IAML DL - Study Guide - Week 7

Sambit Paul, Pavlos Andreadis, Nigel Goddard

January 2022

1 Introduction

Week 7 starts with introducing the Nearest Neighbours method for classification. This involves using a distance metric to determine clusters of training points and determine classes based on which cluster the new data point falls in. A practical introduction to both supervised and unsupervised method can be found in this article.
We also introduce 2 unsupervised learning methods for clustering of data.
The first method being explored in K-means which aims to cluster data into K groups by minimizing a criterion known as *inertia*. K is a parameter that needs to be chosen as a parameter before execution by the user.
Along side that, we will also explore Gaussian mixture models (GMMs) which are a generalisation of K-means to incorporate covariance information Pedregosa et al. [2011]. This model uses a combination of Gaussian distributions to model the data.

2 Nearest Neighbours

- Nearest Neighbours algorithm works under the principle of "*similar things exist in close proximity*".
- A basic mathematical intuition is given in Hastie et al. [2009] Section 2.3.2 where they have given examples of using N-neighbours for creating decision boundaries based on local clustering of data.
- *Voronoi Tessellation*: The smaller the number of neighbours for clustering, the more granularly the decision boundaries are fragmented. For very small numbers like 1-2, this fragmentation is called Voronoi tessellation. You can read more about this in this article.
- To understand few of the issues that Kearest Neighbours method suffers from, you can refer to Barber [2012] Section 14.1 (Pg. 317 - Pg. 318).
-
    - Larger value of K, means all points may be classified as the class with more data points.

  - Smaller value of K, means the model is not generalisable and can cause large fluctuations for small changes in the data.

  The choice of the number of neighbours to use (**K**) can be identified using validation. This can be considered parameter tuning for a machine learning model.

- One of the key differences between a linear decision boundary built using methods like SVM based on least-squares method and nearest neighbours based decision boundary is the fact that there is an underlying assumption about the data distribution being linearly separable. To read further on this, please refer to Hastie et al. [2009] Section 2.3.3.

- Please refer to this video to understand more about Kernels and Parzen Windows.

- To know more about the ongoing research on kNNs and how they are improving upon the existing method, you can refer to Zhang et al. [2017] and Wu et al. [2008] Section 8.4.

- - *KD trees* can be considered a combination of decision trees and kNN algorithm in which each split in the tree is based on the median value of a specific feature. Each leaf node contains "k" points against which the nearest neighbours calculation can be done.
    To know more about this, please refer to Section 6.3 and Section 6.4 of this article
  - This article provides a clear and succinct explanation of *Locality-Sensitive Hashing*. You can refer to this paper Zhang et al. [2013] to understand how LSH improves the efficiency of kNNs. The abstract of the paper gives a clear overview of the idea, but to get a deeper understanding, it might be useful to refer to Algorithm 2 in the paper.
  - This video provides a good explanation of inverted list based nearest neighbours search.

- For a probabilistic perspective on nearest neighbours, a very succinct explanation is provided in Barber [2012] Section 14.3.

3 K-Means Clustering

- Why is it called **K-Means**?
  In K-Means the term K refers to the number of clusters that need to be identified; and, *means* refers to the process of averaging of data to find the centroid of each cluster.

- **Monothetic** and **Polythetic Clustering**: In a monothetic scheme, cluster membership is based on the presence or absence of a single characteristic. Polythetic schemes use more than one characteristic. For example, classifying people solely on the basis of their gender is a monothetic classification, but if both gender and handedness (left or right handed) are used, the classification is polythetic.

- To read about hard and soft clustering, please refer to this article.
- The objective of K-means as defined in Bishop [2006] Section 9.1 is the minimisation of the cost function J where $J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{n,k} ||x_n - \mu_k||^2$ such that, $r_{n,k}$ denotes if point n belongs to cluster k and $||x_n - \mu_k||^2$ is the squared error.
- To understand the K-means algorithm, please refer to Wu et al. [2008] Section 2.1. The basic steps can be elucidated as:

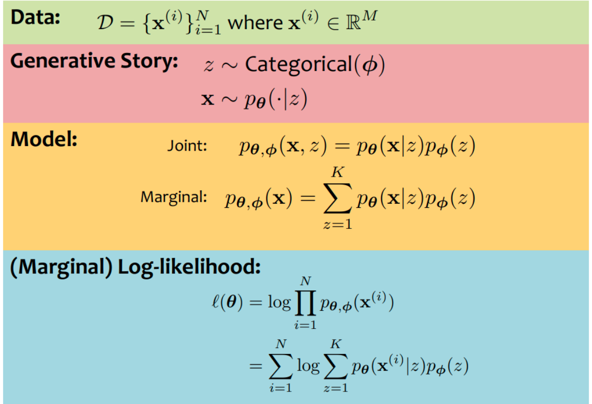
  1. Specify number of clusters K.
  2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
  3. Keep iterating until there is no change to the centroids or maximum iterations has been reached.
- An improvement on the basic K-means algorithm is to introduce a kernel on top of the data to project it into a high-dimensional space Dhillon et al. [2004]. Although the boundaries will be linear in the high-dimensional space, on projecting back to the lower dimensions, it becomes non-linear.
- To read about the limitations of K-means, please refer to Wu et al. [2008] Section 2.2.
- To get a quick overview of the K-means algorithm, please refer to Barber [2012] Section 20.3.5. [Requires an understanding of Expectation Maximization]

4 Gaussian Mixture Models

- This topic requires an intuition about Maximum Likelihood Estimation. To get a quick refresher, please refer to this article.
- What is **Expectation-Maximization**?
  Expectation maximization is an iterative process of improving the probability of a model to predict if an observation belongs to a specific distribution in the presence of latent variables.
  - E-Step ⇒ Estimate the missing variables in the dataset
  - M-Step ⇒ Maximize the parameters of the model in the presence of the data

Maximum Likelihood estimate the same probability in the absence of latent variables.

- This can be used good starter video to understand the intuition about Expectation-Maximization (EM).
- To get a deeper understanding of the mathematics behind the general EM algorithm, please refer to Bishop [2006] Section 9.4. Another approach to EM, based on mathematical derivations, is provided in Section 2 of this document.
- Basic Representation of Mixture Models is provided in Figure 1



Data: $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ where $\mathbf{x}^{(i)} \in \mathbb{R}^M$

Generative Story: $z \sim \text{Categorical}(\boldsymbol{\phi})$

$\mathbf{x} \sim p_{\boldsymbol{\theta}}(\cdot|z)$

Model:

Joint: $p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}, z) = p_{\boldsymbol{\theta}}(\mathbf{x}|z)p_{\boldsymbol{\phi}}(z)$

Marginal: $p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) = \sum_{z=1}^{K} p_{\boldsymbol{\theta}}(\mathbf{x}|z)p_{\boldsymbol{\phi}}(z)$

(Marginal) Log-likelihood:

$$\ell(\boldsymbol{\theta}) = \log \prod_{i=1}^{N} p_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}^{(i)})$$

$$= \sum_{i=1}^{N} \log \sum_{z=1}^{K} p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|z)p_{\boldsymbol{\phi}}(z)$$

Figure 1: These are the basic steps that need to be followed to build a Mixture Model

- An intuitive concept of Gaussian Mixture model is provided in this article
- Section 2 and 3 from this document provides an elaborate explanation of Gaussian Mixture models and Expectation Maximization.
- A thorough and clear explanation of Gaussian Mixture Models (albeit, slightly lengthy) is also provided in Bishop [2006] Section 9.2.

5 Comparison between K-means and GMM

Criterion	**K-Means**	**GMM**
Convergence	Faster than GMM	Slower than K-Means
Speed	Computationally less intensive	Computationally intensive
Initialization	Random Initialisation	Use K-means to determine the means of the Gaussian
Output	Single hard assignment to clusters	Probability distribution over the cluster assignment

Table 1: This table provides a comparative analysis of K-Means clustering and Gaussian Mixture Models over 4 criteria

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.

Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: spectral clustering and normalized cuts. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 551–556, 2004.

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

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, S Yu Philip, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14 (1):1–37, 2008.

Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Ruili Wang. Efficient knn classification with different numbers of nearest neighbors. *IEEE transactions on neural networks and learning systems*, 29(5):1774–1785, 2017.

Yan-Ming Zhang, Kaizhu Huang, Guanggang Geng, and Cheng-Lin Liu. Fast knn graph construction with locality sensitive hashing. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 660–674. Springer, 2013.