# APMA1210 Recitation 10
# Introduction to Supporting Vector Machine (SVM)

## 1 Setup

Here are 6 points on the 2D plane (shown in Figure 1), and they are divided into two groups.

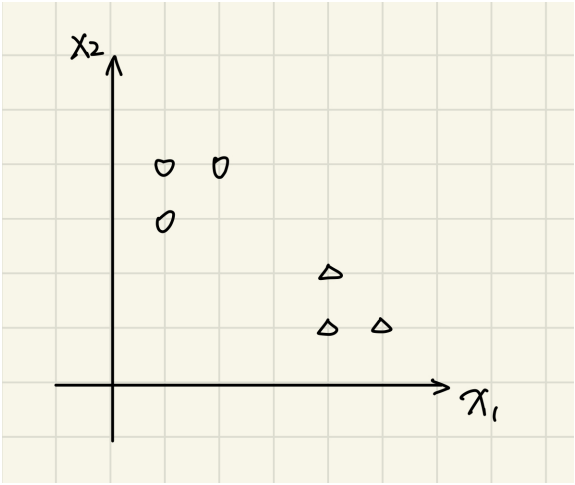


Figure 1: Dataset for classification. Here the points denoted by circles are $(1, 4)$, $(1, 3)$ and $(2, 4)$, which are labeled with $-1$. The points denoted by the triangles are $(4, 2)$, $(4, 1)$ and $(5, 1)$, which are labeled with 1.

We want to find a linear classifier, i.e. a linear function $f : \mathbb{R}^2 \to \mathbb{R}$, s.t. $f(x_1, x_2) > 0$ for $(x_1, x_2)$ in one group and $f(x_1, x_2) < 0$ for $(x_1, x_2)$ in the other group. Let $f = w_1x_1 + w_2x_2 + b$, then we want to find $w_1$, $w_2$ and $b$.

## 2 Maximizing margin distance

We also want $f$ to be "robust", that is, the "distance" between the two groups should be as large as possible. This can be considered as maximizing the distance between two lines

$$\mathbf{w}^T\mathbf{x} + b = -1 \quad \text{anything on or above this boundary is of one class, with label } -1$$

and

$$\mathbf{w}^T\mathbf{x} + b = 1 \quad \text{anything on or below this boundary is of one class, with label } 1$$

where $\mathbf{w} = (w_1, w_2)$, $\mathbf{x} = (x_1, x_2)$. Two possible lines are shown in Figure 2.

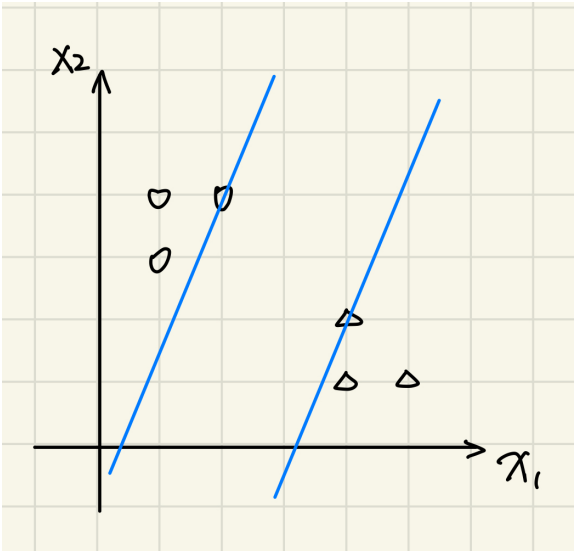


Figure 2: The blue lines are possible margins for the two groups.

The distance between the two lines is

$$d = \frac{2}{\sqrt{w_1^2 + w_2^2}} = \frac{2}{\|\mathbf{w}\|}$$

So this leads to a constrained minimization problem

$$\begin{aligned} \min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\| \\ \text{s.t.} \quad & y_i f_{\mathbf{w}}(\mathbf{x}_i) \geq 1, \quad i = 1, \ldots, n \end{aligned}$$

## 3 Solving with Lagrange multiplier

This constrained minimization problem can be solved with the Lagrange multiplier method. First, we need to determine which constraints are active. From Figure 1, we know that points $(1, 3)$, $(2, 4)$, and $(4, 2)$ are on the margins, and on these points, the inequality constraints become equality constraints. So the objective function now becomes

$$J(\mathbf{w}) = \frac{1}{2}(w_1^2 + w_2^2) + \lambda_1(w_1 + 3w_2 + b + 1) + \lambda_2(2w_1 + 4w_2 + b + 1) + \lambda_3(4w_1 + 2w_2 + b - 1)$$

So we just need to compute the Jacobian of $J$ and solve the following linear system

$$\begin{aligned} \frac{\partial J}{\partial w_1} &= w_1 + \lambda_1 + 2\lambda_2 + 4\lambda_3 = 0 \\ \frac{\partial J}{\partial w_2} &= w_2 + 3\lambda_1 + 4\lambda_2 + 2\lambda_3 = 0 \\ \frac{\partial J}{\partial b} &= b + \lambda_1 + \lambda_2 + \lambda_3 = 0 \\ \frac{\partial J}{\partial \lambda_1} &= w_1 + 3w_2 + b = 0 \\ \frac{\partial J}{\partial \lambda_2} &= 2w_1 + 4w_2 + b = 0 \\ \frac{\partial J}{\partial \lambda_3} &= 4w_1 + 2w_2 + b = 0 \end{aligned}$$

And thus we have

$$w_1 = \frac{1}{2} \quad w_2 = -\frac{1}{2} \quad b = 0$$
$$\lambda_1 = 0 \quad \lambda_2 = \frac{1}{2} \quad \lambda_3 = -\frac{1}{2}$$

We see that the coefficient of constraint corresponding to $(1, 3)$ is 0, which means it is not active. So the supporting vector in this problem is actually $(2, 4)$ and $(4, 2)$, which agrees with our intuition. The margins with maximized distance are shown in

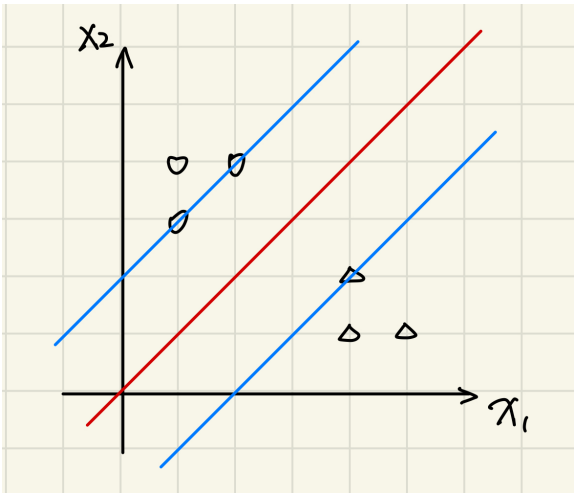


Figure 3: The blue lines are margins with maximized distance. The red line is the classifier $f(x_1, x_2) = \frac{1}{2}x_1 - \frac{1}{2}x_2$.

## 4 Classifying dataset which is not linearly separable

In some cases, the dataset cannot be split into two groups by a line (hyperplane, in general). Here is an example.

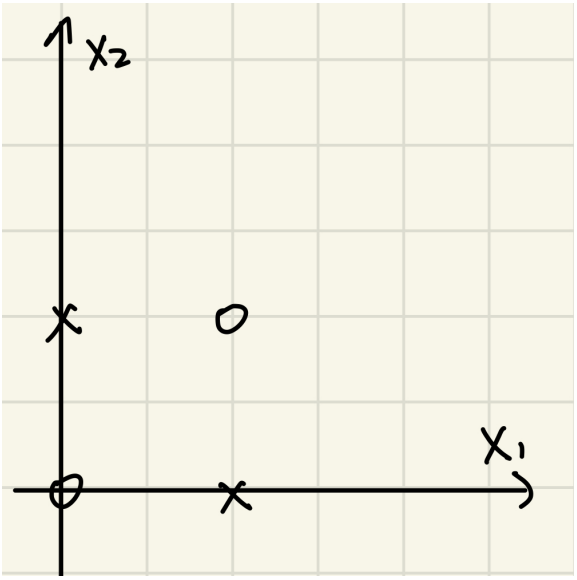


Figure 4: The points denoted with o are in one group, and the points denoted with x are in the other. Assume the coordinates of the o points are $(0, 0)$ and $(1, 1)$, and the coordinates of the x points are $(1, 0)$ and $(0, 1)$.

If we try to use the Lagrange multiplier method, then we will need to minimize

$$J(\mathbf{w}) = \frac{1}{2}(w_1^2 + w_2^2) + \lambda_1(b-1) + \lambda_2(w_1 + w_2 + b - 1) + \lambda_3(w_1 + b + 1) + \lambda_4(w_2 + b + 1)$$

and we need to solve solve the following linear system

$$\begin{aligned}
\frac{\partial J}{\partial w_1} &= w_1 + \lambda_2 + \lambda_3 = 0 \\
\frac{\partial J}{\partial w_2} &= w_2 + \lambda_2 + \lambda_4 = 0 \\
\frac{\partial J}{\partial b} &= b + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 0 \\
\frac{\partial J}{\partial \lambda_1} &= b - 1 = 0 \\
\frac{\partial J}{\partial \lambda_2} &= w_1 + w_2 + b - 1 = 0 \\
\frac{\partial J}{\partial \lambda_3} &= w_1 + b + 1 = 0 \\
\frac{\partial J}{\partial \lambda_3} &= w_2 + b + 1 = 0
\end{aligned}$$

This linear system is singular, which validates our observation.

There are two approaches to dealing with problems like this: soft margin and kernel trick. These are advanced topics, and we will only give some basic ideas. Details can be found in any modern textbook about machine learning.

## 4.1 Soft margin

The idea of the soft margin method is that we can allow some "mistakes" with cost. The formulation is

$$\begin{aligned}
\min_{\mathbf{w},b} \quad & \frac{1}{2}\|\mathbf{w}\| + C\sum_{i=1}^{N}\zeta_i \\
\text{s.t.} \quad & y_i f_{\mathbf{w}}(\mathbf{x}_i) \geq 1 - \zeta_i, \quad i = 1, \ldots, N
\end{aligned}$$

where $C$ is a parameter we can tune. The $\zeta_i$ here are the cost of not classifying all points correctly. By minimizing them we apply the "soft" constraints, which is a generalization of the constraints. Because if all the $\zeta_i$ are 0, then the problem degenerates to the hard margin problem we have just discussed in the preceding section.

## 4.2 Kernel trick

Roughly speaking, we can find nonlinear transformation $\phi$, so for $\mathbf{x}_i$, $\phi(\mathbf{x}_i)$ will be linearly separable, then we just solve the linear SVM for the transformed data.