Convex formulations of radius-margin based Support Vector Machines Appendix

Huyen Do and Alexandros Kalousis

Contents

1 Appendix

In this appendix, we give more details for:

- Proof the inequalities of the new radius approximation $R_O \leq R \leq \frac{1+\sqrt{3}}{2} R_O$ (Lemma 1)
- Proof of Lemma 2
- Details of solving the optimization problems of $R\text{-}SVM^+_{\boldsymbol{\mu}}$ and $R\text{-}SVM^+$.
- Another way to kernelize $R\text{-}SVM^+$, using representer theorem
- Proof that Shivaswamy&Jebara's approach [5] cannot replace the radius-margin approach

1.1 Proof of Lemma 1 (inequalities of the new radius approximation $R_O \leq R \leq \frac{1+\sqrt{3}}{2} R_O$)

The furthest possible instance from $\mathbf{x}_A$ and $\mathbf{x}_B$ will be the one lying on the intersection I of the two spheres C_A and C_B, or to its symmetric point G with respect to AB: $IA = IB = d$ and $GA = GB = d$. Moreover if there are instances in the region P_1, defined by the arcs AI, IB and the upper hemisphere AB - the gray area, there will be no instance in the region P_2, defined by the arcs AG, GB and the lower hemisphere AB - the blue area, since distance between instances should be smaller than $d = AB$. Therefore all instances are lying in the region P_1 and on the sphere C_O or in the region P_2 and on the sphere C_O. Without loss of generality we assume that they lie on P_1 and C_O.

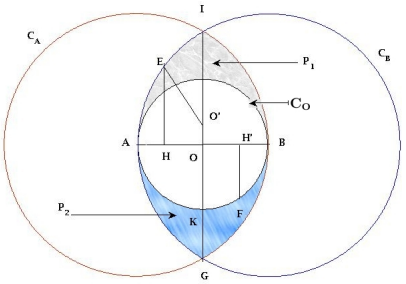

Figure 1: Demonstrating the radius relations.

Let $C_{O'}$ is the sphere which has center O', radius $R_{O'}$ and diameter IK, where K is the intersection of line IG and the lower arc AB (see Figure 1). Since $IK = (1+\sqrt{3})R_O$ we have $R_{O'} = \frac{1+\sqrt{3}}{2}R_O$. We show that this sphere $C_{O'}$ covers all the regions P_1 and C_O. To do so we prove that the distance from any instances of that regions to the center O' is smaller than the radius $R_{O'}$ of the circle $C_{O'}$. Let E be a point that lies on the IA arc, hence $EB = d$. Let F be a point that lies on the lower arc AB (see Figure 1). To prove that the distance from any instances to the center O' is smaller than the radius $R_{O'}$, it is sufficient to prove that $O'E \leq R_{O'}$ and $O'F \leq R_{O'}$.

1.1.1 Proof that $O'E \leq R_{O'}$ and $O'F \leq R_{O'}$.

Let H be the projection E on AB. We first show that $EH \leq OI = \sqrt{3}R_O$. Indeed, since $HB \geq d/2$, we have:

$$\begin{aligned} EH^2 &= EB^2 - HB^2 = d^2 - HB^2 \\ &\leq d^2 - (d/2)^2 = \frac{3}{4}d^2 \end{aligned} \quad (1)$$

The equality happens when $E \equiv I$.
Moreover we have $O'E^2 = (EH - O'O)^2 + OH^2$, where:

$$\begin{aligned} OH^2 &= (\sqrt{(EB^2 - EH^2)} - OB)^2 \\ &= d^2 - EH^2 + R_O^2 - 2R_O\sqrt{d^2 - EH^2} \end{aligned} \quad (2)$$

and $O'O = OI - O'I = \frac{\sqrt{3}-1}{2}R_O$.
Therefore:

$$O'E^2 = (6 - \frac{\sqrt{3}}{2})R_O^2 - (\sqrt{3}-1)R_O.EH - 2R_O\sqrt{4R_O^2 - EH^2}$$

Since $EH \leq \sqrt{3}R_O$, we have $O'E^2 \leq (1 - \frac{\sqrt{3}}{2})R_O^2 = R_{O'}^2$.
Let H' is the projection of F on AB, we have

$$\begin{aligned} O'F^2 &= (FH' + O'O)^2 + H'O^2 \\ &= OF^2 + O'O^2 + 2OH'.O'O \\ &= (2 - \frac{\sqrt{3}}{2})R_O^2 + (\sqrt{3}-1)R_O.OH' \\ &\leq (1 - \frac{\sqrt{3}}{2})R_O^2 = R_{O'}^2 \end{aligned}$$

Thus we proved that $R_O \leq R \leq R_{O'} = \frac{1+\sqrt{3}}{2} R_O$. This also holds true for high dimensional spaces. ■

1.2 Proof of Lemma 2, section 4.4

We will prove that for each optimal solution of the ratio form (10), there exists a value of λ for which the sum form (9) has the same optimal solution. Indeed, let $\mathbf{w}^*, R$ be the optimal solution of (10). We have $\|\mathbf{w}^*\|^2 R^{*2} \leq \|\mathbf{w}\|^2 R^2, \forall \mathbf{w}, R \in F$, we have to show that there exists a $\lambda > 0$ such that $\|\mathbf{w}^*\|^2 + \lambda R^{*2} \leq \|\mathbf{w}\|^2 + \lambda R^2, \forall \mathbf{w}, R \in F$. To do so we find a λ that satisfies the above inequality. That λ has to satisfy $\lambda \leq B = \frac{\|\mathbf{w}\|^2 - \|\mathbf{w}^*\|^2}{R^{*2} - R^2}$ for $\forall \{\mathbf{w}, R \in F | R < R^*\}$, and $\lambda \geq A = \frac{\|\mathbf{w}^*\|^2 - \|\mathbf{w}\|^2}{R^2 - R^{*2}}$ for $\forall \{\mathbf{w}, R \in F | R > R^*\}$. Since $\|\mathbf{w}^*\|^2 R^{*2} \leq \|\mathbf{w}\|^2 R^2, \forall \mathbf{w}, R \in F$, it is easy to show that $B = \frac{\|\mathbf{w}\|^2 - \|\mathbf{w}^*\|^2}{R^{*2} - R^2} \geq \frac{\|\mathbf{w}\|^2}{R^{*2}}$ for $\forall \{\mathbf{w}, R \in F | R < R^*\}$ and $A = \frac{\|\mathbf{w}^*\|^2 - \|\mathbf{w}\|^2}{R^2 - R^{*2}} \leq \frac{\|\mathbf{w}*\|^2}{R^2}$ for $\forall \{\mathbf{w}, R \in F | R > R^*\}$. In addition, we also have $\frac{\|\mathbf{w}*\|^2}{R^2} \leq \frac{\|\mathbf{w}\|^2}{R^{*2}}$, therefore $\max_{R<R^*} A \leq \min_{R>R^*} B$. This means that we can always find λ which satisfies $A \leq \lambda \leq B$. We can estimate λ using cross validation.

1.3 Solving the convex optimization problems of *R-SVM*$^+_{\boldsymbol{\mu}}$ and *R-SVM*$^+$

We will now show how to solve the optimization problems of *R-SVM*$^+_{\boldsymbol{\mu}}$ and *R-SVM*$^+$.

1.3.1 Solving *R-SVM*$^+_{\boldsymbol{\mu}}$

We propose a two step algorithm to solve the optimization problem of *R-SVM*$^+_{\boldsymbol{\mu}}$ given in (6) - section 4.2. In the first step we fix $\boldsymbol{\mu}$ and we solve a simplified optimization problem, in the second step we optimize over $\boldsymbol{\mu}$ using gradient descent. So the original *R-SVM*$^+_{\boldsymbol{\mu}}$ given in (6) is now reformulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{\mu}} J(\boldsymbol{\mu}) && s.t. \ \sum_k \mu_k = 1, \mu_k \geq 0 \\ \text{where: } J(\boldsymbol{\mu}) = \quad &\min_{\mathbf{w},b,\boldsymbol{\xi},r} && \frac{1}{2}\sum_k^d \frac{w_k^2}{\mu_k} + \lambda r + C\sum_i^l \xi_i \\ &s.t. && y_i(\sum_k^d \langle w_k, \mathbf{x}_{ik}\rangle + b) \geq 1 - \xi_i, \forall i \\ & && \frac{1}{2}\|\mathbf{D}_{\sqrt{\boldsymbol{\mu}}}\mathbf{x}_i - \mathbf{D}_{\sqrt{\boldsymbol{\mu}}}\mathbf{x}_j\|^2 \leq r, \forall i,j \end{aligned} \tag{3}$$

With a fixed $\boldsymbol{\mu}$, (3) is similar to standard *SVM* and can be optimized using its dual form:

$$\begin{aligned} J(\boldsymbol{\mu}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \sum_i \alpha_i - \frac{1}{2}\sum_{ij}\alpha_i\alpha_j y_i y_j \sum_k \mu_k \mathbf{x}_{ik}\mathbf{x}_{jk} + \lambda \sum_{ij}\beta_{ij}\sum_k \mu_k(\mathbf{x}_{ik} - \mathbf{x}_{jk})^2 \\ s.t. \quad & \sum_i \alpha_i y_i = 0, \ \ 0 \leq \alpha_i \leq C, \forall i, \ \ \sum_{ij}\beta_{ij} = 1, \ \ \beta_{ij} \geq 0 \end{aligned} \tag{4}$$

Let $\boldsymbol{\alpha}^*, \mathbf{x}_i^*, \mathbf{x}_j^*$ be the optimal solution of the above optimization problem, the gradient of $J(\boldsymbol{\mu})$ of (4) with respect to $\boldsymbol{\mu}$ is: $\frac{\partial J}{\partial \mu_k} = \lambda(\mathbf{x}_{ik}^* - \mathbf{x}_{jk}^*)^2 - \frac{1}{2}\sum_{ij}\alpha_i^*\alpha_j^* y_i y_j \mathbf{x}_{ik}^*\mathbf{x}_{jk}^*$

Concerning the number of constraints of the radius, we have one way to compute them efficiently. We compute vector V of pairwise instance distances with respect to *one* feature: $d(\mathbf{x}_i, \mathbf{x}_j, k) = \mathbf{x}_i(k).\mathbf{x}_j(k)$ where $\mathbf{x}_i(k)$ is the kth feature of instance $\mathbf{x}_i$. We have to compute

this vector V only once, and at each iteration of a fixing $\boldsymbol{\mu}$ we can compute all the constraints in the second line of equation (3) by simply computing the product of V and $\boldsymbol{\mu}$ which can be very efficient. The same trick can be applied to solving $R\text{-}SVM^+$.

The inner solver can employ Liblinear [2] which is a fast, large scale, linear SVM solver.

1.3.2 Solving $R\text{-}SVM^+$

The dual form of (7) - section 4.3, is:

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \sum_i \alpha_i - \frac{1}{2}\sum_{ij}\alpha_i\alpha_j y_i y_j \sum_k \frac{\mathbf{x}_{ik}\mathbf{x}_{jk}}{1+\sum_{tl}\beta_{tl}(\mathbf{x}_{tk}-\mathbf{x}_{lk})^2} \\ s.t. \quad & \sum_{ij}\beta_{ij} = \lambda, \ \sum_i \alpha_i y_i = 0, \ \alpha_i, \beta_{ij} \geq 0, \forall i,j \end{aligned} \tag{5}$$

Problem (5) is convex [1](see proof in section 1.3.3). As with $R\text{-}SVM^+_{\boldsymbol{\mu}}$ we also use a similar two step algorithm to solve the optimization problem (5). In the first step we fix $\boldsymbol{\beta}$ and in the second step we optimize the problem over $\boldsymbol{\beta}$ using gradient descent. (5) is now reformulated as follows:

$$\max_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) \qquad s.t. \ \sum_{ij}\beta_{ij} = \lambda, \beta_{ij} \geq 0, \forall i,j \tag{6}$$

where

$$\begin{aligned} J(\boldsymbol{\beta}) \quad = \quad & \max_{\boldsymbol{\alpha}} \sum_i \alpha_i - \frac{1}{2}\sum_{ij}\alpha_i\alpha_j y_i y_j \sum_k \frac{\mathbf{x}_{ik}\mathbf{x}_{jk}}{1+\sum_{tl}\beta_{tl}(\mathbf{x}_{tk}-\mathbf{x}_{lk})^2} \\ s.t. \quad & \sum_i \alpha_i y_i = 0, \ 0 \geq \alpha_i \geq C, \forall i \end{aligned} \tag{7}$$

Let $\boldsymbol{\alpha}^*$ is the optimal solution of (7). The gradient of $J(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ is:

$$\frac{\partial J}{\partial \beta_{tl}} = \frac{1}{2}\sum_{ij}\alpha_i^*\alpha_j^* y_i y_j \sum_k \frac{\mathbf{x}_{ik}\mathbf{x}_{jk}(\mathbf{x}_{tk}-\mathbf{x}_{lk})^2}{(1+\sum_{tl}\beta_{tl}(\mathbf{x}_{tk}-\mathbf{x}_{lk})^2)^2}$$

Similarly, the inner solver can employ Liblinear [2] which is a fast, large scale, linear SVM solver.

1.3.3 Proof of convexity of problem (5)

Let $t_k := 1 + \sum_{ij}\beta_{ij}(\mathbf{x}_{ik}-\mathbf{x}_{jk})^2$. Problem (5) can be reformulated as:

$$\begin{aligned} \max_{\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{t}} \quad & \sum_i \alpha_i - \frac{1}{2}\sum_{ij}\alpha_i\alpha_j y_i y_j \sum_k \frac{\mathbf{x}_{ik}\mathbf{x}_{jk}}{t_k} \\ s.t. \quad & \sum_{ij}\beta_{ij} = \lambda, \quad \sum_i \alpha_i y_i = 0, \quad \alpha_i, \beta_{ij} \geq 0, \forall i \\ & 1 + \sum_{ij}\beta_{ij}(\mathbf{x}_{ik}-\mathbf{x}_{jk})^2 = t_k, \forall k \end{aligned}$$

The factor $\frac{1}{2}\sum_{ij}\alpha_i\alpha_j y_i y_j \sum_k \frac{\mathbf{x}_{ik}\mathbf{x}_{jk}}{t_k}$ can be written as $\frac{\boldsymbol{\alpha}^T\mathbf{B}\boldsymbol{\alpha}}{t_k}$ where $\mathbf{B}$ is a positive semidefinite matrix. Since $g(\mathbf{x}) = \boldsymbol{\alpha}^T\mathbf{B}\boldsymbol{\alpha}$ is convex, $g(\mathbf{x}/t_k), t_k > 0, t \in \mathcal{R}$ is also convex [1], i.e $g(\mathbf{x}/t_k) = \frac{\boldsymbol{\alpha}^T\mathbf{B}\boldsymbol{\alpha}}{t_k}$ is convex. Therefore the above optimization problem is convex since the objective function is convex and its constraints are linear. ■

[1]It is worth to note that if the primal form is convex, the dual form is not necessary convex [1].

1.3.4 Implementation

In our implementation we have used the following packages cvx [1], SimpleMKL [4], LibLinear [2]. Matlab code can be downloaded at:
`https://sourceforge.net/projects/svmradiusmargin/`

1.4 Another way to kernelize *R-SVM*$^+$, using representer theorem

We cannot kernelize *R-SVM*$^+$, problem (9), directly as it is the case with the *SVM* but we need instead to use a kernel trick. We use the same kernel trick as in [3], where $\mathbf{w}$ is expressed as a linear combination of the training points $\mathbf{w} = \sum_i \alpha_i \mathbf{\Phi}(\mathbf{x}_i)$, following the representer theorem, then we can reformulate (9) as:

$$\begin{aligned} \min_{\boldsymbol{\alpha},b,r} \quad & \sum_{i,j} \alpha_i \alpha_j \mathbf{K}_{ij} + \frac{\lambda}{2} r \\ s.t. \quad & y_i(\sum_j \alpha_j \mathbf{K}_{ji} + b) \geq 1, \forall i \\ & \frac{1}{2} \sum_{k,t} \alpha_k \alpha_t (\mathbf{K}_{ik} - \mathbf{K}_{jk})(\mathbf{K}_{it} - \mathbf{K}_{jt}) \leq r, \forall i, j \end{aligned} \tag{8}$$

This is a QCQP optimization problem since $\sum_{k,t} \alpha_k \alpha_t (\mathbf{K}_{ik} - \mathbf{K}_{jk})(\mathbf{K}_{it} - \mathbf{K}_{jt}) = \boldsymbol{\alpha}^T \mathbf{Q} \boldsymbol{\alpha} \geq 0, \forall \boldsymbol{\alpha}$.

1.5 Proof that Shivaswamy&Jebara's approach cannot replace the radius-margin approach

[5] proposed to optimize the margin and some measure of the data spread. However, this approach can not replace the radius-margin one since they are not equivalent (as it is claimed in that paper). While it is not clear why any other measure of the data spread would be better than the radius (this would probably depend on the specificity of any given learning problem, and can only be seen on a case by case basis), using the radius has the advantage of the theoretical support it enjoys through its direct reliance on the SVM theoretical error bound.

We now prove that Shivaswamy&Jebara's claim of the equivalence of their approach to the radius-margin approach is not correct. This claim is founded by equation (2), Lemma 1, and the discussion after Lemma 1 in Shivaswamy&Jebara's paper, page 753-754.

First, we show that their equation (2) is not well-posed. Indeed, if the number of features is bigger than the number of instances (which is usually encountered in feature selection problems), i.e $m > n$, we can always find a linear transformation $\mathbf{A}$ which maps all instances of one class to one point and all other instances to some other point (a system of linear equation system with number of variables bigger than number of equations can always find a solution). Thus the radius is 2 times the margin, the training set now degenerates to two points. The objective value become $(2E - 1)\rho$ where E is a constant and ρ is the value of the margin. It means that if $2E \geq 1$ we can always choose ρ as close to zero as possible; if $2E < 1$ we can choose ρ infinitely big and the objective can be infinitely (negative) small. Therefore the optimization problem in equation (2) is ill- posed, the results based on this equation are not reliable.

Second, we show that their Lemma 1 is not correct. Lemma 1 is based on equation (2), moreover even if (2) is well posed, the conclusion of Lemma 1 is still insufficient. The solution

matrix $\widetilde{\mathbf{A}}^*$ can be rank one, however it can have any rank. If $\mathbf{BC} = \mathbf{D}$ where $\mathbf{D}$ is a rank one matrix, and $\mathbf{C}$ is also a rank one, then it is impossible to conclude that $\mathbf{B}$ is also rank one. Therefore, Lemma 1 is not correct.

An anonymous has pointed out that the problem with this lemma can be fixed by replacing the radius in the objective with $trace(A * A')$ and the radius in the constraint with $1.tr(A * A')$. However, in that case, Lemma 1 is correct but ones cannot use it to conclude that the approach of Shivaswamy&Jebara is equivalent to the radius-margin approach, therefore, it can not replace the radius-margin approach.

Therefore, Shivaswamy&Jebara's approach is just one way to solve the problem, but it cannot replace completely the radius-margin one. Although their approach is simpler because it does not require the radius computation, it is not clear whether there is any advantage of using a data spread measure other than the radius.

In addition our radius-margin approach also leads to convex optimization problems as Shivaswamy&Jebara's approach. In our paper we also show that our method, a convex variant of the radius-margin approach in fact outperforms RMM of Shivaswamy&Jebara.

1.6 Matlab code is available

Matlab code can be downloaded at:
`https://sourceforge.net/projects/svmradiusmargin/`

References

[1] Boyd, S., Vandenberghe, L.: Convex optimization. Cambridge University Press (2004)

[2] Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin., C.J.: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874. (2008)

[3] Mika, S., Ratsch, G., Weston, J., Schoelkopf, B., Muller, K.R.: Fisher discriminant analysis with kernels. In: Neural Networks for Signal Processing (1999)

[4] Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: SimpleMKL. Journal of Machine Learning Research 9, 2491–2521 (2008)

[5] Shivaswamy, P.K., Jebara, T.: Maximum relative margin and data-dependent regularization. Journal of Machine Learning Research 11 (2010)