Sequential Minimal Optimization

SMO (Platt, 1998)

Problem: Given sets $A^{\pm} = \{x_1^{\pm}, \ldots, x_{n_{\pm}}^{\pm}\}$, minimize

$$Q(\lambda^+, \lambda^-) = \| \sum_{i=1}^{n_+} \lambda_i^+ x_i^+ - \sum_{i=1}^{n_-} \lambda_i^- x_i^- \|^2 - \sum_{i=1}^{n_+} \lambda_i^+ - \sum_{i=1}^{n_-} \lambda_i^-$$

subject to the constraints $\lambda_i^{\pm} \geq 0$ for all i and

$$\sum_{i=1}^{n_+} \lambda_i^+ = \sum_{i=1}^{n_-} \lambda_i^- = \alpha > 0.$$

Strategy:

1. Pick a pair λ_i^+ and λ_j^-.

2. Holding all other $\lambda^{\pm}$ constant, change λ_i^+ and λ_j^- by *the same amount* δ so that

$$\lambda_i^+ + \delta \geq 0 \text{ and } \lambda_j^- + \delta \geq 0$$

   and $Q(\lambda^+, \lambda^-)$ decreases when you replace λ_i^+ by $\lambda_i^+ + \delta$ and λ_j^- by $\lambda_j^- + \delta$.

3. Repeat this process until the $\lambda^{\pm}$ change by an amount less than some preset tolerance.

SMO algorithm continued

Following this strategy, consider Q as a function of λ_i^+ and λ_j^-, with all other $\lambda^\pm$ treated as constants. Recall that

$$w(\lambda^+, \lambda^-) = \sum_{i=1}^{n_+} \lambda_i^+ x_i^+ - \sum_{i=1}^{n_-} \lambda^- x_i^-$$

and that

$$Q(\lambda^+, \lambda^-) = \|w(\lambda^+, \lambda^-)\|^2 - \sum_{i=1}^{n_+} \lambda^+ - \sum_{i=1}^{n_-} \lambda^-.$$

Changing $\lambda_i^+ \mapsto \lambda_i^+ + \delta$ and $\lambda_j^- \mapsto \lambda_j^- + \delta$ amounts to computing

$$w_{\delta,i,j}(\lambda^+, \lambda^-) = w(\lambda^+, \lambda^-) + \delta(x_i^+ - x_j^-).$$

To make the change that makes Q get as much smaller as possible, we want to choose δ to minimize

$$Q_{new} = \|w_{\delta,i,j}(\lambda^+, \lambda^-)\|^2 - 2\alpha$$

subject to the constraint that $\delta \geq \max\{-\lambda^+, -\lambda^i\}$.

This is a one variable minimization problem.

SMO continued.

To minimize Q_{new} we compute the derivative with respect to δ:

$$\frac{d}{d\delta}(\|w_{\delta,i,j}(\lambda^+, \lambda^-)\|^2 - 2\alpha - 2\delta) = 2w_{\delta,i,j}(\lambda^+, \lambda^-) \cdot (x_i^+ - x_j^-) - 2.$$

Setting this equal to zero yields the formula

$$\delta_{i,j} = \frac{1 - w(\lambda^+, \lambda^-) \cdot (x_i^+ - x_j^-)}{\|x_i^+ - x_j^-\|^2}.$$

To reduce the size of Q_{new} as much as possible, while preserving the constraints:

- If $\delta_{i,j} \geq \max\{-\lambda_i^+, -\lambda_j^-\}$, then replace λ_i^+ and λ_j^- by $\lambda_i^+ + \delta_{i,j}$ and $\lambda_j^- + \delta_{i,j}$ respectively.
- Otherwise let $M = \max\{-\lambda_i^+, -\lambda_j^-\}$ and replace λ_i^+ and λ_j^- by $\lambda_i^+ + M$ and $\lambda_j^- + M$ respectively – so one of the λ's will become zero.

The SMO algorithm summarized.

Algorithm (SMO)

Given: Two linearly separable sets of points $A^+ = \{x_1^+, \ldots, x_{n_+}^+\}$ and $A^- = \{x_1^-, \ldots, x_{n_-}^-\}$ in $\mathbf{R}^k$.

Find: Points p and q belonging to $C(A^+)$ and $C(A^-)$ respectively such that

$$\|p - q\|^2 = \min_{p' \in C(A^+), q' \in C(A^-)} \|p' - q'\|^2$$

Initialization: Set $\lambda_i^+ = \frac{1}{n_+}$ for $i = 1, \ldots, n_+$ and $\lambda_i^- = \frac{1}{n_-}$ for $i = 1, \ldots, n_-$. Set

$$p(\lambda^+) = \sum_{i=1}^{n_+} \lambda_i^+ x_i^+$$

and

$$q(\lambda^-) = \sum_{i=1}^{n_-} \lambda_i^- x_i^-$$

Notice that $w(\lambda^+, \lambda^-) = p(\lambda^+) - q(\lambda^-)$. Let $\alpha = \sum_{i=1}^{n_+} \lambda^+ = \sum_{i=1}^{n_-} \lambda^-$. These sums will remain equal to each other throughout the operation of the algorithm.

SMO continued

Iteration: Repeat the following steps until maximum value of δ^* computed in each iteration is smaller than some tolerance (so that the change in all of the λ's is very small):

- For each pair i, j with $1 \le i \le n_+$ and $1 \le j \le n_-$, compute

$$M_{i,j} = \max\{-\lambda_i^+, -\lambda_j^-\}$$

and

$$\delta_{i,j} = \frac{1 - (p(\lambda^+) - q(\lambda^-)) \cdot (x_i^+ - x_j^-)}{\|x_i^+ - x_j^-\|^2}.$$

If $\delta_{i,j} \ge M$ then set $\delta^* = \delta_{i,j}$; otherwise set $\delta^* = M$. Then update the $\lambda^\pm$ by the equations:

$$\begin{aligned} \lambda_i^+ &= \lambda_i^+ + \delta_{i,j}^* \\ \lambda_j^+ &= \lambda_j^- + \delta_{i,j}^* \end{aligned}$$

When this algorithm finishes, $p \approx p(\lambda^+)$ and $q \approx q(\lambda^-)$ will be very good approximations to the desired closest points.

SMO conclusion

Recall that if we set $w = p - q$, then the optimal margin classifier is

$$f(x) = w \cdot x - \frac{B^+ + B^-}{2} = 0$$

where $B^+ = w \cdot p$ and $B^- = w \cdot q$. Since $w = p - q$ we can simplify this to obtain

$$f(x) = (p - q)\dot{x} - \frac{\|p\|^2 - \|q\|^2}{2} = 0.p$$

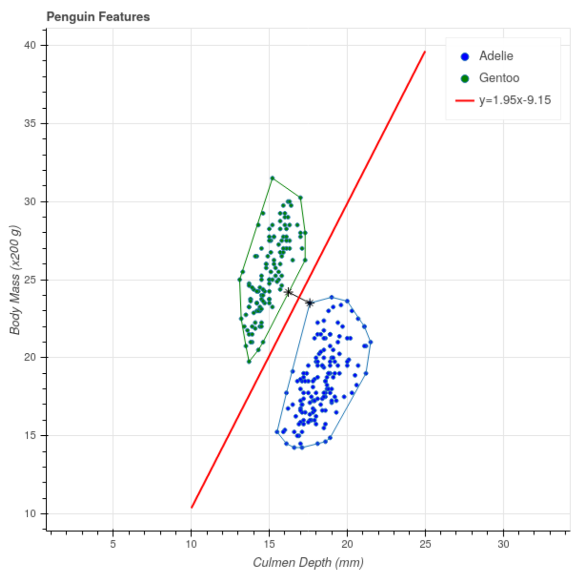

Figure 1: Closest points in convex hulls of penguin data