IAML - Study Guide - Week 2

Sambit Paul, Pavlos Andreadis, Nigel Goddard

January 2022

1 Introduction

On Week 2 of the course we will learn about *Naive Bayes*, which is often used as a *baseline* model across applications. A baseline model being a model that is easy to set up (often with no learning involved) and is used as to perform a 'sanity' check on your experiment results.

Next we introduce 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.

2 Naive Bayes

- Bayes' rule is a probabilistic method to update our belief about a certain variable provided more evidence. It is given by this equation:

$$P(E_2|E_1) = \frac{P(E_1|E_2) \times P(E_2)}{P(E_1)} \tag{1}$$

  where,

  - E_1 and E_2 are two mutually exclusive events
  - $P(E_2|E_1)$ is the posterior probability
  - $P(E_1|E_2)$ is the likelihood
  - $P(E_2)$ is the prior probability
  - $P(E_1)$ is the marginal probability

  To understand the Bayes' rule in more detail, you can use Bishop [2006] Section 1.2.3.

- Conditional Independence is a concept which states that for 2 events **A** and **B** to be conditionally independent given an event **C**, knowledge of whether

A occurs provides no information on the likelihood of **B** occurring, and knowledge of whether **B** occurs provides no information on the likelihood of **A** occurring. This is explained on a more mathematical basis on Barber [2012] Section 1.1.1 under definitions 1.6 and 1.7.

- The idea behind naive Bayes classification is to model the joint distribution of an event belonging to a class and the features related to that event. So, in the form of an equation, we can say:

$$P(y|x_1, x_2...x_n) = \frac{P(x_1, x_2...x_n|y) \times P(y)}{P(x)}$$
$$\Rightarrow P(y|x_1, x_2...x_n) = \frac{P(y) \times \prod_{i=1}^{n} P(x_i|y)}{P(x)}$$

  So, what we are fundamentally modelling is: $P(y) \times \prod_{i=1}^{n} P(x_i|y)$.

  This video provides a very good intuition for Naive Bayes' Classification. For further reading, the course textbook Barber [2012] Chapter 10 contains all the information you will need for a thorough understanding of the Naive Bayes' model.

3 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

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.

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.