BIOS 835: Support Vector Classifier

Alexander McLain

October 11, 2023

Outline

Introduction

Perfect classifier

No perfect classifier

Constrained Optimization

Notation and problem set up.

Recall two of our goals in subset selection of linear models

- Let $\boldsymbol{X} \in \mathbb{R}^p$ and $Y \in \{-1, 1\}$, and we want to predict what class each observation comes from $\boldsymbol{X}$.
- **Separating hyperplane:** exists if a plane can perfectly separate the data into classes.

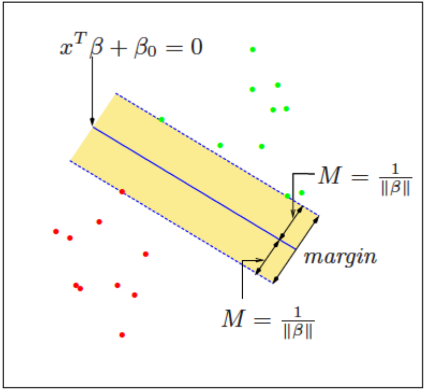

Figure: Data with a separating hyperplane, but which hyperplane to use?

Which hyperplane to use?

- d_- = distance from the SH to the nearest negative point
- d_+ = distance from the SH to the nearest positive point
- The margin is defined as

$$d = d_- + d_+$$

- If the data are linearly separable, then there exists a β_0 and $\boldsymbol{\beta}$ such that

$$\begin{aligned} \beta_0 + \boldsymbol{X}_i'\boldsymbol{\beta} &\geq +1 \qquad \text{if } y_i = +1 \\ \beta_0 + \boldsymbol{X}_i'\boldsymbol{\beta} &\leq -1 \qquad \text{if } y_i = -1 \end{aligned}$$

Which hyperplane to use?

- Define two hyperplanes:

$$
\begin{aligned}
H_{+1} : (\beta_0 - 1) + \boldsymbol{X}'\beta &= 0 \\
H_{-1} : (\beta_0 + 1) + \boldsymbol{X}'\beta &= 0
\end{aligned}
$$

- Points that lie on H_{+1} or H_{-1} are said to be support points.

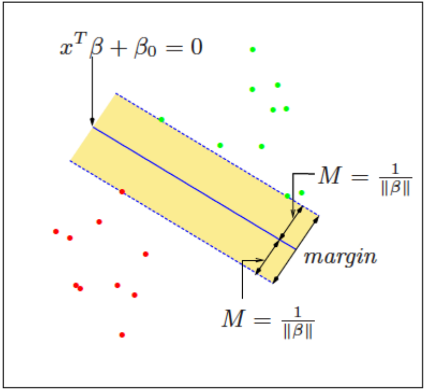

Figure: Support points are those that lie on the dotted line.

Which hyperplane to use?

- Suppose, $\boldsymbol{X}_{-1}$ lie on H_{-1} and $\boldsymbol{X}_{+1}$ lie on H_{+1}.
- Then,

$$\begin{aligned} \beta_0 + \boldsymbol{X}'_{+1}\boldsymbol{\beta} &= +1 \\ \beta_0 + \boldsymbol{X}'_{-1}\boldsymbol{\beta} &= -1 \end{aligned}$$

- The perpendicular distance from $\boldsymbol{X}_{-1}$ and $\boldsymbol{X}_{+1}$ to the hyperplane $\beta_0 + \boldsymbol{X}'\beta = 0$ is

$$\begin{aligned} d_+ &= \frac{|\beta_0 + \boldsymbol{X}'_{+1}\boldsymbol{\beta}|}{||\boldsymbol{\beta}||} = \frac{1}{||\boldsymbol{\beta}||} \\ d_- &= \frac{|\beta_0 + \boldsymbol{X}'_{-1}\boldsymbol{\beta}|}{||\boldsymbol{\beta}||} = \frac{1}{||\boldsymbol{\beta}||} \end{aligned}$$

Which hyperplane to use?

- Thus $d = \frac{2}{||\beta||}$ and a criteria for choosing a hyperplane is to maximize $d = \frac{2}{||\beta||}$
- Similarly, we seek to minimize $\frac{||\beta||^2}{2}$ such that

$$y_i(\beta_0 + \boldsymbol{X}_i'\beta) \geq 1$$

- This is a constrained optimization problem, which we'll address later

What if there are not separating hyperplanes?

- Be slack!! (i.e., use a "soft margin" solution)
- Let $\xi_i \geq 0$ be the slack variable for each data point.

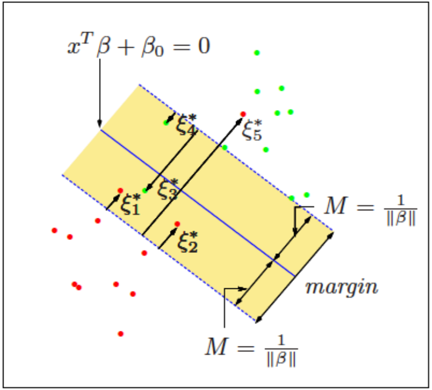

Figure: Here, there does not exist a hyperplane that perfectly separates the data.

What if there are not separating hyperplanes?

- We now look to control $d = \frac{2}{||\beta||}$ and $\sum_{i=1}^{n} \xi_i$.
- That is we seek to

$$\min ||\boldsymbol{\beta}|| \text{ such that } y_i(\beta_0 + \boldsymbol{X}_i'\boldsymbol{\beta}) \geq 1 - \xi_i \quad \forall i,$$

  such that $\xi_i \geq 0$ and $\sum_{i=1}^{n} \xi_i <$ Constraint.
- We can re-write this as

$$\min_{\beta_0, \beta} \left\{ \frac{1}{2} ||\boldsymbol{\beta}||^2 + C \sum_{i=1}^{n} \xi_i \right\}$$

  such that $\xi_i \geq 0$ and $y_i(\beta_0 + \boldsymbol{X}_i'\boldsymbol{\beta}) \geq 1 - \xi_i \quad \forall i$.

Side-bar on constrained optimization

- Use Lagrangian multipliers.
- The constraints are

$$\xi_i \geq 0 \quad y_i(\beta_0 + \boldsymbol{X}_1'\boldsymbol{\beta}) - 1 \geq 0 \quad i = 1, 2, \ldots, n$$

- This gives the Lagrangian (primal) function as

$$F_P(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2}||\boldsymbol{\beta}||^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\{y_i(\beta_0 + \boldsymbol{X}_i'\boldsymbol{\beta}) - (1 - \xi_i)\} - \sum_{i=1}^{n}\mu_i\xi_i$$

where $\boldsymbol{\alpha} = (\alpha_1 \ldots, \alpha_n)$ are the Lagrangian coefficients.

Side-bar on constrained optimization

- After solving the derivatives of the Lagrangian (primal) function wrt $(\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha})$ we get the Lagrangian (Wolfe) dual objective function

$$F_D(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\boldsymbol{X}_i' \boldsymbol{X}_j)$$

- We maximize $F_D(\boldsymbol{\alpha})$ such that $\boldsymbol{\alpha} \geq 0$ and $\boldsymbol{\alpha}'\boldsymbol{y} = 0$.
- This together with the Karush-Kuhn-Tucker (KKN) conditions, results in

$$\hat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \boldsymbol{X}_i$$

where $\hat{\alpha}_i > 0$ iff $y_i(\beta_0 + \boldsymbol{X}_i'\boldsymbol{\beta}) = (1 - \xi_i)$ (the "support" points).