

Maximum margin classifiers

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:

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

Now we need to figure out how to solve this. Enter convex optimization and Lagrange theory...

Introduce Lagrange-multipliers $\alpha_{1:N}$, one for each constraint. Leads to Lagrangian:

$$L(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_n \alpha_n \{y_n [\mathbf{w}^\top \mathbf{x}_n + b] - 1\}$$

Now, we want to minimize $L(\mathbf{w}, \boldsymbol{\alpha})$ with respect to *both* \mathbf{w} and α . Take derivatives with respect to \mathbf{w} :

$$\begin{aligned} \nabla_{\mathbf{w}} L &= \mathbf{w} - \sum_n \alpha_n y_n \mathbf{x}_n = 0 \\ \implies \mathbf{w} &= \sum_n \alpha_n y_n \mathbf{x}_n \\ \nabla_b L &= \sum_n y_n \alpha_n = 0 \end{aligned}$$

So, given $\boldsymbol{\alpha}$, \mathbf{w} is deterministic... plug back in to L :

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \frac{1}{2} \left\| \sum_n \alpha_n y_n \mathbf{x}_n \right\|^2 - \sum_n \alpha_n \left\{ y_n \left[\left(\sum_m \alpha_m y_m \mathbf{x}_m \right)^\top \mathbf{x}_n + b \right] - 1 \right\} \\ &= \frac{1}{2} \sum_m \sum_n \alpha_m \alpha_n y_m y_n \mathbf{x}_m^\top \mathbf{x}_n - \sum_m \sum_n \alpha_m \alpha_n y_m y_n \mathbf{x}_m^\top \mathbf{x}_n - b \sum_n \alpha_n y_n + \sum_n \alpha_n \\ &= -\frac{1}{2} \sum_m \sum_n \alpha_m \alpha_n y_m y_n \mathbf{x}_m^\top \mathbf{x}_n + \sum_n \alpha_n \end{aligned}$$

So now just solve:

$$\begin{aligned} \text{minimize}_{\boldsymbol{\alpha}} \quad & \sum_n \alpha_n - \frac{1}{2} \sum_{n,m} y_n y_m \alpha_n \alpha_m \mathbf{x}_n^\top \mathbf{x}_m \\ \text{subject to} \quad & \sum_n y_n \alpha_n = 0 \\ & \alpha_n \geq 0 \quad , \quad (\forall n) \end{aligned}$$

Then compute the bias:

$$b = -\frac{1}{2} \left[\max_{n:y_n=-1} \mathbf{w}^\top \mathbf{x}_n + \min_{n:y_n=+1} \mathbf{w}^\top \mathbf{x}_n \right]$$

This leads to a *sparse* solution: most $\boldsymbol{\alpha}$ are zero. Why? The Karush-Kuhn-Tucker conditions say that at the optimum:

$$\alpha_n [y_n (\mathbf{w}^\top \mathbf{x}_n + b) - 1] = 0 \quad (\forall n)$$

This means that $\alpha_n \neq 0$ only when the point is right on the margin. These points are the *support vectors*.

Not linearly-separable data...

Introduce slack parameters: ξ_n is how far “on the wrong side” $y_n \mathbf{x}_n$ is from the margin. Then:

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

High C means “fit data” while low C means “have a large margin.”

Following the same dual formulation, we get:

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, r) = \frac{1}{2} \mathbf{w}^\top \mathbf{w} + C \sum_n \xi_n - \sum_n \alpha_n [y_n (\mathbf{w}^\top \mathbf{x}_n + b) - 1 + \xi_n] - \sum_n r_i \xi_i$$

Differentiate:

$$\begin{aligned} \nabla_{\mathbf{w}} L &= \mathbf{w} - \sum_n y_n \alpha_n \mathbf{x}_n = 0 \\ \implies \mathbf{w} &= \sum_n y_n \alpha_n \mathbf{x}_n \\ \nabla_b L &= \sum_n y_n \alpha_n = 0 \\ \nabla_{\xi_n} L &= C - \alpha_n - r_n = 0 \end{aligned}$$

Thus:

$$L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, r) = \sum_n -\frac{1}{2} \sum_{n,m} y_n y_m \alpha_n \alpha_m \mathbf{x}_n^\top \mathbf{x}_m$$

Which is the same as before, but now we have constraints: $\alpha_n \geq 0$ and $r_n \geq 0$ and $C - \alpha_n - r_n = 0$. This means: (1) $\alpha_n \leq C$. (2) if $r_n = 0$ then $\alpha_n = C$. Geometrically, support vectors are now also the “noisy” points.

Why large margins? Because they mean simple solutions.