Support Vector Machines

Ali Gooya

Introduction

- We assume linearly separable data for binary classification.
- Training data $\mathrm{x}_1, \cdots, \mathrm{x}_N$, with target variables $\mathrm{t}_1, \cdots, \mathrm{t}_N$, $\mathrm{t}_i \in \{-1,1\}$
- We view the sign of for classification: $h(\mathrm{x}) = \mathrm{sign}(y(\mathrm{x}))$
- Note that we can scale w and b without changing $h(\mathrm{x})$
- Choose w and b so that for the closest point to the decision boundary we satisfy: $y(\mathrm{x}) = \pm 1$
- Therefore we assume:
  - For any nearest point x_1 with $t_1 = -1$, we have $\mathrm{w}^T\mathrm{x}_1 + b = -1$
  - For any nearest point x_2 with $t_2 = 1$, we have $\mathrm{w}^T\mathrm{x}_2 + b = -1$

Introduction

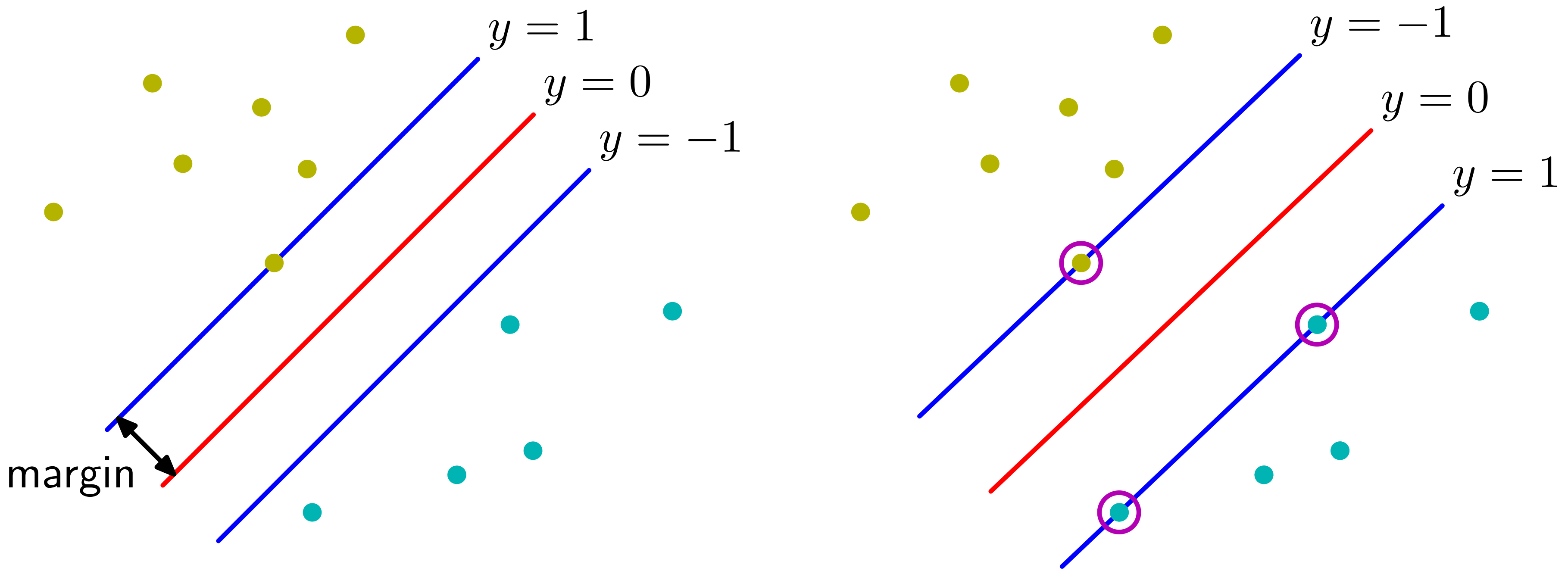
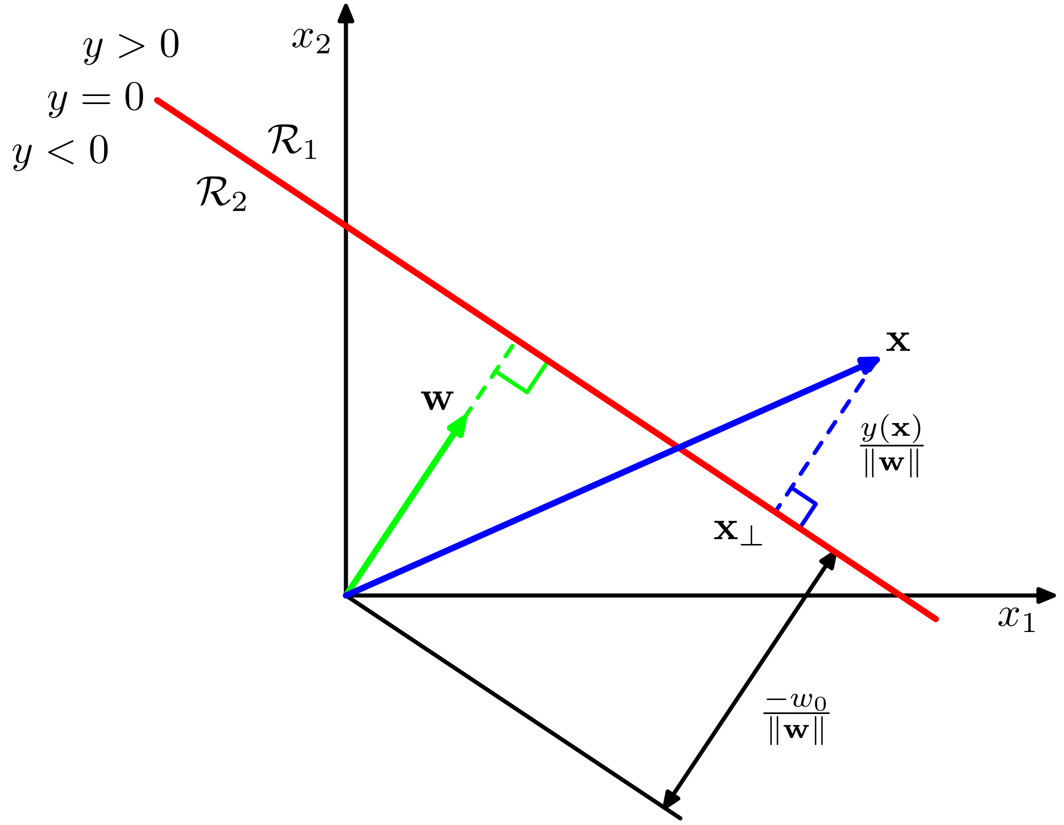


Figure 7.1 from Bishop: The margin is defined as the perpendicular distance between the decision boundary and the closest of the data points, as shown on the left figure. Maximizing the margin leads to a particular choice of decision boundary, as shown on the right. The location of this boundary is determined by a subset of the data points, known as support vectors, which are indicated by the circles.

Review: Geometry of the linear classification

- For a linear predictor function $y(\mathrm{x}) = \mathrm{w}^T\mathrm{x} + b$, recall that:
- w is normal to the decision boundary.
- Distance of x from decision boundary $y = 0$ is proportional to $y(\mathrm{x})$



SVM Basics

- So, the distance of the data points from $y = 0$ is: $\dfrac{t_n y_n}{\|\mathrm{w}\|}$
- Thus, for the closet points, this distance (margin) is: $\dfrac{1}{\|\mathrm{w}\|}$
- To maximise the margin is to minimise $f(\mathrm{w}) = \dfrac{1}{2}\|\mathrm{w}\|^2$
- Subject to constraints that: $1 \leq t_n y_n$

Primal problem

- For the linear separable SVM, the constrained optimisation is specified

$$\min_{\mathbf{w},b} \quad f(\mathbf{w}) = \tfrac{1}{2}\mathbf{w}^\top \mathbf{w} \tag{1}$$

$$\text{subject to} \quad t_i\left(\mathbf{w}^\top \mathbf{x}_i + b\right) \geq 1, \quad i = 1, \ldots, n \tag{2}$$

- Applying the Karush-Kuhn-Tucker conditions, the primal problem is:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^\top \mathbf{w} + \sum_i^n \alpha_i (1 - t_i(\mathbf{w}^\top \mathbf{x}_i + b))$$

$$\alpha_i \geq 0 \qquad \boldsymbol{\alpha} = (\alpha_1 \cdots, \alpha_n)^\top$$

$$1 - t_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 0$$

$$\alpha_i(1 - t_i(\mathbf{w}^\top \mathbf{x}_i + b)) = 0$$

Slackness conditions

Primal problem

- The primal problem is a convex optimisation problem that can be solved by solvers such as ALGLIB.
- Setting the derivatives of the Lagrangian with respect to the parameters:

$$
\begin{aligned}
\frac{\partial L}{\partial b} &= 0 \Rightarrow \sum_{i=1}^{n} \alpha_i t_i = 0 \\
\frac{\partial L}{\partial \mathbf{w}} &= 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i t_i \mathbf{x}_i
\end{aligned}
$$

- We solve an example problem to further illustrate.

From primal to dual problems

- The primal problem is a convex optimization problem that can be solved by solvers such as ALGLIB.
- But our SVM will remain linear ☹
- Moving towards dual problem can be rewarding and can make SVM non-linear ☺ (-- using kernel trick!)
- To define the dual problem, we first eliminate the primal variables.

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i t_i = 0$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{n} \alpha_i t_i \mathbf{x}_i$$

Dual problem

- Substitute the results in the primal:

$$
\begin{aligned}
L(\mathbf{w}, b; \alpha) &= \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j t_i t_j \mathbf{x}_i^\top \mathbf{x}_j \;+\; \sum_{i=1}^{n} \alpha_i \left(1 - t_i \left(\sum_{j=1}^{n} \alpha_j t_j \mathbf{x}_j^\top \mathbf{x}_i + b\right)\right) \\
&= \sum_{i=1}^{n} \alpha_i \;-\; \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j t_i t_j \mathbf{x}_i^\top \mathbf{x}_j
\end{aligned}
$$

Dual problem

- We arrive at the following dual problem

$$g(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} t_i t_j \alpha_i \alpha_j \mathbf{x}_i^\top \mathbf{x}_j$$

$$\alpha_i \geq 0$$

$$\sum_{i=1}^{n} \alpha_i t_i = 0$$

Complementary slackness condition

- Note that the dual problem is concave (why?) and should be maximized (see the the previous lecture)
- This can be solved by quadratic programming.

Solution for the dual

- Let $\boldsymbol{\alpha}^* = (\alpha_1^*, \cdots, \alpha_n^*)^\top$ be the solution for above then

$$\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* t_i \mathbf{x}_i$$

- To determine b^*, we note that for any $\alpha_i^* > 0$

$$t_i({\mathbf{w}^*}^\top \mathbf{x}_i + b^*) = 1$$

Sparsity in predication

- For any test point such as $\mathbf{x}$ the discriminant will be

$$y(\mathbf{x}) = \mathbf{w}^{*\top}\mathbf{x} + b^*$$
$$= \sum_{i:\alpha_i^*>0} \alpha_i^* t_i \mathbf{x}^\top \mathbf{x}_i + b^*$$

- This shows that prediction uses only a sparse number of $\mathbf{x}_i$'s.
- One interesting point is that only inner products ($\mathbf{x}_i^\top \mathbf{x}_j$) appear in the dual problem and the prediction!
- We can exploit this observation to make SVMs non-linear.