REVIEW

Support vector machines (SVM) are a recent addition to the set of machine learning techniques available. The SVM is based on the theory of structural risk minimization and as such has good generalisation properties that have been demonstrated both theoretically and empirically. The flexibility of the SVM is provided by the use of kernel functions that implicitly map the data to a higher, possibly infinite, dimensional space. A solution linear in the features in the higher dimensional space corresponds to a solution non-linear in the original features. This flexibility does not however lead to overfitting since the SVM performs automatic capacity control. A remaining problem with SVMs is the selection of the kernel and regularization parameters. The commonly used Gaussian kernel benefits from a good choice of the scaling parameter σ. This is usually chosen by training an SVM for a range of values of σ and using an estimate of the generalization error to do model selection. The standard approach of validation becomes impractical for large data sets as SVMs can perform inefficiently. An idea used to initialize σ and this is then used to update σ during training. On benchmark data sets, this algorithm achieves a cross-validated error not significantly different from the optimal value. It is also more computationally efficient than a line search for σ.

The convolution of the resulting hypothesis is controlled by tuning the kernel. With the help of validation set we obtain its value and its choice amounts to selecting the model. To select the model an algorithm is devised in which we do not need validation set but with little additional computational cost. Here, we do not have separate learning and model selection but kernels are adjusted progressively during the learning process and choose the kernel value which provides the best possible upper bound on the generalization error.

Pros

Training of model is relatively easy.

Trade-off amongst classifier complexity and error can be controlled explicitly.

SVM is effective in high dimensions space.

We are updating kernel during the learning process thus SVM is a good option since it is memory efficient.

The clear margin of separation between classes makes it work relatively well.

Cons

We may face problem since SVM algorithm is not suitable for large datasets.

Since it works by putting data points above and below the classifying hyperplane there is no probabilistic explanation for the classification.

The basic limitation is that this method can hardly cope up with the dynamic environment where data changes with time. So, we need an algorithm which takes care of data in the dynamic environment.

This problem can be resolved by using hybrid system for model selection in SVM where two or more techniques can be combined.

Error

By moving to a high dimensional space, we are incurring a penalty on sample complexity.

Model selection

Standard way is to learn the model and test them on a validation set to obtain the optimal value of the kernel parameter. This is a consuming method. In this paper, the approach uses training data making it more efficient.

Experimental results

The Kernel-Adatron (KA) algorithm was recently brought up by two authors and used to train SV machines. For computing the bound, $\epsilon \leq \mathcal{R}^2/m\gamma^2$ radius of the ball needs to be estimated in the feature space. Using convex quadratic programming routines, this could be done explicitly by maximizing the Langrangian with respect to λi provided some constraints.

$$R = \sum_{i,j} \lambda i \lambda j K(xi, xj) - 2 \sum_{i,j} \lambda j K(xi, xj) + \sum_{i} K(xi, xj)$$

Besides this can also be calculated by noting that Gaussian kernels always map training points to the surface of a sphere of radius one centered on the origin of the feature space. We can easily see from the statement that distance of a point from the origin is its norm:

$$|(|\phi(x)|)| = \sqrt{(< \phi(x), \phi(x) >)} = \sqrt{\left(K(x,x)\right)} = \sqrt{\left(e^{\left(|(|x-x|)|/(2\sigma^2)\right)}\right)} = 1$$

A little additional computational cost is needed for determining R from the quadratic programming problem, at least for Gaussian kernels.

Now we plot the bound $\sum_i \alpha i/m$ and generalisation error for two figures from a United States Postal Services dataset of handwritten digits.

It was investigated that the minimum of the bound approximately coincides with the minimum of the generalisation error. This way an optimal σ can be chosen. Also, this estimate of σ is derived only from training data without the need of additional validation set.

Thus, an algorithm is devised which automatically learns the kernel parameter with little additional cost, both in computational and sample-complexity sense. The process of model selection is carried on during the learning phase itself. To provide a good estimate of the correct model complexity, experimental results are provided.