

# Design and Analysis of Local Kernel Machines

Ravi S. Ganti

Alexander G. Gray\*

October 26, 2009

## 1 Introduction

Given a set of  $N$  points  $S = \{(x_i, y_i), \dots, (x_N, y_N)\} \in \mathbb{R}^d \times \{0, 1\}$  sampled from an unknown distribution  $\mathcal{D}$  the problem of classification is to learn a hypothesis  $h : X \rightarrow \{0, 1\}$  which has a small value for  $\mathbb{P}_{(x,y) \sim \mathcal{D}} [h(x) \neq y]$ . Classification is undoubtedly one of the most fundamental problems in machine learning and a plethora of algorithms such as non-parametric Bayes classifier (NBC), decision trees (DT), nearest neighbour based classification (NNC), non-linear/kernelized support vector machines or simply called support vector machine (SVM) have been designed for this problem. While algorithms such as NBC ( $O(N)$ ), DT ( $O(N \log(N))$ ) and nearest neighbour (NN) based classifiers ( $O(N \log(N))$ ) using tree based methods are scalable to large datasets, SVM's [1] are known to be theoretically superior to them as they attain Bayes error (smallest possible error) asymptotically, escape the curse of dimensionality which is known to plague non-parametric estimators such as NBC, NN, DT and also have demonstrated superior empirical performance on a wide variety of classification problems. SVM's work by finding a separating hyperplane in the RKHS of a kernel function  $K(\cdot, \cdot)$  by solving a quadratic programming problem (QP) that tradesoff complexity of the fit to the goodness-of-the-fit to the data. While the convexity of the optimization problem makes it amenable to finding a global optimum solution in  $O(N^3)$  using interior point methods, it also hampers the scalability of SVM's to large scale data-sets. In order to counter scalability issues of SVM's a number of algorithms have been proposed. One popularly used algorithm is the SMO algorithm [2].

*Our goal is to achieve SVM like statistical performance as well as scalability.* Motivated by local methods in the problem of regression such as local linear regression [3], we designed local kernel machines. In local linear regression a non linear curve is fit to the data by making local linear fits to different parts of space. The idea behind this technique is the fact that by Taylor's theorem any smooth function can be approximated by a linear function in a small interval. Carrying this idea to classification means that we now fit a non-linear decision surface by a smooth decision surface that can be locally well approximated by a linear decision surface. Hence we now solve a large number of smaller and much less expensive local linear SVM (LLSVM) problems instead of a single, large SVM problem. The LLSVM problem works by using a smoothing kernel  $k(\cdot, \cdot, \cdot)$  (say Epanechnikov kernel) and solves the following dual convex QP at every new test point  $x_0$ :

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j k(x_i, x_0, \sigma_s) k(x_j, x_0, \sigma) \langle x_i, x_j \rangle - \sum_{i=1}^n \alpha_i k(x_i, x_0, \sigma_s) \quad (1)$$

subject to:

$$w_{x_0} = \sum_{i=1}^N \alpha_i y_i k(x_i, x_0, \sigma_s) x_i, \quad \sum_{i=1}^N \alpha_i y_i K(x_i, x_0, \sigma_s) = 0, \quad 0 \leq \alpha_i \leq \frac{1}{N \lambda_N}. \quad (2)$$

## 2 Preliminary Results

**Theorem 1 (Consistency of LLSVM)** *The LLSVM is consistent if  $N \rightarrow \infty$ ,  $\sigma \rightarrow 0$ ,  $\lambda_N \rightarrow 0$  and  $N \lambda_N \sigma^d \rightarrow \infty$ .*

---

\*advisor/mentor

**Theorem 2 (Risk bound for LLSVM)** Let  $\phi(\cdot)$  be a  $L$  Lipschitz convex loss function and  $R = \mathbb{E}_{(x,y) \sim \mathcal{D}}$ , and  $\hat{R} = \frac{1}{n} \sum_{i=1}^N \phi(y_i w_{x_i} \cdot x_i)$  where  $w_{x_i}$  is the LLSVM learnt at  $x_i$  by solving the optimization problem (1-2). We have

$$R \leq \hat{R} + \frac{4L^2}{n\lambda_N\sigma^d} + \left[ \frac{9L^2}{\lambda_n\sigma^d} + \phi(0) \right] \sqrt{\frac{\log(\frac{1}{\delta})}{2N}}.$$

### 3 Goals of the project

The idea of LLSVM opens up a lot of interesting theoretical issues which are worth investigating. While statistical literature is pregnant with results for non-parametric estimators [4] and also with results that explain the good high dimensional performance of kernel machines [5, 6, 1, 7], we are not aware of work which involves both these approaches. While the risk bound in theorem (2) investigates the LLSVM from the point of view of kernel machines, it should also be possible to obtain more results for the LLSVM by viewing it purely as a non-parametric estimator and using the whole arsenal of non-parametric statistics. e.g. it would be really interesting to see a bias-variance decomposition for the LLSVM estimator. Such a decomposition will help resolve the question if LLSVM's escapes the curse of dimensionality unlike other non-parametric estimators. Another interesting idea worth investigating is a kernelized local SVM (KLSVM) where for each new point  $x_0$  a local (the extent of the locality being decided by the smoothing kernel) kernelized SVM problem is solved. It would be extremely interesting to see how these two sources of non linearity namely the bandwidth ( $\sigma_s$ ) of the smoothing kernel  $k(\cdot, \cdot, \cdot)$  and the bandwidth ( $\sigma_r$ ) of the reproducing kernel  $K$  will affect the risk bounds of the KLSVM. The local nature of the LLSVM allows us to work with data where global euclidean distances are not reliable but local euclidean distances are. This is particularly true with data lying on manifolds. Even for such data we expect LLSVM to work well. It would be nice to exploit differential geometric aspects of the manifold to show advantage(s) of an LLSVM over SVM's.

### 4 Conclusions

To conclude we expect LLSVM and KLSVM to give state-of-the-art classification performance and scalability on a variety of real world datasets and open up a whole body of work relating kernel machines and non-parametric estimators. We plan to work with our astrophysics collaborators on quasar identification problems in astronomy. We also plan to draw upon the expertise of Prof Vladimir Koltchinskii and Prof Nina Balcan to help us in tackling the various issues that we have proposed to attack in this proposal.

### References

- [1] Bernhard Schoelkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.
- [2] J. Platt. Using sparseness and analytic QP to speed training of support vector machines. *Advances in NIPS*, 1999.
- [3] C. Loader. *Local regression and likelihood*. Springer Verlag, 1999.
- [4] A.B. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2008.
- [5] V. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.
- [6] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(3):463–482, 2003.
- [7] V. Koltchinskii. Rejoinder: Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6):2697–2706, 2006.