

Margins

Large margin principle: want \mathbf{w}, b that separates the classes maximally. We assume that \mathbf{w}, b separates the data and achieves a *functional margin* of at least 1. That is, $\mathbf{w}^\top \mathbf{x} + b \geq 1$ for all positive \mathbf{x} and $\mathbf{w}^\top \mathbf{x} + b \leq -1$ for all negative \mathbf{x} . We defined the *geometric margin* based on the *normalized* weight vector: $\gamma = \min_n y_n (\mathbf{u}^\top \mathbf{x}_n + b)$, where $\mathbf{u} = \mathbf{w} / \|\mathbf{w}\|$. Compute margin as a function of normalized weight vector \mathbf{u} , for positive point \mathbf{x}^+ and negative point \mathbf{x}^- :

$$\begin{aligned} \gamma &\leq \frac{1}{2} [\mathbf{u}^\top \mathbf{x}^+ + b - \mathbf{u}^\top \mathbf{x}^- - b] \\ \gamma &= \frac{1}{2} \left[\frac{\mathbf{w}}{\|\mathbf{w}\|}^\top \mathbf{x}^+ - \frac{\mathbf{w}}{\|\mathbf{w}\|}^\top \mathbf{x}^- \right] \\ &= \frac{1}{2\|\mathbf{w}\|} [\mathbf{w}^\top \mathbf{x}^+ - \mathbf{w}^\top \mathbf{x}^-] \\ &= \frac{1}{\|\mathbf{w}\|} \end{aligned}$$

This shows that the *margin* is inversely proportional to the norm of the weight vector, and independent of the bias. Moreover, having a large margin is equivalent to having a small weight vector norm (why does this make intuitive sense?).

Now, we write the learning problem as an optimization problem:

$$\begin{aligned} &\text{minimize}_{\mathbf{w}} \quad \frac{1}{2} \|\mathbf{w}\|^2 \\ &\text{subject to} \quad y_n [\mathbf{w}^\top \mathbf{x}_n + b] - 1 \geq 0 \quad (\forall n) \end{aligned}$$

Why large margins? Because they mean simple solutions.