A robust regularization path for the Doubly Regularized Support Vector Machine

Antoine Lachaud, David Mercier, S. Canu, Frédéric Suard

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

A robust regularization path for the Doubly Regularized Support Vector Machine

Antoine Lachaud[1], David Mercier[1], Stephane Canu[2], and Frederic Suard [1]

1- CEA,LIST,
Gif sur Yvette - 91191 - France

2- INSA de Rouen - LITIS
Avenue de l'Universite - Saint-Etienne-du-Rouvray - 76800 - France

Abstract. The Doubly Regularized SVM (DrSVM) is an extension of SVM using a mixture of L_2 and L_1 norm penalties. This kind of penalty, sometimes referred as the elastic net, allows to perform variable selection while taking into account correlations between variables. Introduced by Wang [1], an efficient algorithm to compute the whole DrSVM solution path has been proposed. Unfortunately, in some cases, this path is discontinuous, and thus not piecewise linear. To solve this problem, we propose here a new sub gradient formulation of the DrSVM problem. This led us to propose an alternative L_1 regularization path algorithm. This reformulation efficiently addresses the aforementioned problem and makes the initialization step more generic. The results show the validity of our sub-gradient formulation and the efficiency compared to the initial formulation.

1 Introduction

The Support Vector Machine (SVM) has been extensively used over years for classification tasks, thanks to its high-rated performances and its robustness. However SVM is not a sparse model since the number of support vectors increases linearly with data. To address this issue, the SVM L_1 [2], replaces the L_2 norm by L_1 norm as penalty to induce more sparsity. One drawback of the L_1 norm resides in its property to keep only one element among a correlated variable set, even if all the variables are relevant. Indeed in some high dimensional problems where $p \gg n$, it can be interesting to keep all the variables of interest (even if they are correlated) and discard the other ones. The elastic net penalty, a mix of (L_1, L_2) penalties, addresses this compromise between sparsity and variable selection and has been applied for regression tasks [3]. The L_2 penalization term balances the effect of L_1 norm penalization by reducing the difference of coefficients associated with correlated variables. Wang [1] studied an elastic net equivalent formulation for classification problems called Doubly Regularized SVM (DrSVM) and proposed an algorithm to solve it via a regularization path method. The DrSVM problem is stated as follows, for some given regularization parameters λ_1 and λ_2:

$$\min_{\beta_0,\beta} J(\beta_0,\beta) = \sum_{i=1}^{n} \max\big(0, 1 - y_i(\beta_0 + \beta^T x_i)\big) + \tfrac{1}{2}\lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \qquad (1)$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

The main issue is that the initial problem (1) is not differentiable due to the Hinge function and L_1 norm. Wang chose to cast it into an equivalent constrained problem and built a regularization path according to a parameter that controls the L_1 constraint. Since the regularization path is not directly built from λ_1, the λ_1-regularization path has to be deduced afterwards. Moreover, for very sparse configuration ($\lambda_2 \ll 1$), the algorithm suffers from a lack of stability. However we decided to solve the initial problem via sub-gradient theory and built a regularization path directly from λ_1. This path construction improves the robustness of DrSVM for sparse models. In this paper, we introduce the initial DrSVM formulation and detail the novel formulation using the sub-gradient theory. Then, in order to improve the robustness and validity of this new formulation, we propose to apply it on both toy dataset and real data.

2 DrSVM algorithm

Here we detail the initial DrSVM resolution and introduce some notations: let $X = (x_i)_{1\leq i\leq n}$ be the learning database and $Y = (y_i)_{1\leq i\leq n}$ the label vector, (λ_1,λ_2) the regularization parameters associated respectively to L_1 and L_2 norms, $f(x) = \beta_0 + \beta^T x$ a linear classifier, $r_i = 1 - y_i f(x_i)$ the residual error at point x_i. Due to the non differentiability of the hinge loss and the L_1 norm, the definition of the following sets arises naturally: $\mathcal{R} = \{i, r_i < 0\}$, $\mathcal{E} = \{i, r_i = 0\}$, $\mathcal{L} = \{i, r_i > 0\}$, $\mathcal{V}_\beta = \{j, \beta_j \neq 0\}$, $\mathcal{V}_0 = \{j, \beta_j = 0\}$. We note $\beta_{\mathcal{V}_\beta}$ the vector β projected on the subscript set $\mathcal{V}_\beta$ and $|\mathcal{V}_\beta|$ the cardinal of $\mathcal{V}_\beta$. Wang proposed a regularization path algorithm to solve the DrSVM problem (1). To this end, the initial DrSVM problem (1) has been recast as:

$$\begin{cases} \min\limits_{\beta_0,\beta,\epsilon} & J_2(\beta_0,\beta,\epsilon) = \sum\limits_{i=1}^{n} \epsilon_i + \frac{1}{2}\lambda_2 \left\|\beta\right\|_2^2 \\ \text{with} & \forall i, r_i \leq \epsilon_i \ ; \ \forall i, \epsilon_i \geq 0 \ ; \ \left\|\beta\right\|_1 \leq s, \end{cases} \tag{2}$$

where the parameter s controls the influence of the L_1 constraint, and ϵ is the slack variable vector. The regularization path proposed by Wang is built according to s and will be called the s-path algorithm. The Lagrangian function is defined by: $L(\beta_0,\beta,\alpha,\eta) = J_2(\beta,\epsilon) + \sum_{i=1}^{n} \alpha_i(r_i\text{-}\epsilon_i) + \eta(\left\|\beta\right\|_1 - s) + \sum_{i=1}^{n} \mu_i\epsilon_i$, where α, η and μ are the Lagrangian coefficients. Writing the KKT conditions of (2) leads to the following $|\mathcal{E}|+|\mathcal{V}_\beta|+2$ unknowns/equations system:

$$\begin{cases} \forall j \in \mathcal{V}_\beta \ , & \lambda_2 \dfrac{\Delta\beta_j}{\Delta s} - \displaystyle\sum_{i\in\mathcal{E}} \frac{\Delta\alpha_i}{\Delta s} y_i x_{ij} + \frac{\Delta\eta}{\Delta s}\,\mathrm{sign}(\beta_j) = 0 & (a) \\ & \displaystyle\sum_{i\in\mathcal{E}} \frac{\Delta\alpha_i}{\Delta s} y_i = 0 & (b) \\ \forall i \in \mathcal{E} \ , & \dfrac{\Delta\beta_0}{\Delta s} + \displaystyle\sum_{j\in\mathcal{V}_\beta} \frac{\Delta\beta_j}{\Delta s} x_{ij} = 0 & (c) \\ & \displaystyle\sum_{j\in\mathcal{V}_\beta} \mathrm{sign}(\beta_j)\frac{\Delta\beta_j}{\Delta s} = 1 & (d) \end{cases} \tag{3}$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

The s-path algorithm proceeds into two steps. First step is the computation of $(\alpha_{\mathcal{E}}, \beta_0, \beta_{\mathcal{V}_\beta}, \eta)$ derivatives by inverting 3. Second step is break points detection (changes of the system 3). The breaking points are computed by considering all the events that could occur: a point leaves the set $\mathcal{E}$, a point hits the set $\mathcal{E}$, a variable becomes inactive, a variable becomes active or $\eta = 0$ (stopping condition). During each iteration, α and η are updated, the Lagrangian parameter η can be assimilated to the parameter λ_1 of the problem (1). The regularization path according to λ_1 can be easily deduced from the s regularization path.
An helpful property of the DrSVM is its ability to approximate the L_1 SVM for relatively lows values of λ_2. So, if (β_0,β) are piecewise linear according to the threshold parameter s, this is not generally the case for the Lagrangian coefficients α, especially for low values of λ_2. Low λ_2 values tend to prevent points to leave $\mathcal{E}$ and variable activation which is likely to lead to $|\mathcal{E}|>|\mathcal{V}|+1$ and over-determination of (β_0,$\beta_{\mathcal{V}_\beta}$). In this situation it is necessary to update α and η while keeping s constant in order to reach a state where the regularization path can be continued. To overcome this problem, we studied the initial problem (1) via the sub-gradient theory and decided to build the regularization path according to λ_1 rather than s, which is more efficient. Moreover we will demonstrate that it automatically avoids the over determination of (β_0,$\beta_{\mathcal{V}_\beta}$). Note that the Lagrange formulation and associated KKT conditions could have been used to derive equations but at the price of more tedious computations.

3 Formulation of DrSVM via sub-gradient theory

The sub-gradient theory extends the concept of gradient in the case of non differentiability. More precisely, for some vector space X, if $h : X \to \mathbb{R}$ is a convex function, its sub-gradient $\partial h(a)$ at the point a is defined by $\partial h(a) = \{g \in X,\ \forall x \in X,\ f(x) \geq h(a) + g^T(x-a)\}$. By definition, a sub-gradient is a set of vectors that respects some constraints but can be taken for an unknown constrained parameter. We will keep this assimilation for the rest of the paper. The non differentiable problem (1) involved only real convex functions so the sub-gradient is well defined. The sub-gradient of $\mathcal{L}$ according to respectively β_0 and β, are calculated by composition using the formula of the sub-gradient of max and absolute functions: $\partial_{\beta_0} J(\beta_0, \beta, \eta) = -\sum_{i=1}^{n} y_i \alpha_i$ and $\partial_{\beta} J(\beta_0, \beta, \eta) = -\sum_{i=1}^{n} y_i \alpha_i x_i + \lambda_2 \beta + \lambda_1 \gamma$, with γ and α the sub-gradients associated respectively with the hinge loss and the L_1 norm. They fulfil the following conditions: $\alpha_{\mathcal{R}} = 0$, $\alpha_{\mathcal{E}} \in [0,1]$, $\alpha_{\mathcal{L}} = 1$ and $\gamma_{\mathcal{V}_\beta} = \text{sign}(\beta)_{\mathcal{V}_\beta}$, $\gamma_{\mathcal{V}_0} \in [-1,1]$. Note also that α is proportional to Lagrange coefficients of (2). The differentiation with respect to λ_1 of the optimality conditions and $r_i = 0$ gives a $|\mathcal{E}|+|\mathcal{V}_\beta|+1$ unknowns/equations system:

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

$$\begin{cases} \forall j \in \mathcal{V}_\beta\ , & \lambda_2 \dfrac{\Delta\beta_j}{\Delta\lambda_1} - \displaystyle\sum_{i\in\mathcal{E}} \frac{\Delta\alpha_i}{\Delta\lambda_1} y_i x_{ij} = -\text{sign}(\beta_j) & (a) \\ & \displaystyle\sum_{i\in\mathcal{E}} \frac{\Delta\alpha_i}{\Delta\lambda_1} y_i = 0 & (b) \\ \forall i \in \mathcal{E}\ , & \dfrac{\Delta\beta_0}{\Delta\lambda_1} + \displaystyle\sum_{j\in\mathcal{V}_\beta} \frac{\Delta\beta_j}{\Delta\lambda_1} x_{ij} = 0 & (c) \end{cases} \tag{4}$$

Parameters ($\alpha_\mathcal{E}$,β_0,$\beta_{\mathcal{V}_\beta}$) are piecewise linear according to λ_1 and thus a λ_1-regularization path algorithm can be derived. As seen before, the computation of the whole λ_1-path is a two steps procedure. The computation of ($\alpha_\mathcal{E}$, β_0, $\beta_{\mathcal{V}_\beta}$) derivatives with respect to λ_1 are given by inverting (4) and break points detection (changes of the system (4)). The possible events are the same than as those for the s-path but the detection's conditions are different. Let's study the case of sparse models which leads naturally to $|\mathcal{E}|=|\mathcal{V}|+1$. In this situation (a,b) implies $\Delta\alpha_\mathcal{E}/\Delta\lambda_1 \neq 0$ (variation) while (c) implies $\Delta\beta_0/\Delta\lambda_1 = \Delta\beta_{\mathcal{V}_\beta}/\Delta\lambda_1 = 0$, which prevents the following events: a variable becomes inactive or a point hits $\mathcal{E}$. The only events possible are: a point leaves from $\mathcal{E}$ or a variable becomes active which leads to $|\mathcal{E}|<|\mathcal{V}|+1$ so the λ_1 path automatically avoids the problem of over determination of (β_0,$\beta_{\mathcal{V}_\beta}$).

The initialization of the λ_1-path algorithm requires a particular attention. Initially $\lambda_1 = +\infty$, all variables are inactive. Two cases have to be considered, depending on the balanced or unbalanced nature of the classes.

Balanced case: the mathematical analysis shows that any value of $\beta_0 \in [-1, 1]$ is solution, though the $\mathcal{R}$, $\mathcal{E}$ and $\mathcal{L}$ might not be the same. The choice $\beta_0 = 0$ $\Rightarrow \mathcal{R} = \mathcal{E} = \varnothing$, $\mathcal{L} = [1, n]$, $\alpha = \mathbf{1}$, $\gamma = \mathbf{0}$ and avoids to solve a linear problem to start the main procedure. Then we decrease λ_1 until a sub-gradient component $\gamma_j = \frac{1}{\lambda_1}\sum_{i=1}^n y_i\alpha_i x_{ij}$ violates the constraint: $-1 \leq \gamma_j \leq 1$. This leads to:

$$\lambda_1^0 = \max_{j\in V_0} |\sum_{i=1}^n y_i\alpha_i x_{ij}|\ ;\ j_0 = \arg\max_{j\in V_0} |\sum_{i=1}^n y_i\alpha_i x_{ij}|$$

The first activated variable is β_{j_0}, λ_1 is set to λ_1^0 and the main procedure can be started.

Unbalanced case: lets I^+ and I^- be the plus and minus classes. We suppose $|I^+| > |I^-|$ (the other case is symmetric). The solution is unique: $\beta_0 = 1$, $\mathcal{R} = \varnothing$, $\mathcal{E} = I^+$ and $\mathcal{L} = I^-$. The value λ_1 is infinite and decreases to λ_1^0 evaluated as follows:

$$\begin{cases} \min\limits_{\lambda_1, \alpha_i\in\mathcal{E}} \lambda_1 \\ \displaystyle\sum_{i=1}^n y_i\alpha_i = 0\ ;\ \forall i \in \mathcal{E}\ ,\ 0 \leq \alpha_i \leq 1\ ;\ \forall j \in \mathcal{V}_0\ ;\ -\lambda_1 \leq |\sum_{i=1}^n \alpha_i y_i x_{ij}| \leq \lambda_1 \end{cases} \tag{5}$$

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Once (5) is solved, the main procedure can start. One can note that Wang proposed a different way by solving a linear system depending on $s > 0$, to determine the state of $\mathcal{R}$, $\mathcal{E}$ and $\mathcal{L}$. Although this initialization is valid, s has to be chosen appropriately. This can be problematic if no prior is known for the dynamic of the regularization path. Indeed if the value of s is too high a large part of the regularization path can be skipped, which can be prejudicial to sparse models building.

4 Experiments

To demonstrate the validity of our λ_1 regularization path, we reproduced some results provided by Wang [1] on a toy data protocol and real MNIST data. Then, we show an exemple of discontinuous behavior of the s-path, while our λ_1-path algorithm provides the whole set of solutions.

4.1 Toy data

This dataset is composed of two balanced classes of dimension p. The n^- points of minus class are drawn from a normal distribution of mean $\mu_- = (-\mathbf{1}_{5,1}, \mathbf{0}_{p-5,1})$ with covariance matrix Σ. The n^+ points of plus class are drawn from a normal distribution of mean $\mu_+ = (\mathbf{1}_{5,1}, \mathbf{0}_{p-5,1})$ with the same covariance matrix Σ. A sub part of Σ called Σ^* fits the covariance between the five variables. As Wang proposed we test three configurations: Σ_1^*, Σ_2^* and Σ_3^*. We use a validation set of 20000 points to find the optimal parameters and a test set of 20000 points. The analysis of table (1) shows that the two algorithms provide the same results.

Σ^*	n	p	p_0	DrSVM(s)	DrSVM(λ_1)
Σ_1^*	100	10	5	0.1410	0.1413
	50	300	5	0.1658	0.1655
Σ_2^*	100	10	5	0.140	0.140
	50	300	5	0.144	0.145
Σ_3^*	100	10	5	0.1175	0.1175
	50	300	5	0.1220	0.125

Table 1: Comparison of the test error of the s-path with our λ_1-path algorithm.

4.2 MNIST data

The MNIST data set is composed of 28×28 pixels images. We have selected the 6 and 9 digits images. Then for each class, we choose 250 images for the learning data, 750 images for the validation data and 1000 images for the test data. Note that Wang used some characteristics for the classification which are not described so we decided to ignore them and to work with the linear discriminant functions, so that $p = 784$. We retained for each regularization formulation the optimal set (λ_1,λ_2) which are very closed and both test errors are equal to 0.5%.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

4.3 An example of discontinuous s-path

This toy data is used to highlight our choice to build a regularization λ_1 path. The minus and plus distributions are generated respectively from normal distributions of respective mean $\mu_- = -\mathbf{1}_{5,1}$, $\mu_+ = \mathbf{1}_{5,1}$, with $\Sigma = (\sigma_{i,j})_{1 \leq i,j \leq 5}$, for $i \neq j$: $\sigma_{i,j} = 0.8$ and for $i = j$: $\sigma_{i,j} = 1$. The value of λ_2 is set to 10^{-3}. This is a low value, so the case $|\mathcal{E}| = |\mathcal{V}| + 1$ happened naturally. We have plotted the path of α coefficients for both s and λ_1 regularization path. The discontinuous points represent the part of the path that keeps s constant while decreasing λ_1.

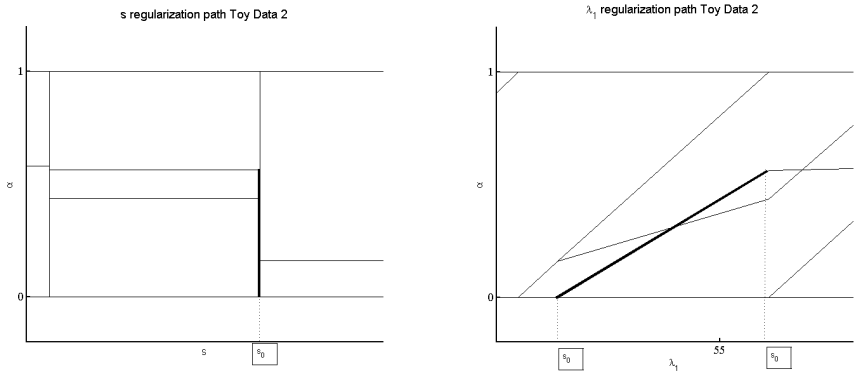

Figure 1: Zoom of regularization paths with respect to s (left) and λ_1 (right). On the s path some solutions α show discontinuities. The discontinuous point s_0 in the s-path corresponds to a segment in the λ_1-path

5 Conclusion

The study of the initial DrSVM problem (1) via the sub-gradient theory led us to build a regularization path directly from L_1 regularization parameter: λ_1. This reformulation of DrSVM problem makes the regularization path more robust to sparse models which makes the DrSVM more generic.
The DrSVM problem, by its own nature (the ability to keep all the pertinent variables), is suited to interpretable models. So it may be interesting to merge the DrSVM with a kernel approach. The last issue would be to determine some early stopping conditions, in order to choose the most appropriate model.

References

[1] Li Wang, Ji Zhu, and Hui Zou. The doubly regularized support vector machine. *Statistica Sinica*, 16(2):589, 2006.

[2] Paul S Bradley and Olvi L Mangasarian. Feature selection via concave minimization and support vector machines. In *ICML*, volume 98, pages 82–90, 1998.

[3] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.