of problems, which are not data-rich, we use cross validation which you can read about Hastie et al. [2009] Section 7.10.
- Model evaluation is done in the form of metrics. To know more about such metrics, you can refer to this article.
- Reading List:
  Goodfellow et al. [2016] pp. 105 - 117
  Witten et al. [2011] pp. 31 - 35

References

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.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Gilles Louppe. Understanding random forests: From theory to practice, 2014.

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