IAML DL - Study Guide - Week 06

Pavlos Andreadis, Sambit Paul

November 2019

1 Introduction

Week 5 continues with the concepts of Support Vector Machines introducing overlapping class distributions and how to modify the algorithm to deal with them and make them more robust. We will also deal with the usage of kernels to create non-linear SVM classifiers. This article provides a high level understanding of the importance of kernels in the field of machine learning

The next section introduces 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.

2 Support Vector Machines 2

- Derivation of optimal parameters using Lagrange multipliers:

  $g(x) = |w|^2 \Rightarrow \frac{dg}{dw} = 2|w|$
  $f(x) = \sum\{y_i(w^T x_i + w_0) - 1\} \Rightarrow \frac{df}{dw} = \sum y_i x_i$

  Using Lagrange Multiplier, we can say:
  $g(x) = \lambda f(x)$
  $\Rightarrow 2|w| = \sum \lambda_i y_i x_i$
  $\Rightarrow |w| = \sum \alpha_i y_i x_i$ assuming $\alpha_i = \frac{\lambda_i}{2}$

  This concept has also been explained thoroughly in Bishop [2006]. Please refer to Section 7.1 Pg 328.

- For linearly non-separable data, creating a solution which gives an exact separation will not be generalisable. This requires the need to allow some data points to be misclassified. Please refer to Bishop [2006] Section 7.1.1 for more details on this.

- For a quicker introduction to the use of ξ_n for making SVMs more robust, please refer to Section 17.5.1 from Barber [2012].
- To understand the influence of the parameter C in SVM classification, this StackOverflow article provides a very good explanation.
- Following the 2-Norm Soft-margin subsection under Section 17.5.1 from Barber [2012], it will be clear that the optimisation problem requires only the inner product. A simpler derivation for the optimisation equation is provided here:
  $f(x) = \frac{1}{2}w^T w$ and $g(x) = y_n(w^T x_n + w_0) - 1$ Hence, using Lagrange multipliers, we can say that the optimisation problem is:
  $L(w, w_0) = f(x) + \sum \alpha_n g_n(w, w_0)$
  This implies,
  $\frac{dL}{dw} = w - \sum_n \alpha_n y_n x_n = 0$ and $\frac{dL}{dw_0} = 0 - \sum_n \alpha_n y_n = 0$
  Filling in the values in the original optimisation equation, we get:
  $L(w, w_0) = \sum_n \alpha_n - \frac{1}{2}w^T w \Rightarrow L(w, w_0) = \sum_n \alpha_n - \frac{1}{2}\sum_{n,m} \alpha_m \alpha_n y_n x_n^T x_m y_m$
  Hence, the optimisation equation depends on the $x_n^T x_m$ which is basically an inner product. If x is replaces with the basis function, we get $\phi(x_n)^T \phi(x_m)$. This can be represented as $K(x_n, x_m)$ and are called kernel functions. Kernel function basically calculate the inner product in the transformed space.
- The conditions to determine which functions can be considered as Kernel functions is defined using Mercer's theorem.

3 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.

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.

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.