IAML DL - Study Guide - Week 02

Sambit Paul, Pavlos Andreadis

January 2020

1 Introduction

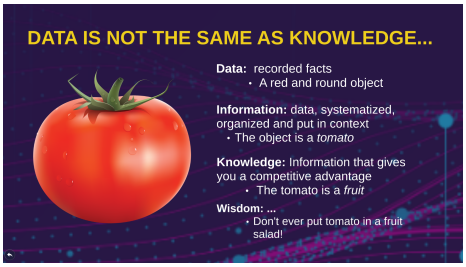
On Week 2 of the course you will first be learning how to *pre-process* (i.e. 'prepare') your data for using them to train Machine Learning models. Following that, we will learn of *Naive Bayes*, which is often used as a *baseline* model across applications. A baseline model being a model that is easy to set up (often with no learning involved) and is used as to perform a 'sanity' check on your experiment results.

2 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 rewards 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.

3 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.

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