1 Support Vector Machines

1.1 Basic linear geometry

A hyperplane in $\mathbb{R}^p$ is given by

$$\mathcal{H} = \{X \in \mathbb{R}^p : h(X) = \beta_0 + \beta^\top X = 0\}$$

1. The vector β is normal to $\mathcal{H}$ where $||\beta||_2 = 1$. To see this, let $X, X' \in \mathcal{H}$. Then $\beta^T(X - X') = 0$
2. Important: For any point $X \in \mathbb{R}^p$, the (signed) length of its orthogonal complement to $\mathcal{H}$ is $h(X)$.
   ◦ For $X_0 \in \mathcal{H}$ and $X \in \Re^p$, it is

$$\begin{aligned} \langle \beta, X - X_0 \rangle &= \beta^T(X - X_0) + \beta_0 - \beta_0 \\ &= (\beta^T X + \beta_0) + (\beta^T X_0 + \beta_0) = h(X) + h(X_0) = h(X). \end{aligned}$$

1.2 Support vector machines(SVM)

Let $Y_i \in \{-1, 1\}$

A classification rule induced by a hyperplane is

$$g(X) = \text{sgn}(X^\top \beta + \beta_0)$$

where $\text{sgn}(X)$ is 1 if $X > 0$ and -1 if $X < 0$.

1.3 Seperating hyperplanes

Our classification rule is based on a hyperplane $\mathcal{H}$

$$g(X) = \text{sgn}(X^\top \beta + \beta_0)$$

A correct classification is one such that $h(X)Y > 0$. The larger the quantity $Yh(X)$, the more "sure" the classification. (Reminder: The signed distance to $\mathcal{H}$ is $h(X)$.) Under classical separability, we can find a function such that $Y_i h(X_i) > 0\ \forall i$. That is, makes perfect training classifications via g.

Note: Figure 1 shows that we are able to find the hyperplane that creates the biggest margin between the training points for class -1 and 1.[1]

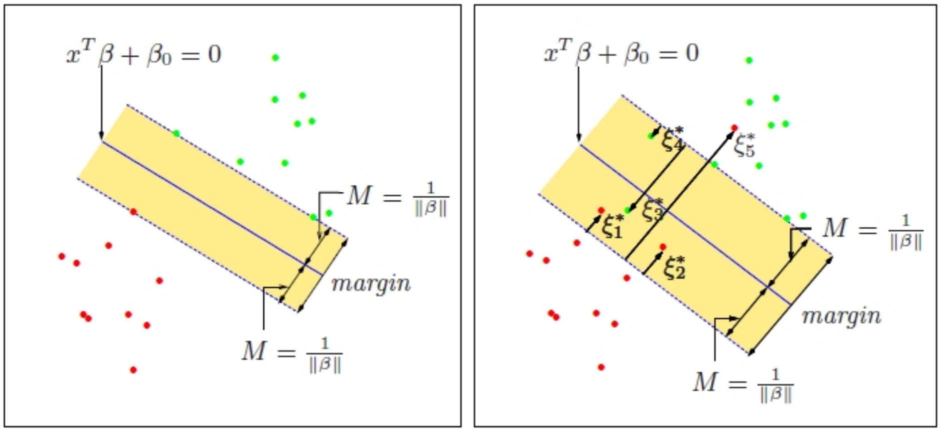

Figure 1: Support vector classifiers. The left panel shows the separable case. The decision boundary is the solid line, while broken lines bound the shaded maximal margin of width 2M. The right panel shows the nonseparable (overlap) case.

1.4 Optimal separating hyperplane

This idea can be encoded in the following convex program

$$\max_{\beta_0,\beta} M \text{ subject to}$$

$$Y_i h(X_i) \geq M \text{ for each } i \text{ and } ||\beta||_2 = 1$$

Intuition:

- We know that $Y_i h(X_i) > 0 \ \Rightarrow g(X_i) = Y_i$. Hence, larger $Y_i h(X_i) \ \Rightarrow$ "more" correct classification.
- For "more" to have any meaning, we need to normalize β, thus the other constraint.

References

[1] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. New York: Springer.