SVMs Part 1: Maximum Margin Classifiers

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

Support vector classifiers-linearly separable cases

Consider a 2-class classification problem. For notational convenience, we'll suppose that output y takes values in $\mathcal{G} = \{-1, 1\}$. We further develop linear classification methodology.

For $\boldsymbol{\beta} \in \Re^p$ and $\beta_0 \in \Re$ we'll consider the form

$$g(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} + \beta_0 \tag{1}$$

and a theoretical predictor/classifier

$$f(\mathbf{x}) = \text{sign}(g(\mathbf{x})) \tag{2}$$

We will approach the problem of choosing $\boldsymbol{\beta}$ and β_0 to in some sense provide a maximal cushion around a hyperplane separating between $\mathbf{x}_i$ with corresponding $y_i = -1$ and $\mathbf{x}_i$ with corresponding $y_i = 1$.

Optimization problem

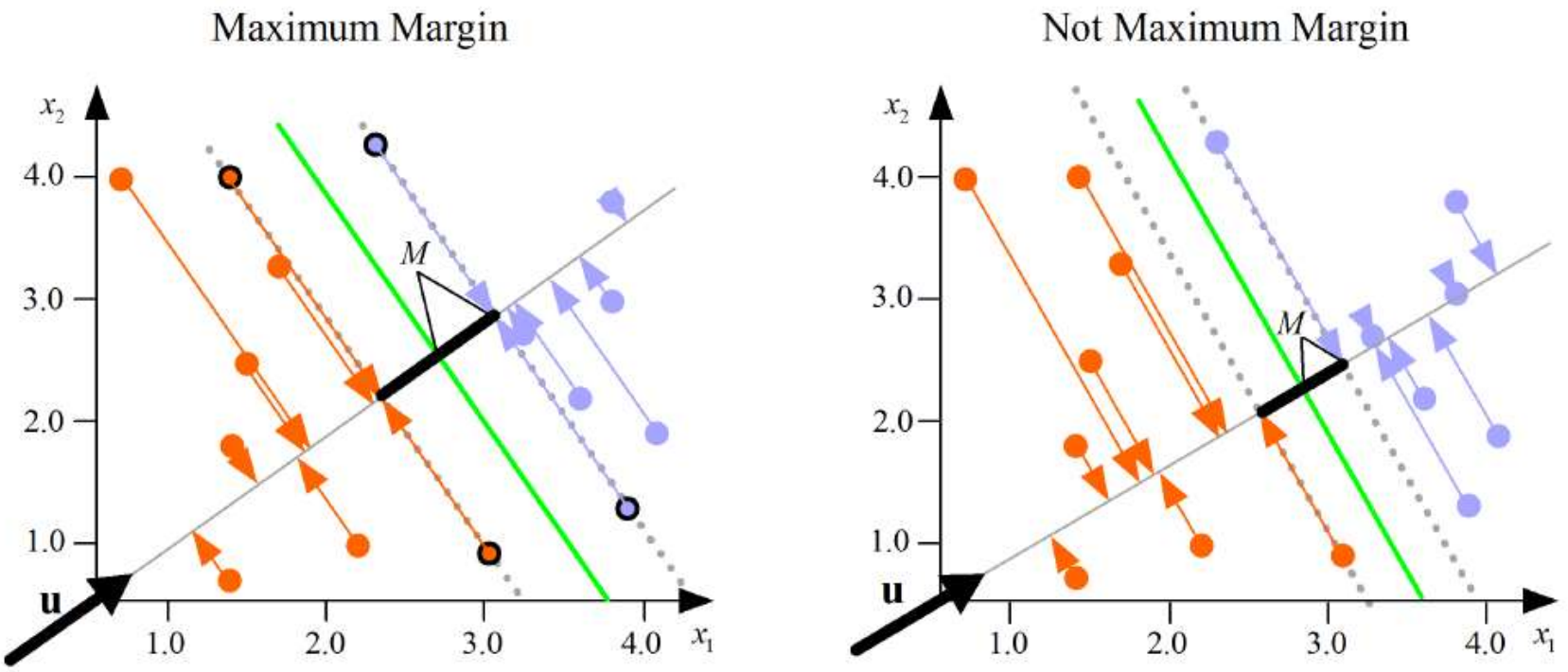
In the case that there is a classifier of form (2) with 0 training error rate, we consider the optimization problem

$$\underset{\mathbf{u} \text{ with } \|\mathbf{u}\| = 1 \text{ and } \beta_0 \in \Re}{\text{maximize}} \quad M \quad \text{subject to } y_i \left(\mathbf{x}_i'\mathbf{u} + \beta_0\right) \geq M \quad \forall i \qquad (3)$$

This can be thought of in terms of choosing a unit vector $\mathbf{u}$ (or direction) in $\Re^p$ so that upon projecting the training input vectors $\mathbf{x}_i$ onto the subspace of multiples of $\mathbf{u}$ there is maximum separation between convex hull of projections of the $\mathbf{x}_i$ with $y_i = -1$ and the convex hull of projections of $\mathbf{x}_i$ with corresponding $y_i = 1$. (The sign on $\mathbf{u}$ is chosen to give the latter larger $\mathbf{x}_i'\mathbf{u}$ than the former.) The resulting classifier might be termed a **maximum margin classifier**.

Maximum margin goal

Here is a $p = 2$ cartoon illustrating the basic "maximum margin" goal



Properties of a solution

Notice that if $\mathbf{u}$ and β_0 solve maximization problem (3), the margin is

$$M = \frac{1}{2}\left(\min_{\substack{\mathbf{x}_i \text{ with} \\ y_i = 1}} \mathbf{x}_i'\mathbf{u} - \max_{\substack{\mathbf{x}_i \text{ with} \\ y_i = -1}} \mathbf{x}_i'\mathbf{u} \right)$$

and the constant that makes the voting function (1) take the value 0 on the separating hyperplane is

$$\beta_0 = -\frac{1}{2}\left(\min_{\substack{\mathbf{x}_i \text{ with} \\ y_i = 1}} \mathbf{x}_i'\mathbf{u} + \max_{\substack{\mathbf{x}_i \text{ with} \\ y_i = -1}} \mathbf{x}_i'\mathbf{u} \right)$$

Second version of the optimization problem

For purposes of applying standard optimization theory and software, it is useful to reformulate the basic problem (3) several ways. First, note that (3) may be rewritten as

$$\underset{\mathbf{u} \text{ with } \|\mathbf{u}\| = 1 \text{ and } \beta_0 \in \Re}{\text{maximize}} \quad M \quad \text{subject to } y_i \left(\mathbf{x}_i' \left(\frac{\mathbf{u}}{M} \right) + \frac{\beta_0}{M} \right) \geq 1 \quad \forall i \qquad (4)$$

Then if we let

$$\boldsymbol{\beta} = \frac{\mathbf{u}}{M}$$

it's the case that

$$\|\boldsymbol{\beta}\| = \frac{1}{M} \quad \text{or} \quad M = \frac{1}{\|\boldsymbol{\beta}\|}$$

Third version of the optimization problem

So (4) can be rewritten

$$\underset{\substack{\boldsymbol{\beta} \in \Re^p \\ \text{and } \beta_0 \in \Re}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2 \quad \text{subject to } y_i\left(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0\right) \geq 1 \quad \forall i \qquad (5)$$

This formulation is that of a convex (quadratic criterion, linear inequality constraints) optimization problem for which there exists standard theory and algorithms.

The so-called primal functional corresponding to (5) is (for $\boldsymbol{\alpha} \in \Re^N$)

$$F_{\mathrm{P}}\left(\boldsymbol{\beta}, \beta_0, \boldsymbol{\alpha}\right) \equiv \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_{i=1}^{N} \alpha_i \left(y_i\left(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0\right) - 1\right) \quad \text{for} \quad \boldsymbol{\alpha} \geq \mathbf{0}$$

To solve (5), one may for each $\boldsymbol{\alpha} \geq \mathbf{0}$ choose $\left(\boldsymbol{\beta}\left(\boldsymbol{\alpha}\right), \beta_0\left(\boldsymbol{\alpha}\right)\right)$ to minimize $F_{\mathrm{P}}\left(\cdot, \cdot, \boldsymbol{\alpha}\right)$ and then choose $\boldsymbol{\alpha} \geq \mathbf{0}$ to *maximize* $F_{\mathrm{P}}\left(\boldsymbol{\beta}\left(\boldsymbol{\alpha}\right), \beta_0\left(\boldsymbol{\alpha}\right), \boldsymbol{\alpha}\right)$.

Karush-Kuhn-Tucker conditions

The "Karush-Kuhn-Tucker conditions" are necessary and sufficient for solution of this optimization problem.

These are the *gradient conditions*

$$\frac{\partial F_{\mathrm{P}}\left(\boldsymbol{\beta}, \beta_0, \boldsymbol{\alpha}\right)}{\partial \beta_0} = -\sum_{i=1}^{N} \alpha_i y_i = 0 \tag{6}$$

and

$$\frac{\partial F_{\mathrm{P}}\left(\boldsymbol{\beta}, \beta_0, \boldsymbol{\alpha}\right)}{\partial \boldsymbol{\beta}} = \boldsymbol{\beta} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = \mathbf{0} \tag{7}$$

the *feasibility conditions*

$$y_i\left(\mathbf{x}_i'\boldsymbol{\beta} + \beta_0\right) - 1 \geq 0 \quad \forall i \tag{8}$$

Karush-Kuhn-Tucker conditions cont.

the *non-negativity conditions*

$$\boldsymbol{\alpha} \geq \mathbf{0}$$

and the *orthogonality conditions*

$$\alpha_i \left(y_i \left(\mathbf{x}_i' \boldsymbol{\beta} + \beta_0 \right) - 1 \right) = 0 \quad \forall i \tag{9}$$

Now (6) and (7) are

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \quad \text{and} \quad \boldsymbol{\beta} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i \equiv \boldsymbol{\beta}(\boldsymbol{\alpha}) \tag{10}$$

and plugging these into $F_{\mathrm{P}}(\boldsymbol{\beta}, \beta_0, \boldsymbol{\alpha})$ gives a function of $\boldsymbol{\alpha}$ only.

Standard optimization theory

That is, write

$$
\begin{aligned}
F_{\mathrm{D}}(\boldsymbol{\alpha}) &\equiv \frac{1}{2}\left\|\boldsymbol{\beta}(\boldsymbol{\alpha})\right\|^2 - \sum_{i=1}^{N} \alpha_i \left(y_i \mathbf{x}_i' \boldsymbol{\beta}(\boldsymbol{\alpha}) - 1\right) \\
&= \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x}_i' \mathbf{x}_j - \sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x}_i' \mathbf{x}_j + \sum_i \alpha_i \\
&= \sum_i \alpha_i - \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j y_i y_j \mathbf{x}_i' \mathbf{x}_j \\
&= \mathbf{1}'\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}'\mathbf{H}\boldsymbol{\alpha}
\end{aligned}
$$

for

$$
\underset{N \times N}{\mathbf{H}} = \left(y_i y_j \mathbf{x}_i' \mathbf{x}_j\right)
$$

Standard optimization theory

Then the "dual" problem is

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

and apparently this problem is easily solved.

Properties of the solution

Now condition (9) implies that if $\alpha_i^{\text{opt}} > 0$

$$y_i \left(\mathbf{x}_i' \boldsymbol{\beta} \left(\boldsymbol{\alpha}^{\text{opt}} \right) + \beta_0 \left(\boldsymbol{\alpha}^{\text{opt}} \right) \right) = 1$$

so that

1. by (8) the corresponding $\mathbf{x}_i$ has minimum $\mathbf{x}_i' \boldsymbol{\beta} \left(\boldsymbol{\alpha}^{\text{opt}} \right)$ for training vectors with $y_i = 1$ or maximum $\mathbf{x}_i' \boldsymbol{\beta} \left(\boldsymbol{\alpha}^{\text{opt}} \right)$ for training vectors with $y_i = -1$ (so that $\mathbf{x}_i$ is a **support vector** for the "slab" of thickness $2M$ around a separating hyperplane),
2. $\beta_0 \left(\boldsymbol{\alpha}^{\text{opt}} \right)$ may be determined using the corresponding $\mathbf{x}_i$ from

$$y_i \beta_0 \left(\boldsymbol{\alpha}^{\text{opt}} \right) = 1 - y_i \mathbf{x}_i' \boldsymbol{\beta} \left(\boldsymbol{\alpha}^{\text{opt}} \right) \quad \text{i.e.} \quad \beta_0 \left(\boldsymbol{\alpha}^{\text{opt}} \right) = y_i - \mathbf{x}_i' \boldsymbol{\beta} \left(\boldsymbol{\alpha}^{\text{opt}} \right)$$

(apparently 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 support vectors in order to evaluate $\beta_0 \left(\boldsymbol{\alpha}^{\text{opt}} \right)$) and,

Properties of the solution cont.

3.

$$1 = y_i \beta_0 \left(\boldsymbol{\alpha}^{\text{opt}}\right) + y_i \left(\sum_{j=1}^{N} \alpha_j^{\text{opt}} y_j \mathbf{x}_j \right)' \mathbf{x}_i$$

$$= y_i \beta_0 \left(\boldsymbol{\alpha}^{\text{opt}}\right) + \sum_{j=1}^{N} \alpha_j^{\text{opt}} y_j y_i \mathbf{x}_j' \mathbf{x}_i$$

The fact (10) that $\boldsymbol{\beta}(\boldsymbol{\alpha}) \equiv \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$ implies that only the training cases with $\alpha_i > 0$ (typically corresponding to a relatively few support vectors) determine the nature of the solution to this optimization problem.

Properties of the solution cont.

Further, for $\mathcal{SV}$ the set indices of support vectors in the problem,

$$\begin{aligned}
\left\|\boldsymbol{\beta}\left(\boldsymbol{\alpha}^{\text{opt}}\right)\right\|^2 &= \sum_{i\in\mathcal{SV}}\sum_{j\in\mathcal{SV}} \alpha_i^{\text{opt}}\alpha_j^{\text{opt}} y_i y_j \mathbf{x}_i'\mathbf{x}_j \\
&= \sum_{i\in\mathcal{SV}} \alpha_i^{\text{opt}} \sum_{j\in\mathcal{SV}} \alpha_j^{\text{opt}} y_i y_j \mathbf{x}_j'\mathbf{x}_i \\
&= \sum_{i\in\mathcal{SV}} \alpha_i^{\text{opt}} \left(1 - y_i\beta_0\left(\boldsymbol{\alpha}^{\text{opt}}\right)\right) \\
&= \sum_{i\in\mathcal{SV}} \alpha_i^{\text{opt}}
\end{aligned}$$

the next to last of these following from 3. above, and the last following from the gradient condition (6). Then the **margin** for this problem is simply computed in terms of the α_i^{opt}s as

$$M = \frac{1}{\left\|\boldsymbol{\beta}\left(\boldsymbol{\alpha}^{\text{opt}}\right)\right\|} = \frac{1}{\sqrt{\sum_{i\in\mathcal{SV}} \alpha_i^{\text{opt}}}}$$