Abstract

The capacity of a learning machine is characterized by various measures such as the Vapnik-Chervonenkis (VC) dimension, Rademacher complexity, and covering numbers. All these measures define a notion of complexity or capacity of a learning machine in either how well it can fit a random labelling of data points or how many samples it can classify with all possible labelings of the data points. Several generalization bounds are proposed in the literature using these measures. Large capacity can lead to overfitting, while a small one can lead to under-fitting. Researchers have worked on developing several algorithms and paradigms to find the right balance of empirical error and capacity. One such work in this direction is the Minimal Complexity Machine (MCM). The MCM was formulated as a way to tightly bound the VC dimension. The MCM was shown to outperform the Support Vector Machine (SVM) in generating sparse solutions and resulting in good generalization properties. In this thesis, we extend the MCM and develop several algorithms to scale up the MCM to large datasets while maintaining sparsity, lower complexity, and having a small generalization error. In doing so, we identified four major problems with the state-of-the-art SVM and the least squares SVM (LS-SVM). First, they have a large space complexity. Second, they do not generate highly sparse solutions. Third, they cannot use indefinite kernels and last they are not suitable for IoT devices due to large model size even after quantization.

We develop solutions to the problems mentioned above around the recently proposed MCM. We initially develop a Stochastic Gradient Descent (SGD) variant of the MCM that can scale to large datasets via the use of explicit feature maps. However, the use of explicit feature maps does not provide the information about the points that make up the decision boundary and hence, we develop algorithms in the Empirical Feature Space (EFS). The EFS is defined to be all linear combinations of vectors in the kernel or succinctly, the span of the kernel. We define the MCM in the EFS, the least squares MCM and their respective input space margin maximization variants. The advantage of using EFS is that the VC dimension of a classifier in the EFS is bounded by the number of training samples. We use prototype vector selection of data points in scaling the EFS variants to large datasets.

The EFS variants developed use One-vs-One (OVO) classification scheme for multiclass classification scheme, which results in slower training and inference time for datasets with a large number of classes. To tackle this issue, we develop novel multiclass MCM variants along with their least squares and margin maximization in EFS. We show that these variants offer the advantage of faster training and prediction time while maintaining sparsity and accuracies higher than the OVO variants. The next part of the thesis focuses on extending the multiclass MCM variants to using...
indefinite kernels. The EFS formulations allow for the use of indefinite kernels in the optimization problem of the multiclass MCM. We transition from the Hilbert spaces to Banach and Krein spaces to present analysis for the indefinite kernel multiclass MCM and the least squares MCM. We demonstrate that indefinite kernel MCM variants have similar generalization properties as the positive definite kernel MCM variants.

In the last part of the thesis, we solve the final issue identified by us in extending the algorithms to memory constraint IoT devices. We present weight quantization comparison of the multiclass MCM and the SVM variants such as the least squares SVM and the LIBSVM. We evaluate all our algorithms on 13 benchmark datasets, spanning small and large datasets both in terms of the number of samples and classes. We show that the proposed multiclass MCM variants can retain accuracies even using 3 bits as compared to 12 bits of used by the SVM and the LS-SVM.