1 Optimal Separating Hyperplane

The idea can be encoded in the following *convex program*

$maxM_{\beta_0,\beta}$ subject to $Y_ih(X_i) \geq M$ for each i and $||\beta||_2 = 1$

Intuition:

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

We can rewrite the original program :

$$min_{\beta_0,\beta}\frac{1}{2}||\beta||_2^2 \ subject\ to\ Y_ih(X_i) \geq 1,\ for\ each\ i \tag{1}$$

Now, we can convert this constrained optimization problem into the Lagrangian (primal) form

$$min_{\beta_0,\beta}\frac{1}{2}||\beta||_2^2 - \sum_{i=1}^{n}\alpha_i[Y_i(X_i^T\beta + \beta_0) - 1] \tag{2}$$

Derivatives with respect to β_0 and β

- $\beta = \sum_{i=1}^n \alpha_i Y_i X_i$
- $0 = \sum_{i=1}^n \alpha_i Y_i$

A side condition, known as complementary slackness states :
$\alpha_i[1 - Y_ih(X_i)] = 0$ for all i
This implies either:

- $\alpha_i = 0$, which happens if the constraint $Y_ih(X_i) > 1$, That is, when the constraint is non binding.
- $\alpha_i > 0$, which happens if the constraint $Y_ih(X_i) = 1$, That is, when the constraint is binding.

2 Support vector classifier

In general, we can't realistically assume that the data are linearly separated (even in a transformed space). In, this case, the previous program has no feasible solution. We need to introduce slack variables, ξ, that allow for overlap among the classes. These slack variables allow for us to encode training misclassifications into the optimization problem.

$$max_{\beta_0,\beta,\xi_1,\ldots,\xi_n} M \ subject \ to \ Y_i(X_i) \geq M(1-\xi_i), \xi_i \geq 0, \sum \xi_i \leq t, \ for \ each \ i \tag{3}$$

Note that,

- t is a tuning parameter. This literature usually refer to t as a budget
- The separable case corresponds to t=0

We can rewrite the original program by converting $\sum \xi \leq t$ to the Lagrangian:

$$min_{\beta_0,\beta} \frac{1}{2}||\beta||_2^2 + \lambda \sum \xi \ subject \ to \ Y_i h(X_i) \geq 1-\xi, \xi \geq 0, \ for \ each \ i \tag{4}$$

The slack variables give us insight into the problem

- If $\xi = 0$, then that observation is on correct the side of the margin.
- If $\xi \in (0,1]$, then that observation is on the incorrect side of the margin, but still correctly classified.
- If $\xi > 1$, then the observation is incorrectly classified.

Continuing to convert constraints to the Lagrangian:

$$min_{\beta_0,\beta,\xi} \frac{1}{2}||\beta||_2^2 + \lambda \sum \xi - \sum_{i=1}^{n} \alpha_i [Y_i(X_i^T \beta + \beta_0) - (1-\xi)] - \sum_{i=1}^{n} \gamma_i \xi_i \tag{5}$$

Necessary conditions (taking derivatives)

- $\beta = \sum_{i=1}^{n} \alpha_i Y_i X_i$
- $0 = \sum_{i=1}^{n} \alpha_i Y_i$
- $\alpha_i = \lambda - \gamma_i$

We can think of t as a budget for the program.

- If t=0, then there is no budget and we won't tolerate any margin violations
- If $t > 0$, then no more than $\lfloor t \rfloor$ observations can be misclassified.
- A larger t then leads to larger margins.

Further Intuition:

- Like the optimal hyperplane, only observations that violate the margin determine H.
- A large t allows for many violations, hence many observations factor in to the fit.
- A small t means only a few observations do.
- t calibrates a bias/variance trade-off, as expected
- In practice, t gets selected via cross-validation.