SVMs Part 2: Support Vector Classifiers

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Relaxing maximum margin classifier constraints

In a linearly non-separable case, the convex optimization problem

$$\underset{\substack{\boldsymbol{\beta} \in \Re^p \\ \text{and } \beta_0 \in \Re}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 \quad \text{subject to } y_i\left(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0\right) \geq 1 \ \ \forall i$$

has no solution (no pair $\boldsymbol{\beta} \in \Re^p$ and $\beta_0 \in \Re$ provides $y_i\left(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0\right) \geq 1$ $\forall i$). So in looking for good choices of $\boldsymbol{\beta} \in \Re^p$ and $\beta_0 \in \Re$) one might relax the constraints of the problem slightly.

That is, suppose that $\xi_i \geq 0$ and consider the set of constraints

$$y_i\left(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0\right) + \xi_i \geq 1 \ \forall i$$

The ξ_i are "slack" variables and provide some "wiggle room" in search for a hyperplane that "nearly" separates the two classes.

Budget for total slack

The total amount of slack allowed might be controlled by setting a limit

$$\sum_{i=1}^{N} \xi_i \leq C$$

for some positive "budget" C.

If $y_i \left(\mathbf{x}_i' \boldsymbol{\beta} + \beta_0 \right) \geq 0$, training case i is correctly classified. So if for some pair $\boldsymbol{\beta} \in \Re^p$ and $\beta_0 \in \Re$ this holds for all i, the problem is separable. Any non-separable problem must then have at least one negative $y_i \left(\mathbf{x}_i' \boldsymbol{\beta} + \beta_0 \right)$ for any $\boldsymbol{\beta} \in \Re^p$ and $\beta_0 \in \Re$ pair. This in turn requires that the budget C must be at least 1 for a non-separable problem to have a solution even with the addition of slack variables. In fact, a budget C allows for at most C mis-classifications in the training set. And in a non-separable case, C must be large enough so that some choice of $\boldsymbol{\beta} \in \Re^p$ and $\beta_0 \in \Re$ produces a classifier with training error rate no larger than C/N.

Support vector classifier problem

So consider the optimization problem

$$\underset{\substack{\boldsymbol{\beta} \in \Re^p \\ \text{and } \beta_0 \in \Re}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 \quad \text{subject to} \begin{cases} y_i\left(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0\right) + \xi_i \geq 1 \quad \forall i \\ \text{for some } \xi_i \geq 0 \text{ with } \sum_{i=1}^N \xi_i \leq C \end{cases} \tag{1}$$

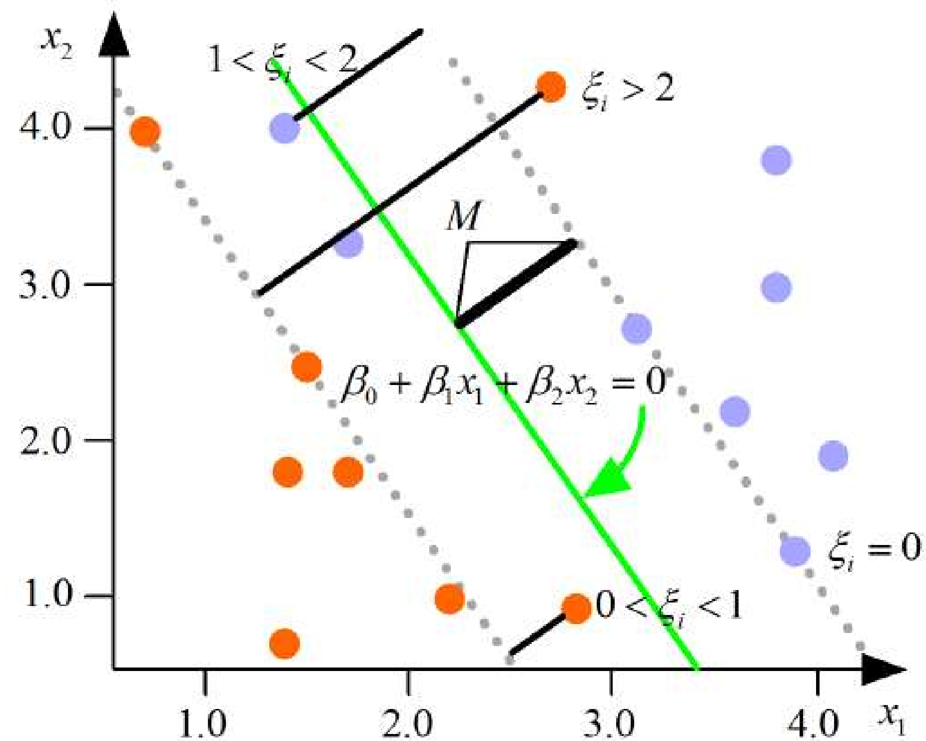
generalizing the third form of the separable problem. Now (1) is equivalent to

$$\underset{\substack{\mathbf{u} \text{ with } \|\mathbf{u}\|=1 \\ \text{and } \beta_0 \in \Re}}{\text{maximize}} \quad M \quad \text{subject to} \begin{cases} y_i\left(\mathbf{x}_i'\mathbf{u} + \beta_0\right) \geq M\left(1 - \xi_i\right) \quad \forall i \\ \text{for some } \xi_i \geq 0 \text{ with } \sum_{i=1}^N \xi_i \leq C \end{cases}$$

generalizing the original form of the separable problem. Here ξ_i is a fraction of the margin M that input $\mathbf{x}_i$ is allowed to be on the "wrong side" of its cushion around the hyperplane defined by $\mathbf{x}'\boldsymbol{\beta} + \beta_0 = 0$.

Small *p=2* problem

The ideas and notation of this development are illustrated in the Figure below for a small $p = 2$ problem.



Penalized (rather than constrained) budget

A more convenient version of (1) is

$$\underset{\substack{\boldsymbol{\beta} \in \Re^p \\ \text{and } \beta_0 \in \Re}}{\text{minimize}} \quad \frac{1}{2}\left\|\boldsymbol{\beta}\right\|^2 + C^* \sum_{i=1}^{N} \xi_i \quad \text{subject to } \begin{cases} y_i\left(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0\right) + \xi_i \geq 1 \quad \forall i \\ \text{for some } \xi_i \geq 0 \end{cases} \tag{2}$$

A nice development on pages 376-378 of Izenman's book provides the following solution to this problem (2) parallel to the development for separable cases.

Dual problem

Generalizing the separable case dual problem

$$\text{maximize} \quad \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\mathbf{H}\boldsymbol{\alpha} \quad \text{subject to } \boldsymbol{\alpha} \geq \mathbf{0} \text{ and } \boldsymbol{\alpha}'\mathbf{y} = 0$$

the present dual problem is for $\underset{N\times N}{\mathbf{H}} = (y_i y_j \mathbf{x}_i' \mathbf{x}_j)$

$$\text{maximize} \quad \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\mathbf{H}\boldsymbol{\alpha} \quad \text{subject to } \mathbf{0} \leq \boldsymbol{\alpha} \leq C^*\mathbf{1} \text{ and } \boldsymbol{\alpha}'\mathbf{y} = 0 \qquad (3)$$

The constraint $\mathbf{0} \leq \boldsymbol{\alpha} \leq C^*\mathbf{1}$ is known as a "box constraint" and the "feasible region" prescribed in (3) is the intersection of a hyperplane defined by $\boldsymbol{\alpha}'\mathbf{y} = 0$ and a "box" in the positive orthant. The $C^* = \infty$ version of this reduces to the "hard margin" separable case.

Properties of the solution

Upon solving (3) for $\boldsymbol{\alpha}^{\text{opt}}$, the optimal $\boldsymbol{\beta} \in \Re^p$ is of the form

$$\boldsymbol{\beta}\left(\boldsymbol{\alpha}^{\text{opt}}\right) = \sum_{i \in \mathcal{SV}} \alpha_i^{\text{opt}} y_i \mathbf{x}_i \tag{4}$$

for $\mathcal{SV}$ the indices of set of **support vectors** $\mathbf{x}_i$ which have $\alpha_i^{\text{opt}} > 0$. The points with $0 < \alpha_i^{\text{opt}} < C^*$ lie on the "edge of the margin" (have $\xi_i = 0$ and lie on the surface of a "slab" of thickness $2M$ around the hyperplane) and the ones with $\alpha_i^{\text{opt}} = C^*$ have $\xi_i > 0$ and lie on the "wrong side" of their surface of the slab. Any of the support vectors on the "edge of the margin" (with $0 < \alpha_i^{\text{opt}} < C^*$) may be used to solve for $\beta_0 \in \Re$ as

$$\beta_0\left(\boldsymbol{\alpha}^{\text{opt}}\right) = y_i - \mathbf{x}_i' \boldsymbol{\beta}\left(\boldsymbol{\alpha}^{\text{opt}}\right) \tag{5}$$

(For reasons of numerical stability it is common practice to average values $y_i - \mathbf{x}_i' \boldsymbol{\beta}\left(\boldsymbol{\alpha}^{\text{opt}}\right)$ for such support vectors in order to evaluate $\beta_0\left(\boldsymbol{\alpha}^{\text{opt}}\right)$.)

Classifier "complexity"

C^* is a regularization parameter and large C^* in (2) corresponds to small C in (1). Identification of a classifier requires only solution of the dual problem (3), evaluation of (4) and (5) to produce linear form

$$g(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} + \beta_0$$

and then classifier $f(\mathbf{x}) = \text{sign}(g(\mathbf{x}))$.

Even when a problem is linearly separable, there may be good reason to use the present formulation with $C^* < \infty$ (and a correspondingly larger margin and more support vectors). Small C^* (large C) corresponds to a "low complexity" classifier with many support vectors contributing to the ultimate form. The exact form of the classifier is less sensitive to a few key data cases than for large C^*. (If the problem were SEL prediction, small C^* would be the "low variance/high bias" case.) Cross-validation is used in practice to choose an appropriate C^*.

Examples for *p=2*

Below is a figure illustrating the impact of C^* on a support vector classifier.