Degeneracy in Model Selection for SVMs with Radial Gaussian Kernel

Tobias Glasmachers

Institut für Neuroinformatik
Ruhr-Universität Bochum - Germany

Abstract. We consider the model selection problem for support vector machines applied to binary classification. As the data generating process is unknown, we have to rely on heuristics as model section criteria. In this study, we analyze the behavior of two criteria, radius margin quotient and kernel polarization, applied to SVMs with radial Gaussian kernel. We proof necessary and sufficient conditions for local optima at the boundary of the kernel parameter space in the limit of arbitrarily narrow kernels. The theorems show that multi-modality of the model selection objectives can arise due to insignificant properties of the training dataset.

1 Introduction

Assume we are given ℓ training examples $(x_i, y_i) \in X \times Y$, where X is the input space and $Y = \{-1, +1\}$ is the space of class labels. The examples are assumed to be drawn i.i.d. from an unknown probability distribution ν on $X \times Y$. The binary classification task is to construct a predictive classifier $c : X \to Y$ from these examples, that is a classifier which minimizes a risk functional, usually the probability of misclassifying a pattern sampled from ν. As the distribution ν itself is unknown, we can not minimize this risk directly.

The support vector machine (SVM) has become a standard tool for this task, see [1, 2, 3] for an introduction. It requires a positive definite kernel function $k : X \times X \to \mathbb{R}$. This kernel implicitly defines a map $\Phi : X \to \mathcal{H}$ into a feature Hilbert space $\mathcal{H}$ where $\langle \Phi(x), \Phi(z) \rangle = k(x, z)$. The SVM constructs an affine linear function $f : \mathcal{H} \to \mathbb{R}$. The class labels of points $x \in X$ are predicted as $c(x) = \text{sign}(f(x))$. The minimum distance in $\mathcal{H}$ of the correctly classified training examples from the set $\{f = 0\}$, known as the separating hyperplane, is called the (geometric) margin of the classifier. In the C-SVM setting the affine linear function optimizes an objective which is a trade off between margin maximization and model complexity reduction, controlled by a parameter $C > 0$. Following [4], we consider C as a kernel parameter within the 2-norm SVM framework.

The SVM exhibits high classification accuracy and allows for the incorporation of prior knowledge about the problem at hand via the kernel function. However, its performance crucially depends on the choice of a kernel that makes it easy to separate the data linearly. The choice of the kernel function is the model selection problem for the support vector machine.

In practice prior knowledge may lead to a parameterized family of kernel functions. Then the model selection problem is reduced to a finite dimensional

search space. For at most two or three dimensional parameter manifolds it is possible to examine the objective landscape via nested grid search. Otherwise more sophisticated techniques such as gradient descent or evolutionary algorithms are the methods of choice.

2 Model Selection Objectives

The goal of model selection is to ensure a good generalization performance of the classifier, that is, a minimal risk. This performance can be estimated on a separate dataset not used for training, or by cross-validation. We can define other quality measures for candidate models, for example if we need continuous or differentiable objectives for optimization. These should be roughly monotonic in the generalization ability of the classifier with high probability. Furthermore it is important to ensure that optimization algorithms (direct search or gradient descent algorithms) do not get stuck in local extrema. Therefore it is desirable that objective functions tends to smoothen the generalization error landscape. Ideally, they should not possess multiple local optima.

For the (kernel based) support vector machine, a variety of learning theoretical generalization error bounds have been proven. These bounds have been proposed for model selection [4]. For example, the minimization of the radius margin quotient $(R/\gamma)^2$ is a useful objective. Here, R denotes the radius of the smallest (closed) ball in feature space containing all training examples and γ is the margin by which the SVM separates the classes.

Another approach is to maximize the kernel polarization [5]

$$P = \sum_{i,j=1}^{\ell} y_i y_j K_{ij}$$

or the scaling invariant kernel target alignment $\hat{A} = P/\sqrt{\sum_{i,j=1}^{\ell}(K_{ij})^2}$ introduced in [6], both measuring the capability of the kernel function to separate the classes on the examples.

The kernel allows for the computation of inner products between pairs of training examples in $\mathcal{H}$. Usually, this is the only computation which is affordable in the feature space, and thus model selection algorithms are restricted to the information represented in the positive definite symmetric kernel (Gram) matrix

$$K = \big(k(x_i, x_j)\big)_{1 \le i,j \le \ell} \ .$$

Hence we consider quantities which depend on the training data only through the kernel matrix K. This condition is fulfilled for the model selection objective functions presented above. Each of them can be written as a function of the kernel matrix. In case of the radius margin quotient the functional relation is only implicitly available as the computation of R and γ each requires the solution of an optimization problem depending on K.

3 Properties of the Radial Gaussian kernel

In the following we consider the radial Gaussian kernel[1]

$$k_\sigma(x,z) = \exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right) + \delta_{x,z}\frac{1}{C} \tag{1}$$

operating on the input space $X = \mathbb{R}^n$. Although in this formulation the SVM parameter C is considered a kernel parameter, we will concentrate on the adaptation of the single parameter $\sigma > 0$. The theoretical analysis will hold for any fixed value of C. The derivative of the Kernel (1) with respect to σ is

$$k'_\sigma(x,z) = \frac{\partial}{\partial\sigma}k_\sigma(x,z) = \frac{\|x-z\|^2}{\sigma^3}\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right) \quad .$$

First it becomes clear from definition (1) that all training feature vectors are of the same length in $\mathcal{H}$, as it holds $\|\Phi_\sigma(x)\|^2 = \langle\Phi_\sigma(x), \Phi_\sigma(x)\rangle = k_\sigma(x,x) = 1 + 1/C$. Thus, the diagonal entries of K do not depend on σ. This is not the case for the off-diagonal entries. For every pair of different training examples $x \neq z$ it holds $k_\sigma(x,z) \in (0,1)$. As σ decreases, the feature vectors $\Phi_\sigma(x)$ and $\Phi_\sigma(z)$ become more and more orthogonal (that is, if σ is divided by $\sqrt{2}$ the kernel values are squared and thus become smaller). Let us now consider two pairs of training examples $x_1 \neq z_1$ and $x_2 \neq z_2$ fulfilling the strict[2] inequality $\|x_1 - z_1\| < \|x_2 - z_2\|$. It follows $k_\sigma(x_1,z_1) > k_\sigma(x_2,z_2)$ for all $\sigma > 0$. An interesting property to observe here is

$$1 > \frac{k_\sigma(x_2,z_2)}{k_\sigma(x_1,z_1)} = \left(\frac{k_1(x_2,z_2)}{k_1(x_1,z_1)}\right)^{\frac{1}{\sigma^2}} \quad , \tag{2}$$

that is by decreasing σ the quotient becomes arbitrarily small. It is important to note that not only the absolute values of the off-diagonal entries of K decrease during this process, but that they become more and more different in the sense that their quotients decay (or explode if the pairs are switched).

We can compute a similar quotient using the derivative of the kernel function instead of the kernel function itself:

$$1 > \frac{k'_\sigma(x_2,z_2)}{k'_\sigma(x_1,z_1)} = \left(\frac{k'_1(x_2,z_2)}{k'_1(x_1,z_1)}\right)^{\frac{1}{\sigma^2}} \cdot \left(\frac{\|x_1-z_1\|^2}{\|x_2-z_2\|^2}\right)^{\frac{1}{\sigma^2}-1} \quad . \tag{3}$$

This quotient decays even faster for $\sigma \to 0$.

We want to reserve the indices p and q for the pair of training examples with minimum input space distance, that is

$$(p,q) = \underset{(i,j)\text{ s.t. } i<j}{\operatorname{argmin}} \|x_i - x_j\| \quad .$$

[1] We use the Kronecker delta notation $\delta_{a,b} = 1$ if $a = b$ and $\delta_{a,b} = 0$ otherwise.

[2] For simplicity we assume $\|x_1 - z_1\| \neq \|x_2 - z_2\|$.

To stress the dependency of the model selection objectives from the finite number of kernel matrix entries depending on σ, we use the matrix notation

$$K(\sigma) = \big(k_\sigma(x_i, x_j)\big)_{1 \le i,j \le \ell} \quad \text{and} \quad K'(\sigma) = \big(k'_\sigma(x_i, x_j)\big)_{1 \le i,j \le \ell} \ .$$

It is clear from equations (2) and (3) that these matrices become degenerate in the limit case $\sigma \to 0$. The diagonal entries are fixed to $K_{ii}(\sigma) = 1 + 1/C$ and $K'_{ii}(\sigma) = 0$, while the off-diagonal entries quickly decay to zero. Among the off-diagonal entries $K_{pq}(\sigma) = K_{qp}(\sigma)$ and $K'_{pq}(\sigma) = K'_{qp}(\sigma)$ become arbitrarily dominating over all other entries. This is formalized in the following lemma:

Lemma 1. *For the quotients of off-diagonal entries of K and K' it holds*

$$\lim_{\sigma \to 0} \sum_{\substack{1 \le i < j \le \ell \\ (i,j) \neq (p,q)}} K_{ij}(\sigma) \Big/ K_{pq}(\sigma) = 0 \quad \textit{and} \quad \lim_{\sigma \to 0} \sum_{\substack{1 \le i < j \le \ell \\ (i,j) \neq (p,q)}} K'_{ij}(\sigma) \Big/ K'_{pq}(\sigma) = 0 \ .$$

Proof. The lemma is a direct consequence of equations (2) and (3). □

4 Degeneracy Theorems

As a consequence of the degeneracy of the kernel matrix and its derivative we show that both radius margin quotient $(R/\gamma)^2$ and kernel polarization P are governed by the pair of labels (y_p, y_q) if the kernel parameter σ becomes small. For the analysis of the radius margin quotient we introduce the notation $\ell_+ = |\{i \,|\, y_i = +1\}|$ and $\ell_- = |\{i \,|\, y_i = -1\}|$.

Theorem 2. *Under the conditions $\ell_+ > 0$, $\ell_- > 0$ and $\ell > 2$ the radius margin quotient R^2/γ^2 has a local optimum (minimum) at the boundary $\sigma \to 0$ if and only if $y_p \neq y_q$.*

Proof. We will compute the derivative of R^2/γ^2 for $\sigma \to 0$. The radius R can be obtained from the solution β^* of the quadratic problem

$$R^2 = \max_\beta \left(\sum_{i=1}^{\ell} \beta_i K_{ii}(\sigma) - \sum_{i,j=1}^{\ell} \beta_i \beta_j K_{ij}(\sigma) \right)$$

under the constraints $\sum_{i=1}^{\ell} \beta_i = 1$ and $\forall\, i = 1, ..., \ell : \beta_i \ge 0$. In the limit $\sigma \to 0$ we have $K_{ij}(\sigma) = (1 + 1/C)\delta_{ij}$. The resulting objective and the constraints depend on all β_i equally. Thus all β_i^* take on the same value leaving only a one dimensional problem open. From the equality constraint we obtain the feasible solution $\beta_i^* = 1/\ell$, from which we compute $R^2 = (1 + 1/C)(1 - 1/\ell)$.

The maximum α^* of the dual SVM objective

$$\sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K_{ij}(\sigma) \tag{4}$$

under the equality constraint $\sum_{i=1}^{\ell} y_i \alpha_i = 0$ and the inequality constraints $\forall\, i = 1, ..., \ell : \alpha_i \geq 0$ defines the margin by $1/\gamma^2 = \sum_{i,j=1}^{\ell} \alpha_i^* \alpha_j^* y_i y_j K_{ij}(\sigma)$. After plugging $K_{ij}(\sigma) = (1+1/C)\delta_{ij}$ into equation (4), the resulting problem depends equally on all α_i having the same label y_i. From this observation we conclude that the solution is of the form $\alpha_i^* = \tilde{\alpha}_{y_i}$. The two dimensional problem is then reduced to one dimension by the equality constraint and can be solved setting the derivative of the objective (4) to zero. We get the feasible solution

$$\alpha_i^* = \begin{cases} \tilde{\alpha}_{(+1)} = \frac{2C}{1+C}\frac{\ell_-}{\ell} & \text{if } y_i = +1 \\ \tilde{\alpha}_{(-1)} = \frac{2C}{1+C}\frac{\ell_+}{\ell} & \text{if } y_i = -1 \end{cases} \qquad \text{and thus} \qquad 1/\gamma^2 = \frac{4}{1+1/C} \cdot \frac{\ell_+ \ell_-}{\ell} \ .$$

We observe that all training examples are support vectors ($\forall\, i = 1, ..., \ell : \alpha_i^* > 0$). Following [4], the derivative of the radius margin quotient can be computed as

$$\begin{aligned} \frac{\partial}{\partial \sigma}\frac{R^2}{\gamma^2} =& R^2 \left(-\sum_{i,j=1}^{\ell} \alpha_i^* \alpha_j^* y_i y_j K'_{ij}(\sigma) \right) + \frac{1}{\gamma^2} \left(\sum_{i=1}^{\ell} \beta_i^* K'_{ii}(\sigma) - \sum_{i,j=1}^{\ell} \beta_i^* \beta_j^* K'_{ij}(\sigma) \right) \\ =& -2 \sum_{1 \leq i < j \leq \ell} \left(\alpha_i^* \alpha_j^* R^2 y_i y_j + \beta_i^* \beta_j^* / \gamma^2 \right) \cdot K'_{ij}(\sigma) \ . \end{aligned}$$

From Lemma 1 we know that in the limit $\sigma \to 0$ this sum is governed by the term

$$\alpha_p^* \alpha_q^* R^2 y_p y_q + \beta_p^* \beta_q^* / \gamma^2 \ . \tag{5}$$

From the prerequisites it follows $(\ell - 1)(\min(\ell_+, \ell_-))^2 > \ell_+ \ell_-$. Together with the limits of α^* and β^* computed above, we get the inequality

$$\alpha_p^* \alpha_q^* R^2 \ \geq \ \frac{4(\ell-1)(\min(\ell_+, \ell_-))^2}{(1+1/C)\ell^3} \ > \ \frac{4\ell_+\ell_-}{(1+1/C)\ell^3} = \beta_p^* \beta_q^* / \gamma^2$$

which shows that the left summand of expression (5) dominates the right one. Thus the derivative of R^2/γ^2 has the same sign as $-y_p y_q$. At the boundary $\sigma \to 0$, a positive (negative) derivative indicates a minimum (maximum). □

Theorem 3. *The kernel polarization P has a local optimum (maximum) at the boundary $\sigma \to 0$ if and only if $y_p \neq y_q$.*

Proof. We compute the derivative

$$\frac{\partial}{\partial \sigma} P = \sum_{i,j=1}^{\ell} y_i y_j K'_{ij}(\sigma) = 2 \cdot \sum_{1 \leq i < j \leq \ell} y_i y_j K'_{ij}(\sigma)$$

and observe from Lemma 1 that for small σ the term is governed by $y_p y_q K'_{pq}(\sigma)$. It is negative (positive) if and only if the labels y_p and y_q differ (are equal), indicating a maximum (minimum) at the boundary $\sigma \to 0$. □

Although the kernel target alignment $\hat{A}$ differs from kernel polarization only by a normalization term, its analysis is more complicated.

The theorems show that near the boundary $\sigma \to 0$ the term $y_p y_q$ controls the path taken by a gradient descent algorithm. It is clear that this term does not represent much information about the underlying distribution ν. At least for noisy datasets it is a highly random quantity. Further the proofs of the theorems show that lots of local optima may exist near the boundary, governed by label combinations of proximate training examples.

5 Conclusion

In practice, bounds derived from statistical learning theory as well as measures like kernel polarization work well for SVM model selection. Here we want to sensitize for the fact that these functions are in general multi-modal. As parameter search is usually carried out over several orders of magnitude, boundary extrema may play a role. It is possible to sail around the arising difficulties easily. We recommend to use a heuristic as proposed in [7] for an initial choice of the scaling parameter σ of the Gaussian kernel. With this initialization a search algorithm should avoid the boundary optima with very high probability.

We know that our proofs poorly catch the true multi-modality of model selection objectives. Besides the boundary optima computed in this study, useful objective functions can exhibit lots of local optima far from the boundary, see for example [8]. This fact justifies the application of heuristics to the model selection problem.

References

[1] V. N. Vapnik. *Statistical Learning Theory*. Wiley, New-York, 1998.

[2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[3] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002.

[4] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002.

[5] Y. Baram. Learning by kernel polarization. *Neural Computation*, 17(6):1264–1275, 2005.

[6] N. Cristianini, A. Elisseeff, J. Shawe-Taylor, and J. Kandola. On Kernel-Target Alignment. In *Neural Information Processing Systems*, pages 367–373. MIT Press, 2001.

[7] T. Jaakkola, M. Diekhaus, and D. Haussler. Using the fisher kernel method to detect remote protein homologies. *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 149–158, 1999.

[8] F. Friedrichs and C. Igel. Evolutionary tuning of multiple SVM parameters. *Neurocomputing*, 64(C):107–117, 2005.