IAML DL - Study Guide - Week 6

Sambit Paul, Pavlos Andreadis, Nigel Goddard

January 2022

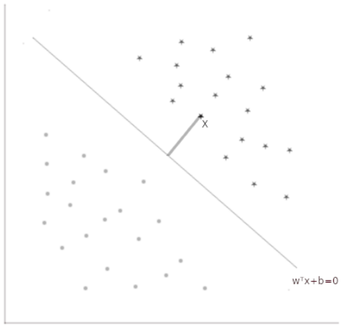
1 Introduction

Week 6 introduces one of the most popular machine learning models - Support Vector Machines (SVMs). As stated by Andrew NG in this article, SVMs are among the best "off-the-shelf" supervised learning algorithm in use.

Alongside, we will also cover an examination of ethical issues in machine learning research and practice. For more material, see these class notes by Shannon Vallor.

2 Support Vector Machines 1

- The intuition is to learn a decision boundary (hyperplane) to separate classes of data such that margin between the training points on either classes is maximised.
- **Intuition behind SVM:**
  For a binary classification problem with 2 classes (say, +1 and -1), we need to:
  - Predict +1 iff $w^T x + b0$
    If $w^T x + b \gg 0$, confidence of class = +1 is very high.
  - Predict -1 iff $w^T x + b < 0$
    If $w^T x + b \ll 0$, confidence of class = -1 is very high.
- **Derivation of Geometric Margin of SVM:**



Let us consider the distance between the training sample X and the decision boundary $w^T x + b = 0$ as γ. Since the *weight vector* **w** is orthogonal to the decision boundary, the projection of the point X on the decision boundary can be written as $(x - \gamma.\frac{w}{w})$. Let this point be called $\hat{X}$.
Therefore, if $\hat{X}$ is put in the equation of the decision boundary, we should have $w^T \hat{X} + b = 0$. Therefore, we get the equation:

$$w^T(x - \gamma.\frac{w}{w}) + b = 0 \tag{1}$$

Solving Equation 1 for γ, we get: $\gamma = \frac{w}{w}\hat{X} + \frac{b}{w}$ This is called the geometric margin of point X with respect to the decision boundary. The division of the hyperplane by the Euclidean norm of the weight vector is to ensure scaling the weights does not affect the margin.

- To understand why weights are always orthogonal to the decision boundary, please take a look at this article.
- To understand how to the math for maximising the margin works, please refer to Barber [2012] Section 17.5.1 and for an more in-depth explanation please refer to, Bishop [2006] Section 7.1.

3 Support Vector Machines 2

Part 2 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

- 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.

4 Ethics and Machine Learning

- **General ethical concerns**

  Data technology is not neutral - it has design choices baked in, and use is a choice too. A key perspective is that ethics in machine learning is a process not a checklist. It is an ongoing conversation amongst stakeholders, including you. Key questions to ask are:

  - Who are the stakeholders?
  - Who benefits and how?
  - Who could be harmed and how?

We can divide the general ethical concerns into three categories: benefits, harms and challenges.

- Benefits - want to maximise these
  - Increase human understanding of nature, society and our personal lives
  - Increased social, institutional and economic efficiency
  - Accurate prediction and personalization
- Harms - want to avoid these
  - Harms to fairness and justice
  - Harms to transparency and accountability
  - Harms to privacy and security
- Data challenges - want to address these
  - Appropriate collection and use
  - Data stewardship
  - Data cleanliness and relevance
  - Ethically harmful data bias
  - Validation/test of models and analytics
  - Human accountability in data systems
  - Understanding personal, social and business impacts of data practices

- **Fairness**

  People should not be discriminated against or disparately impacted based on their membership of a protected group or class, such as race, gender, sexual orientation, *etc.* There can be bias in data collection, bias in data labelling, and outcome bias. Using protected characteristics in models is generally a bad idea; and simply not doing that is not sufficient, as there may be other features that are correlated with a protected characteristic. One way to assess fairness in classification systems is to compare ROC curves for different groups of interest (e.g., partitioned by protected characteristic) - if these differ significantly there may be a problem. How to address it is usually context dependent, and it can be the case that achieving fairness leads to a decrease in overall performance.

- **Accountability**

  Human decision makers can be asked to account for their decisions, including moral and ethical choices. These do not apply to ML models, but they do apply to the designers, implementors, and users of the models.

Some design choices can facilitate accountability - e.g. a Decision Tree is human readable with specific choice points, but a deep neural network may have no accessible human interpretation.

- **Transparency**

  There are several aspects including outcome transparancey - knowing how and why a model produced the outcome it did; and process transparency - being open about the processes and choices that went into building the 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.