Using Duality as a Method to Solve SVM Regression Problems

Langley DeWitt

1. Introduction

Support vector machines (SVM) are a mathematical technique used for data classification and regression. The paper *Regression tasks in machine learning via Fenchel duality* gives an overview of how they work and how optimal solutions to support vector regression problems can be found. The paper starts by introducing the properties that an SVM should have and what it means for a specific SVM function to be good fit for a data set. It then describes how finding a dual problem can make finding the solution to the original regression problem easier due to possible differentiability issues that can arise in minimization problem of finding the best fit for the data.

When solving regression problems, a common approach is to find the gradient of the loss function (A function designed to measure how well the function that we

have generated fits the data) with respect to the optimization parameters in the fit function. The gradient is then used to vary the fit function in a way that will minimizes the loss function and thus give better fit for the data. This technique is called gradient descent. However, a problem arises if the loss function is not differentiable. This will cause problems, because if the loss function is not differentiable the gradient cannot be computed, meaning that it will be impossible for the algorithm to minimize the loss function.

Reproducing Kernel Hilbert Space[8]

Before we can start defining what a support vector machine is we must define reproducing kernel Hilbert space. Hilbert space is a vector space with infinite dimensionality, and a reproducing kernel Hilbert space is a Hilbert space with specific properties. A Hilbert space H of functions defined on set E is said to be a reproducing kernel Hilbert space if there exists a function $k(x,y)$ on $E \times E$ such that

1. $k(\cdot, y) \in H$ for any $y \in E$
2. $< f, k(\cdot, y) >= f(y)$ for all $f \in H$

This function is called the reproducing kernel for H.

3. SVM Definition

A support vector machine is defined to be a function f that is a member of H_k where H_k is the reproducing kernel Hilbert space corresponding to the kernel k(x, y). Here the kernel k is assumed to be both symmetric and finitely positive semidefinite. In this context symmetric means that k(x,y) = k(y,x) and finitely positive semidefinite means that for any

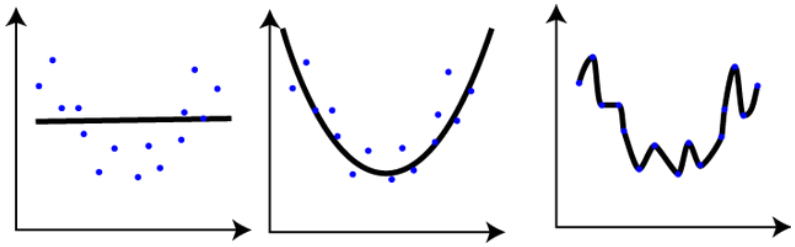
$\{x_1, \ldots, x_m\} \subset \mathbb{R}^n$ and $a \in \mathbb{R}^m$ it is true that $\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j k(x_i, x_j) \geq 0$. We shall take our f to be:

$$f(u) = \sum_{i=1}^{N} c_i k(u, x_i)$$

Where the vectors x_i are the input vectors of the data set. Using the representor theorem we will be able to show that the $f \in H_k$ that best fits our data will take this form. In this definition of f any x_i for which $c_i \neq 0$ is called a support vector of f.

4. Measuring the Quality of an SVM

In order to determine the quality of a support vector machine for modeling a data set we must define a method to quantitatively measure how well the function fits our data set. To do this we must define a loss function *v* that will compare the output of our function at data points x_i with the expected value of the function at those specific inputs given by y_i. This is however not enough to determine if a function is a good fit for the data because it is possible for a match the data perfectly, but not reveal any information about the phenomenon that created the data. For example, many data sets could be matched perfectly be simply creating a piecewise function that would create straight lines connecting the data points, however in the majority of cases this would be a bad fit for the data.



From left to right:

Under fit, just right, over fit

[4]

To avoid this we must introduce a smoothness functional into our measurement of f to insure that it will map points in $\mathbb{R}^n$ that are close to values in $\mathbb{R}$ that are close. We will call this functional $\Omega(f)$. From this we get that the optimal fit for our data will satisfy the Tikhonov regularization problem:

$$\inf\{C\sum_{i=1}^{N} v(f(x_i), y_i) + \Omega(f)/2\} \quad [2]$$

In this expression C>0 is called the *regularization parameter* [2] and is used to set the tradeoff between under and over fitting to the data set.

For the smoothness functional we can take $\Omega(f) = \|f\|_k^2$. Where $\|f\|_k$ signifies the norm of f on the Hilbert space H_k.

Gram Matrix[7]- The square matrix consisting on pairwise scalar products of elements in a Hilbert space or a pre-Hilbert space.

There exists a ϕ such that $k(x, y) = <\phi(x), \phi(y)>$, so k has a Gram matrix and it will take the form:

$$K = \begin{matrix} k(x_1,x_1) & \dots & k(x_N,x_1) \\ \vdots & \ddots & \vdots \\ k(x_1,x_N) & \dots & k(x_N,x_N) \end{matrix}$$

Because *k* is symmetric and positive semidefinite *K must be as well*. From this matrix we get the results $f(x_i) = (Kc)_i$ and $\|f\|_k = c^T Kc$, so the optimization problem becomes:

$$P_{gen} \quad \inf_{c\in\mathbb{R}^n}\{C\sum_{i=1}^{N} v((KC)_i, y_i) + c^T Kc/2\} \quad [2]$$

5. Representor theorem

The representor theorem states that if an SVM optimization problem can be expressed as:

$$\inf\{C\sum_{i=1}^{N} v(f(x_i), y_i) + g(\|f\|_k)\}$$

and g is non-decreasing, then the optimal solution *f* can be expressed as

$$f(u) = \sum_{i=1}^{N} c_i k(u, x_i)$$

Proof [1]

Let f_s(u)be the projection of $f(u)$onto the subspace of H_k defined by:

$$span(k(u, x_i), 1 \le i \le N$$

With f_T(u) being the component of *f* that is perpendicular to f_s

$$\|f\|^2 = \|f_s\|^2 + \|f_T\|^2 \ge \|f_s\|^2$$

Because g is non-decreasing

$$g(\|f\|_k^2) \ge g(\|f_s\|_k^2)$$

Therefore *g* is minimized if *f* lie in the subspace. In addition *k* has the reproducing property so:

$$f(x_i) =< f, k(x_i, u) \ge< f_s, k(x_i, u) > +< f_T, k(x_i, u) \ge$$
$$< f_s, k(x_i, u) >= f_s(x_i)$$

Therefore $\sum_{i=1}^{N} v(f(x_i), y_i)$ depends only on f_s. Together these two result show that the optimal *f* will lie on the subspaces.

6. Using Duality to Solve the Optimization Problem

Because the optimization problem is convex, but not necessarily differentiable it can be difficult to find the solution. To solve this, the paper introduced a dual problem using the Fenchel-Moreau conjugate function defined by:

$$f^*(p) = \sup_{x \in \mathbb{R}^n} \{p^T x - f(x)\} [2]$$

With the property called the Young-Fenchel inequality [2]:

$$f(x) + f^*(p) - p^T x \geq 0$$

And for $\partial f(x) \neq \emptyset$

$$p \in \partial f(x) \quad \leftrightarrow \quad f(x) + f^*(p) = p^T x$$

This gives the dual problem:

$$D_{gen} \quad \sup_{p \in \mathbb{R}^n} \left\{ -C \sum_{i=1}^{N} \left(v(\cdot, y_i)\right)^* \left(\frac{-p_i}{C}\right) - p^T K p \right\} [2]$$

6.1 *Proof for weak duality[2]:*

Because *K* is positive semidefinite and because of the Young-Fenchel inequality we have that:

$$0 \leq C \left[\sum_{i=1}^{N} v(f(x_i), y_i) + \sum_{i=1}^{N} \left(v(\cdot, y_i)\right)^* \left(\frac{-p_i}{C}\right) \right] + \frac{(c-p)^T K (c-p)}{2}$$

$$= C \left[\sum_{i=1}^{N} v(f(x_i), y_i) + \sum_{i=1}^{N} \left(v(\cdot, y_i)\right)^* \left(\frac{-p_i}{C}\right) \right] + P^T K c - P^T K c + \frac{c^T K c}{2} + \frac{p^T K p}{2}$$

$$= C \sum_{i=1}^{N} v(f(x_i), y_i) + \frac{c^T K c}{2} + C \sum_{i=1}^{N} \left(v(f(\cdot), y_i)\right)^* \left(\frac{-p_i}{C}\right) + \frac{p^T K p}{2}$$

Therefore:

$$-C\sum_{i=1}^{N}\left(v(\cdot,y_i)\right)^*\left(\frac{-p_i}{C}\right)-\frac{p^TKp}{2}\leq\ C\sum_{i=1}^{N}v(f(x_i),y_i)+\frac{c^TKc}{2}$$

So weak duality holds.

Now that we have weak duality, if we can show strong duality for the two problems, then the solutions will be the same. This means that we can use the dual problem to find the solution to the original problem. The condition necessary for this is that:

$$(QC)\qquad Im(k)\cap\prod_{i=1}^{N}ri(\ domain(v(u,y_i))\neq\emptyset$$

Where $Im(k)$denotes $\{Kx{:}x\in\mathbb{R}^n\}$ and ri denotes the interior relative to the affine hull.

Affine combination[6]- The affine combination of points $x\in D$ is the set of finite sums with the property

$$\sum_{i=1}^{n}\lambda_i x_i\qquad where\ \ \lambda\in\mathbb{R}^n\ and\ \sum_{i=1}^{n}\lambda_i=1$$

Affine hull[6]- The affine combination all points in the set.

6.2 *Proof of strong duality[2]:*

For the problem

$$P_{gen}\quad\inf_{c\in\mathbb{R}^n}\{C\sum_{i=1}^{N}v((Kc)_i,y_i)+\ c^TKc/2\}$$

Let $g(c)=c^TKc/2$

Because $domain\left(\sum_{i=1}^{N}v(u,y_i)\right)=\prod_{i=1}^{N}ri(\ domain(v(u,y_i))$ we have

$$K(ri(domain(g))\cap\sum_{i=1}^{N}v(u,y_i)=Im(k)\cap\prod_{i=1}^{N}ri(\ domain(v(u,y_i))\neq\emptyset$$

Therefore there exists a $\bar{p}$ such that [3]

$$\inf_{c\in\mathbb{R}^n}\{C\sum_{i=1}^{N} v((Kc)_i, y_i) + c^T Kc/2\} =$$

$$\sup_{p\in\mathbb{R}^n}\left\{-\left(\sum_{i=1}^{N} Cv(\cdot, y_i)\right)^*(-p) - g^*(Kp)\right\} = -\left(\sum_{i=1}^{N} Cv(\cdot, y_i)\right)^*(-\bar{p}) - g^*(K\bar{p})$$

For $q \in \mathbb{R}^n$ we have that

$$g^*(q) = \begin{cases} \frac{1}{2} q^T K^- q \; if q \in Im(K) \\ +\infty \qquad otherwise \end{cases}$$

Where K^- is the Moore-Penrose pseudo –inverse. Therefore we have that

$$g^*(g) = \frac{1}{2}(K\bar{p})^T K^-(K\bar{p}) = \frac{1}{2}\bar{p}^T K K^- K\bar{p} = \frac{1}{2}\bar{p}^T K\bar{p}$$

Therefore

$$\left(\sum_{i=1}^{N} Cv(\cdot, y_i)\right)^*(-\bar{p}) - g^*(K\bar{p}) = C\left(\sum_{i=1}^{N} v(\cdot, y_i)\right)^*(-\frac{\bar{p}}{C}) - g^*(K\bar{p})$$

It follows[5] that there exist $\bar{p}^i \in \mathbb{R}^n$ such that $\sum_{i=1}^{n} \bar{p}^i = \bar{p}$ such that

$$\left(\sum_{i=1}^{N} v(\cdot, y_i)\right)^*\left(\frac{-\bar{p}}{C}\right) = \sum_{i=1}^{N} v^*(\cdot, y_i)\left(\frac{-\bar{p}^i}{C}\right)$$

Therefore

$$\inf_{c\in\mathbb{R}^n}\{C\sum_{i=1}^{N} v((KC)_i, y_i) + c^T Kc/2\} = -C\sum_{i=1}^{N} v^*(\cdot, y_i)\left(\frac{-\bar{p}^i}{C}\right) - \frac{1}{2}\bar{p}^T K\bar{p}$$

Combining this with the weak duality we get that $\bar{p}$ must be the optimal solution to D_{gen}.

The final result that we must have in order to use duality, is to find the conditions for which $P_{gen} - D_{gen} = 0$.

6.3 ***Theorem****[2]:* Let (QC) be satisfied. Then $\bar{c}$ is an optimal solution for P_{gen} if and only if there exists an optimal solution $\bar{p}$ of D_{gen} such that

1. $\frac{-\bar{p}_i}{C} = \partial v(\cdot, y_i)\big((K\bar{c})_i\big)$
2. $K(\bar{c} - \bar{p}) = 0$

Proof[2]:

From strong duality we have that there exists an optimal solution $\bar{p}$ of D_{gen} such that:

$$C\sum_{i=1}^{N} v((K\bar{c})_i, y_i) + C\sum_{i=1}^{N}(K\bar{c})_i\frac{\bar{p}_i}{C} + C\sum_{i=1}^{N} v^*(\cdot, y_i)\left(\frac{-\bar{p}^i}{C}\right) + \frac{1}{2}\bar{p}^T K\,\bar{p} + \frac{\bar{c}^T K\bar{c}}{2} - \bar{p}^T K\bar{c} = 0$$

This is equivalent to

$$\begin{cases} v((K\bar{c})_i, y_i) + v^*(\cdot, y_i)\left(\frac{\bar{p}_i}{C}\right) = (K\bar{c})_i\frac{\bar{p}_i}{C} \\ (\bar{c} - \bar{p})^T K(\bar{c} - \bar{p}) = 0 \end{cases}$$

By the properties of the Young-Fenchel inequality the first statement is equivalent to $-\frac{\bar{p}_i}{C} \in \partial v(\cdot, y_i)\big((K\bar{c})_i\big)$. The second statement is equivalent to $K(\bar{c} - \bar{p}) = 0$.

7. Applying Duality

7.1 The ε-insensitive loss function

$$v_\varepsilon(a, y) = (|a - y| - \varepsilon)_\varepsilon = \begin{cases} 0, & |a - y| \le \varepsilon \\ |a - y|, & else \end{cases}$$

This loss function will not be differentiable, so we cannot use the gradient descent method to find its minimum. Instead we will use the duality method outlined above to create a differentiable dual problem that will give us a minimization of this loss function.

Derivation of the dual problem[2]:

Using the conjugate function as defined earlier we get that for $z, y \in \mathbb{R}$:

$$-\big(v(\cdot, y)\big)^* = -\sup_{a\in\mathbb{R}}\{za - (|a - y| - \varepsilon)_+\} = \inf_{a\in\mathbb{R}}\{-za + (|a - y| - \varepsilon)_+\}$$

From here we can replace $(|a - y| - \varepsilon)_+$ with t such that $t \ge 0$ and $t \ge (|a - y| - \varepsilon)_+$ so we get:

$$-\big(v(\cdot, y)\big)^* = \inf_{a\in\mathbb{R}}\{-za + t\}$$

In order to find the correct value for we use the optimization:

$$-\big(v(\cdot,y)\big)^* = \inf_{a\in\mathbb{R}}\{-za+t\} = \sup_{\lambda,\beta\geq 0}\left\{\inf_{a,t\epsilon\mathbb{R}}\{-za+t+\lambda|a-y|-\lambda\varepsilon-\lambda t-\beta t\}\right\}$$

$$= \sup_{\lambda,\beta\geq 0}\left\{\inf_{a,t\epsilon\mathbb{R}}\{-za+\lambda|a-y|\}+\inf_{a,y\epsilon\mathbb{R}}\{t-\lambda t-\beta t\}-\lambda\varepsilon\right\}$$

Because

$$\inf_{a\epsilon\mathbb{R}}\{-za+\lambda|a-y|\} = \begin{cases}-zy, & \lambda\geq|z|\\ -\infty, & else\end{cases}$$

And

$$\inf_{t\epsilon\mathbb{R}}\{t-\lambda t-\beta t\} = \begin{cases}0, & \lambda+\beta=1\\ -\infty, & else\end{cases}$$

We have that for $|z|\leq 1, \lambda+\beta=1$ so:

$$-\big(v(\cdot,y)\big)^* = \sup_{\lambda,\beta\geq 0}\{-zy-\lambda\varepsilon\}$$

Because we have $|z|$ and a lower bound of λ, and in order for $-zy-\lambda\varepsilon$ to be maximized, λ must be minimized, we get that $\lambda=|z|$. From here we can get a formula for the conjugate loss function:

$$-\big(v(\cdot,y)\big)^*(z) = \begin{cases}-zy-\varepsilon|z|, & |z|\leq 1\\ -\infty, & else\end{cases}$$

This this gives us the dual problem for the ε-insensitive loss function:

$$\sup_{P\in\mathbb{R}^N}\left\{-\frac{P^TKP}{2}+\sum_{i=1}^{N}P_iy_i-\epsilon|P_i|\right\}$$

This problem is differentiable at as long as $P_i\neq 0$, so this optimization problem is much easier to solve than the original problem.

7.2 The Huber Loss Function

This loss function is defined by

$$v_H(a,y) = \begin{cases} \varepsilon|a-y| - \dfrac{\varepsilon^2}{2}, & |a-y| > \varepsilon \\ \dfrac{1}{2}|a-y|^2, & |a-y| \le \varepsilon \end{cases}$$

This loss function is not differentiable on the transition from $|a-y| > \varepsilon$ to $|a-y| \le \varepsilon$, which makes it difficult to solve directly. However, we can generate a differentiable dual problem that will allow us to optimize using this loss function.

Derivation of the dual problem[2]:

The conjugate for this loss function for $z, y \in \mathbb{R}$ will be:

$$-\big(v_H(a,y)\big)^*(z) = -\sup_{a\in\mathbb{R}}\{za + v_H(a,y)\} = \inf_{a\in\mathbb{R}}\{-za - v_H(a,y)\}$$

$$= \min\left\{ \inf_{\substack{a\in\mathbb{R} \\ |a-y|\le\varepsilon}} \left\{-za + \frac{|a-y|^2}{2}\right\}, \inf_{\substack{a\in\mathbb{R} \\ |a-y|>\varepsilon}} \left\{-za + \varepsilon|a-y| - \frac{\varepsilon^2}{2}\right\} \right\}$$

$$= \min\left\{ \begin{array}{c} \inf\limits_{\substack{a\in\mathbb{R} \\ |a-y|\le\varepsilon}} \left\{-za + \dfrac{|a-y|^2}{2}\right\}, \inf\limits_{\substack{a\in\mathbb{R} \\ a>\varepsilon+y}} \left\{-za + \varepsilon(a-y) - \dfrac{\varepsilon^2}{2}\right\}, \\ \inf\limits_{\substack{a\in\mathbb{R} \\ a>y-\varepsilon}} \left\{-za + \varepsilon(y-a) - \dfrac{\varepsilon^2}{2}\right\} \end{array} \right\}$$

In the first infimum we get

$$\inf_{\substack{a\in\mathbb{R} \\ a>\varepsilon+y}} \left\{-za + \varepsilon|a-y| - \frac{\varepsilon^2}{2}\right\} = \begin{cases} \dfrac{\varepsilon^2}{2} - zy - \dfrac{y^2}{2}, & z < -\varepsilon \\ \dfrac{-z^2}{2} - zy - \dfrac{y^2}{2}, & z\epsilon[-\varepsilon,\varepsilon] \\ \dfrac{\varepsilon^2}{2} - zy - z\varepsilon - \dfrac{y^2}{2}, & z > \varepsilon \end{cases}$$

For the second

$$\inf_{\substack{a\in\mathbb{R} \\ a>\varepsilon+y}} \left\{-za + \varepsilon(a-y) - \frac{\varepsilon^2}{2}\right\} = \begin{cases} \dfrac{\varepsilon^2}{2} - zy - z\varepsilon, & z \le \varepsilon \\ -\infty, & else \end{cases}$$

And finally for the third

$$\inf_{\substack{a\in\mathbb{R}\\ a>y-\varepsilon}}\left\{-za+\varepsilon(y-a)-\frac{\varepsilon^2}{2}\right\}=\begin{cases}\frac{\varepsilon^2}{2}-zy+z\varepsilon, & z\geq-\varepsilon\\ -\infty, & else\end{cases}$$

By combining these we get that

$$-\left(v_H(a,y)\right)^*(z)$$
$$=\begin{cases}\min\left\{\frac{-z^2}{2}-zy-\frac{y^2}{2}, \quad -\frac{\varepsilon^2}{2}+zy-z\varepsilon, \quad \frac{\varepsilon^2}{2}-zy-z\varepsilon\right\}, & z\in[-\varepsilon,\varepsilon]\\ -\infty, \quad else\end{cases}$$
$$=\begin{cases}\frac{-z^2}{2}-zy-\frac{y^2}{2}, & z\in[-\varepsilon,\varepsilon]\\ -\infty, \quad else\end{cases}$$

This gives us the dual problem

$$\sup_{P\in\mathbb{R}^N}\left\{-\frac{P^TKP}{2}+\sum_{i=1}^{N}P_iy_i-\frac{P_i^2}{2C}-\frac{y_i^2}{2}\right\}$$

which is differentiable with respect to P, so it will be possible to find a solution.

8. Conclusion

By using dual optimization problems we are able to greatly expand the variety of loss functions that we can use for to fit SVMs to data. Trying to minimize the loss function directly prevents us from using gradient descent for non-differentiable loss functions, but by finding a dual optimization, the problem can become much simpler to solve. This can be seen in the ε-insensitive loss function (7.1). At the points $|a-z|=\varepsilon$ it is not even continuous, so crossing that boundary in the optimization process would be very difficult. On the other hand the dual optimization does not have this issue. Different loss functions behave differently, and as a result will give different results when used for regression. This means that when we fit a function to our data, some loss functions may perform better in different circumstances. Because of this, by expanding the variety of loss functions at our disposal, we have more opportunities to tune the loss function that we use to our specific regression problem.

Sources:

[1]Bartlet, Peter. "The Envelope Theorem: Numerical Examples." Wolfram Demonstrations Project (2008): n. pag. Representertheoremandkernelexamples. UC Berkeley, Apr. 2008. Web. 6 June 2017. <https://people.eecs.berkeley.edu/~bartlett/courses/281b-sp08/8.pdf>.

[2] Boţ, Radu Ioan, and André Heinrich.Heinrich. "Regression Tasks in Machine Learning via Fenchel Duality." Annals of Operations Research 222.1 (2013): 197-211. Web.

[3]Bo¸t, R. I. (2010). Lecture notes in economics and mathematical systems: Vol. 637. Conjugate duality in convex optimization. Berlin/Heidelberg: Springer.

[4]"General Regression and over Fitting." The Shape of Data. N.p., 29 Sept. 2013. Web. 06 June 2017.

[5]Rockafellar, R. T. (1970). Convex analysis. Princeton: Princeton University Press.

[6]Affine hull. Encyclopedia of Mathematics. URL: http://www.encyclopediaofmath.org/index.php?title=Affine_hull&oldid=38757

[7]Gram matrix. Encyclopedia of Mathematics. URL: http://www.encyclopediaofmath.org/index.php?title=Gram_matrix&oldid=35177

[8] Reproducing-kernel Hilbert space. Encyclopedia of Mathematics. URL: http://www.encyclopediaofmath.org/index.php?title=Reproducing-kernel_Hilbert_space&oldid=37590