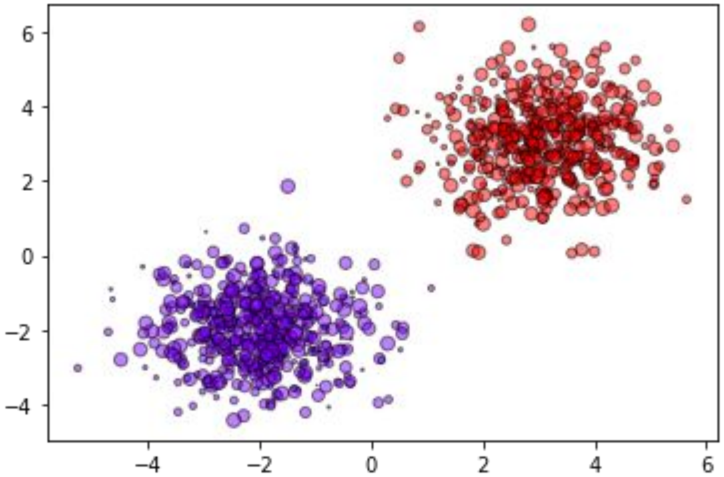
Support Vector Machines

Jayson Cunanan

How do we separate the samples in the figure?

- Draw a line, curves?
- How do you define a line?
  - Can that definition be generalized to higher Dimension?
- What is the best line that separates the current samples?



Support Vector Machine is a **supervised** learning algorithm used in **classification** and **regression** tasks.

Defining the line/hyperplane $\{\boldsymbol{x} \in \mathbb{R}^D : f(\boldsymbol{x}) = 0\}$

$$f : \mathbb{R}^D \to \mathbb{R}$$
$$\boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$$

normal vector,
weight

Intercept,
bias

- Train classifier

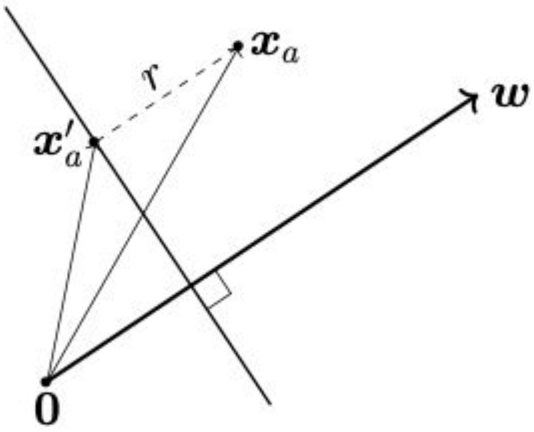
$$\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b \geqslant 0 \quad \text{when} \quad y_n = +1$$

$$\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b < 0 \quad \text{when} \quad y_n = -1$$

The concept of margin / Primal SVM

- **Margin** is the distance of the separating hyperplane to the closest examples in the dataset, assuming that the dataset is linearly separable.

$$
\begin{aligned}
&\max_{\boldsymbol{w},b,r} \quad \underbrace{r}_{\text{margin}} \\
&\text{subject to} \quad \underbrace{y_n(\langle \boldsymbol{w}, \boldsymbol{x}_n \rangle + b) \geqslant r}_{\text{data fitting}}, \; \underbrace{\|\boldsymbol{w}\| = 1}_{\text{normalization}}, \quad r > 0
\end{aligned}
$$



Cons: The number of parameters (the dimension of **w**) of the optimization problem grows linearly with the number of features.

Hiding away duality, Lagrange multipliers,...

We arrive to the **Dual SVM**

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle - \sum_{i=1}^{N} \alpha_i$$

$$\text{subject to} \quad \sum_{i=1}^{N} y_i \alpha_i = 0$$

$$0 \leqslant \alpha_i \leqslant C \quad \text{for all} \quad i = 1, \ldots, N\,.$$

Key observation: **Inner product** occurs only between samples!

This formulation allows us to deal with **non-linear** problems via the **kernel trick**!

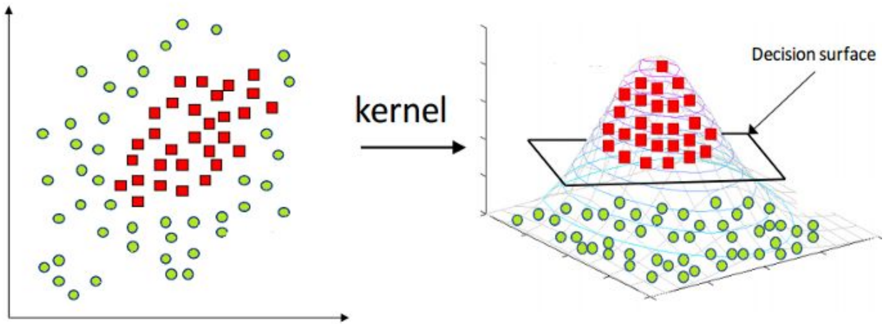
Kernels

Let Φ be a **Feature map**.

Use $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$ instead of $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ to compute **similarities** back in Dual SVM formulation.

Hiding away Reproducing Kernel Hilbert Space…

- We can find a **kernel** k such that $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$
- deal with a dataset that is not linearly separable

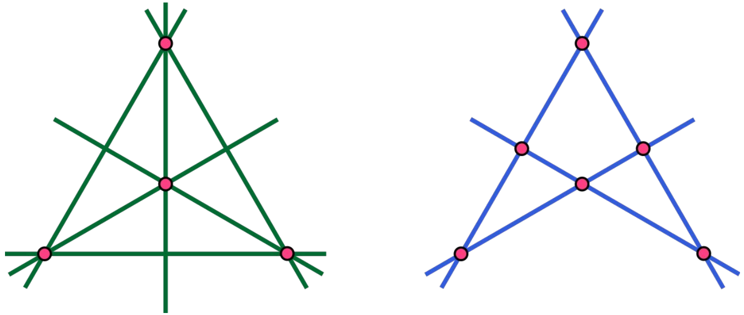


source

THANKS FOR COMING!

Duality

Duality translates concepts, theorems or mathematical structures into other concepts, theorems or structures, in a one-to-one fashion.



Left: four points in a plane, no three of which are on a common line, and of the six lines connecting the six pairs of points

Right: four lines, no three of which pass through the same point, and the six points of intersection of these lines