ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

One-class SVM regularization path and comparison with alpha seeding

Alain Rakotomamonjy[1] and Manuel Davy[2]

1- Laboratoire ITIS EA 4051, Université de Rouen
2- LAGIS/CNRS/INRIA FUTURS, Cité Scientifique, BP 48

Abstract. One-class support vector machines (1-SVMs) estimate the level set of the underlying density observed data. Aside the kernel selection issue, one difficulty concerns the choice of the 'level' parameter. In this paper, following the work by Hastie et. al (2004), we derive the entire regularization path for ν-1-SVMs. Since this regularization path is efficient for building different level sets estimate, we have empirically compared such approach to state of the art approach based on alpha seeding and we show that regularization path is far more efficient.

1 Introduction

One-class support vector machines (1-SVM) [1] are used in a variety of applications, ranging from novelty detection and abrupt change detection to clustering. Though less attention has been paid to 1-SVMs than to their 2-class and multiclass counterparts, they have many strong theoretical and practical properties which make them extremely useful in applications.

The general problem addressed by 1-SVMs is fully exposed in [1, chapter 8], and we do not recall it here for the sake of conciseness. Let $\mathsf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ be a set of m in $\mathcal{X}$, assumed to be distributed i.i.d. according to a pdf $\mathsf{p}(\mathbf{x})$. Considering the so-called ν-1-SVM framework, the problem to be solved is to estimate a level set $\mathsf{S}^\gamma = \{\mathbf{x} \in \mathcal{X} | \mathsf{p}(\mathbf{x}) \geq \gamma\}$ from data by finding a function f^λ in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ of kernel $\mathsf{k}(\cdot,\cdot)$, and an offset $b^\lambda \in \mathbb{R}$ such that

$$\begin{array}{ll} \underset{\mathsf{f},b,\{\xi_i\}}{\text{Minimize}} & \displaystyle\sum_{i=1}^m \xi_i - \lambda b + \frac{\lambda}{2}\|\mathsf{f}(\cdot)\|^2_{\mathcal{H}} \\ \text{with, for all } i = 1,\ldots,m & \mathsf{f}(\mathbf{x}_i) \;\geq\; b - \xi_i \quad \text{and} \quad \xi_i \;\geq\; 0 \end{array} \tag{1}$$

for any λ in $(0, m)$, and letting $\widehat{\mathsf{S}}^\lambda_{\mathsf{X}} = \{\mathbf{x} \in \mathcal{X} | \mathsf{f}^\lambda(\mathbf{x}) - b^\lambda \geq 0\}$. It can be shown [1] that, asymptotically, $1 - \lambda/m$ is the probability mass enclosed inside the level set S^γ, yielding a relation between λ and γ. In practice, tuning λ is even more intuitive that tuning γ. For example, $\lambda = 0.2 \times m$ means that at most (and asymptotically exactly) 20% of the vectors in X are outliers. In the following, we denote by $(\mathsf{f}^\lambda(\mathbf{x}), b^\lambda)$ the minimizer of (1), owing to the equivalence between γ and λ.

From the representer theorem, the solution $\mathsf{f}^\lambda(\cdot)$ belongs to the subspace of $\mathcal{H}$ spanned by the functions $\{\mathsf{k}(\mathbf{x}_i, \cdot), i = 1, \ldots, m\}$, that is $\mathsf{f}^\lambda(\cdot) = \frac{1}{\lambda}\sum_{i=1}^m \alpha_i^\lambda \mathsf{k}(\mathbf{x}_i, \cdot)$, where the coefficient $1/\lambda$ appears for reasons that will be made clearer in the

following. The data in X may be in one of the three sets: 1) *non-support vectors* $\mathcal{L}^\lambda = \{i \in [1,m] : \mathsf{f}^\lambda(\mathbf{x}_i) - b^\lambda > 0 \text{ and } \alpha_i^\lambda = 0\}$; 2) *margin support vectors* $\mathcal{E}^\lambda = \{i \in [1,m] : \mathsf{f}^\lambda(\mathbf{x}_i) - b^\lambda = 0 \text{ and } 0 < \alpha_i^\lambda < 1\}$ and 3) *non-margin support vectors* or *outliers* $\mathcal{R}^\lambda = \{i \in [1,m] : \mathsf{f}^\lambda(\mathbf{x}_i) - b^\lambda < 0 \text{ and } \alpha_i^\lambda = 1\}$.

This paper concerns problems where the 1-SVM problem (1) should be solved for various values of λ. For example, in novelty detection, selecting the right value for λ may require to cross-validate, that is, solve (1) for several λ's.In density estimation by using the level sets [2], problem (1) should be solved for any λ in $[0, m]$. In multiclass classification with a rejection class, several 1-SVMs may be used, with their λ's tuned according to a given criterion. This also requires to solve problem (1) for several λ's for each class. The main contribution of this paper is the derivation of the entire regularization path for ν-1-SVMs, building on the work by Hastie et al. [3].

This paper is organised as follows: in Section 2, we show that the complete regularization path can be computed using simple update rules, following the approach in [3]. Section 3 presents some numerical experiments that illustrates the piecewise linear behaviour of regularization path. Furthermore, we expose some results showing that regularization path method is more efficient than alpha seeding for computing different One-Class SVM solution. Conclusions and future work directions are given in Section 4.

2 Derivation of the entire regularization path

In this section, we derive the entire regularization path for ν-1-SVM. This requires to coma back to the dual optimization problem of problem (1). We suppose in the following that the kernel is such that for any $\mathbf{x}$, $\mathsf{k}\big(\mathbf{x},\mathbf{x}\big) = 1$.

2.1 Dual optimization problem

In order to derive the entire regularization path for 1-SVMs, one writes the Lagrangian for Eq. (1) for some λ, which provides the dual to be solved with numerical optimization technique w.r.t. the α_i's:

$$\underset{\alpha_1,\ldots,\alpha_m}{\text{Minimize}} \quad \frac{1}{2\lambda}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\mathsf{k}\big(\mathbf{x}_i,\mathbf{x}_j\big) \tag{2}$$

$$\text{with} \quad \sum_{i=1}^{m}\alpha_i = \lambda \text{ and } 0 \le \alpha_i \le 1 \text{ for all } i = 1,\ldots,m \tag{3}$$

Once this problem is solved for some λ yielding the solution denoted $\{\alpha_1^\lambda,\ldots,\alpha_m^\lambda\}$, the offset b^λ is computed from Karush-Kuhn-Tucker conditions.

2.2 Initialization of the path

In order to solve this problem for every value of λ, we assume that the problem has been solved for some initial value λ_0. It appears that initializing at $\lambda_0 \approx m$

with $\lambda_0 < m$ requires few computations. First, we note that for $\lambda_0 = m$, the only admissible solution is the Parzen density estimator given by $\alpha_1^m = \ldots = \alpha_m^m = 1$ and $b^m = \sum_{i,j=1}^m \alpha_i^m \alpha_j^m \mathsf{k}(\mathbf{x}_i, \mathbf{x}_j)$ thus all the data belong to $\mathcal{R}^m$. The elbow set $\mathcal{E}^m$ is empty, and we shall see that we cannot start from this situation in practice. Second, set $\lambda_0 = m - \Delta\lambda$ where $\Delta\lambda > 0$ and $\Delta\lambda \ll 1$. Thus, from condition (3), at least one datum (with index k) is such that $\alpha^{\lambda_0} < 1$. Let us assume for instance that $\Delta\lambda$ is small enough, and that the datum are positioned in such a way that this element is unique. Then, problem (2) becomes that of minimizing w.r.t k

$$\frac{1}{2\lambda}\sum_{i=1}^m\sum_{j=1}^m \mathsf{k}(\mathbf{x}_i, \mathbf{x}_j) - \frac{\Delta\lambda}{\lambda}\sum_{i=1}^m \mathsf{k}(\mathbf{x}_i, \mathbf{x}_k) + \frac{\Delta\lambda^2}{\lambda} \tag{4}$$

because $\alpha_i = 1$ $(i \neq k)$, while $\alpha_k = 1 - \Delta\lambda$. Changing k does not change the first and third terms in (4), but it changes the second term which equals $\frac{\Delta\lambda}{\lambda}\mathsf{f}^m(\mathbf{x}_k)$. Thus, the element $\mathbf{x}_k \in \mathsf{X}$ with $k \in \mathcal{E}^m$ as soon as $\Delta\lambda > 0$ is the one such that $\mathsf{f}^m(\mathbf{x}_k) = \langle \mathsf{k}(\mathbf{x}_k, \cdot), \mathsf{f}^m(\cdot)\rangle_{\mathcal{H}}$ is maximum (this is the one closest to the barycenter $\mathsf{f}^m(\cdot)$ in $\mathcal{H}$). With probability zero, it may happen that for any $\Delta\lambda > 0$, several $\mathbf{x}_k$'s have the same value for $\mathsf{f}^m(\mathbf{x}_k)$, meaning that several k's belong to $\mathcal{E}^m$. In that case, a QP optimization procedure should be implemented over $\mathcal{E}^m$ so as to compute the $\alpha_k^{\lambda_0}$'s.

2.3 Running down the path

When running down along the path, it occurs that the indexes of data $\mathbf{x}_i$ change from one of the sets $\mathcal{L}^\lambda$, $\mathcal{R}^\lambda$, $\mathcal{E}^\lambda$ to another. The values of λ that correspond to at least one such change are denoted λ_ℓ in the following, with $\lambda_\ell > \lambda_{\ell+1}$. Moreover, for any value of λ, denote $\mathsf{g}^\lambda(\mathbf{x}) = \frac{1}{\lambda}\sum_{i=1}^m \alpha_i^\lambda \mathsf{k}(\mathbf{x}_i, \mathbf{x}) - b^\lambda = \frac{1}{\lambda}\big(\sum_{i=1}^m \alpha_i^\lambda \mathsf{k}(\mathbf{x}_i, \mathbf{x}) - \alpha_0^\lambda\big)$ with $\alpha_0^\lambda = \lambda b^\lambda$.

2.3.1 Computing α_i^λ's for $\lambda \in [\lambda_{\ell+1}, \lambda_{\ell+1}]$

Let λ be such that $\lambda_\ell \geq \lambda \geq \lambda_{\ell+1}$, and assume $\mathcal{E}^\ell$ is nonempty (where $\mathcal{E}^\ell$ is a shorthand for $\mathcal{E}^{\lambda_\ell}$), we have

$$\mathsf{g}^\lambda(\mathbf{x}) \;=\; \frac{1}{\lambda}\left[\sum_{i\in\mathcal{E}^\ell}(\alpha_i^\lambda - \alpha_i^\ell)\mathsf{k}(\mathbf{x}, \mathbf{x}_i) - (\alpha_0^\lambda - \alpha_0^\ell) + \lambda_\ell \mathsf{g}^\ell(\mathbf{x})\right] \tag{5}$$

Eq.(5) results from applying the same derivation as is [3, Eq. (29)]. Applying Eq. (5) to all the indexes j in $\mathcal{E}^\ell$ yields

$$\sum_{i\in\mathcal{E}^\ell}(\alpha_i^\lambda - \alpha_i^\ell)\mathsf{k}(\mathbf{x}_k, \mathbf{x}_i) - (\alpha_0^\lambda - \alpha_0^\ell) \;=\; 0 \tag{6}$$

because $\mathsf{g}^\lambda(\mathbf{x}_k) = \mathsf{g}^\ell(\mathbf{x}_k) = 0$ for all $k \in \mathcal{E}^\ell$. Eq.(6) being verified for all $k \in \mathcal{E}^\ell$, the α_i^λ can be computed by solving a linear system, and we have

$$\alpha_i^\lambda \;=\; \alpha_i^\ell - (\lambda_\ell - \lambda)\beta_i^\ell \quad \text{for} \quad i = 0, \ldots, m \tag{7}$$

where β_i^ℓ is component $\#i$ of vector $\boldsymbol{\beta}^\ell$ defined as

$$\boldsymbol{\beta}^\ell = \left[\mathbf{A}^\ell\right]^{-1}\mathbf{c} \text{ where } \mathbf{A}^\ell = \begin{bmatrix} \mathbf{K}_{\mathcal{E}^\ell} & -\mathbf{1}_m \\ \mathbf{1}_m^\mathsf{T} & 0 \end{bmatrix} \text{ and } \mathbf{c} = [\mathbf{0}_m^\mathsf{T} \quad 1]^\mathsf{T} \tag{8}$$

In Eq. (8), $\mathbf{K}_{\mathcal{E}^\ell}$ is the kernel matrix of the $\mathbf{x}_i$'s for $i \in \mathcal{E}^\ell$. From Eq.(7), we see that α_i^λ for $i \in \mathcal{E}^\ell$ evolve linearly. Of course, the remaining Lagrange multipliers α_i^λ, $i \in \mathcal{L}^\ell \cup \mathcal{R}^\ell$ do not evolve. Eq.'s (7)- (8) enable to compute $\mathsf{f}^\lambda(\cdot) - b^\lambda$ for any value of λ such that $\lambda_\ell \geq \lambda \geq \lambda_{\ell+1}$.

2.3.2 Finding the boundaries λ_ℓ

When starting from λ_0 and running down the path, boundaries are found recursively. Assume boundary λ_ℓ has been found and $\lambda_\ell \geq \lambda$. The next boundary $\lambda_{\ell+1}$ is met whenever a change occurs in $\mathcal{E}^\ell$, $\mathcal{L}^\ell$ to $\mathcal{R}^\ell$. This happens when

1. one of the α_i, $i \in \mathcal{E}^\ell$ reaches 0 or 1. For each $i \in \mathcal{E}^\ell$, the values that lead to such situations are $\lambda = \frac{1-\alpha_i^\ell}{\beta_i^\ell} + \lambda_\ell$ and $\lambda = -\frac{\alpha_i^\ell}{\beta_i^\ell} + \lambda_\ell$
2. one of the $\mathbf{x}_i$, $i \in \mathcal{L}^\ell \cup \mathcal{R}^\ell$ reaches $\mathsf{g}^\lambda(\mathbf{x}_i) = 0$. This occurs whenever $\lambda = \lambda_\ell\big[1 - \frac{\mathsf{g}^\ell(\mathbf{x}_i)}{\sum_{j\in\mathcal{E}^\ell} \beta_j^\ell \mathsf{k}(\mathbf{x}_i, \mathbf{x}_j)}\big]$

Thus, the next boundary $\lambda_{\ell+1}$ is the largest λ that verifies one of the above conditions.

2.3.3 On the emptiness of $\mathcal{E}^\ell$

It has been shown that for $\lambda = m$, the elbow set $\mathcal{E}^\lambda$ may be empty, that is, $\mathsf{g}^\lambda(\mathbf{x}_i) \neq 0$ for all $i = 1, \ldots, m$. This can also (and only) happen whenever λ is any integer smaller than m, because of condition 3, and because the α_i^λ's with indexes in $\mathcal{L}^\lambda$ and $\mathcal{R}^\lambda$ are 0 and 1 respectively. In this case, since we are in the same situation than the one when $\lambda = m$, we rely on an the initialization step (see section 2.2) in order to figure out the next example $\mathbf{x}_i$ to be integrated in the elbow set.

3 Experiments

3.1 Illustration on toy problem

Firstly, we want to illustrate the result of a OC-SVM regularization path on a simple toy problem. The data we used are samples from a gaussian distribution and we have run our regularization path algorithm using a gaussian kernel. Results are depicted in Figure 1. The left part of the figure shows the evolution of the Lagrangian multipliers α with respect to λ. The piecewise linear behaviour of the α is clearly highlighted. The right part depicts four different level-set estimation of the gaussian distribution that has generated the samples.

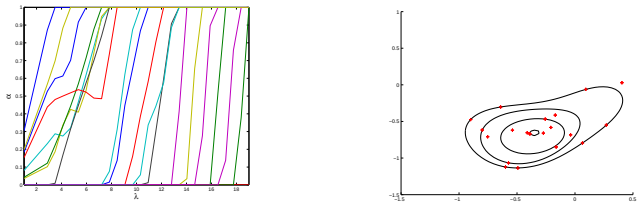

Fig. 1: Illustration of the regularization path of a One-Class SVM on a gaussian toy problem. (left) The entire piecewise linear paths of α_i with respects to λ. (right) Examples of level sets estimations for the same dataset

3.2 Comparing regularization path and alpha seeding

As we have stated in the introduction, there exists several problems which involves One-Class SVM and which may need an efficient model selection with respects to λ. For instance, novelty or change detection algorithms based on One-Class SVM [4, 5] need the tuning of the hyperparameter in order to achieve good performance.

Another example would be the problem of estimating a density estimation through the estimation of several level sets. For such problem, the need of efficient way of computing several $S\gamma$ is clear. In this paragraph, we aim at empirically proving that the regularization path algorithm we propose is very efficient even compared to a warm-restart approach [6]. Remember that a warm-restart approach consists of using the Lagrangian multipliers obtained for a previous value of λ for initializing the QP problem with a new value of λ. DeCoste et al. [6] has proven that such approach allows to considerably speed-up a model selection procedure.

Hence, for several binary classification datasets from the UCI repository, we have compared the computational cost of obtaining different level-sets for each class using alpha seeding approach and regularization path. 19 level sets have been estimated for $\gamma = \{0.05, 0.10, \cdots, 0.90, 0.95\}$. For each dataset, 90% of the examples has been randomly drawn and used for evaluating the computational time. the procedure has been repeated 10 times. The kernel that has been used for the experiment is a gaussian kernel $k(x, y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$ with different values of σ. The computation has been performed on a Pentium D 3GHz and 1 *Gb* of RAM. All the code has been written in Matlab and the One-Class SVM that has been used is available on authors website. Results show that regularization path is more efficient than alpha seeding in most of the cases we analyse. The speed-up gain (which order varies between 0.8 and 25) depends on the kernel and seems to be smaller for large-scale datasets. Note however that the regularization path has provided more level sets than the alpha seeding approach. On the average, the number of boundaries generating a level-set is of the order of twice the number of examples.

Table 1: Comparing computational time in seconds of alpha seeding and a regularization path approach for computing several level sets

Datasets	# examples	σ	**Alpha Seeding**	**Reg. Path**
credit	653	1	18.1	0.7
		5	21.4	3.8
		10	15.8	4.4
pima	768	1	54.3	0.8
		5	39.8	20.7
		10	25.5	11.2
yeast-cyt	1484	1	42.9	49.42
		5	42.6	51.87
		10	42.5	38.9
spamdata	4601	1	18220	7460
		5	2265	1446
		10	1114	1039

4 Conclusion

This paper proposes the derivation of the entire regularization path of the One-Class SVM algorithm according to a parameter λ. This algorithm is able to provide in a efficient way, estimations of different level sets of the probability density function from which examples have been sampled. By building on the work of Hastie et al., we have shown that the Lagrangian multipliers of a One-class SVM vary in piecewise linear way according to parameter. Furthermore, we have provided experimental evidence of the efficiency of regularization path approach compared to alpha seeding heuristics for computing several OC-SVM solutions. The perspectives of this work is to analyze the add-on values of such regularization path approach to different problems such as novelty detection, change detection or multiclass SVM approach.

References

[1] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, USA, 2002.

[2] W. Polonik. Density estimation under qualitative assumptions in higher dimensions. *Journal of Multivariate Analysis*, 55(1):61–81, October 1995.

[3] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu;. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, October 2004.

[4] F. Desobry, M. Davy, and C. Doncarli. An online kernel change detection algorithm. *IEEE transations on Signal Processing*, 53(8):2961–2974, 2005.

[5] F. Desobry, M. Davy, A. Gretton, and C. Doncarli. An online support vector machine for abnormal events detection. *Signal Processing*, 86:2009–2025, 2006.

[6] D. DeCoste and K. Wagstaff. Alpha seeding for support vector machines. In *International Conference on Knowledge Discovery and Data Mining*, 2000.