Department of Mathematics, University of Udine

SVM and their use in IR

Lorenzo Zanolin.
lorenzo.zanolin@spes.uniud.it

May x, 2023

Outline

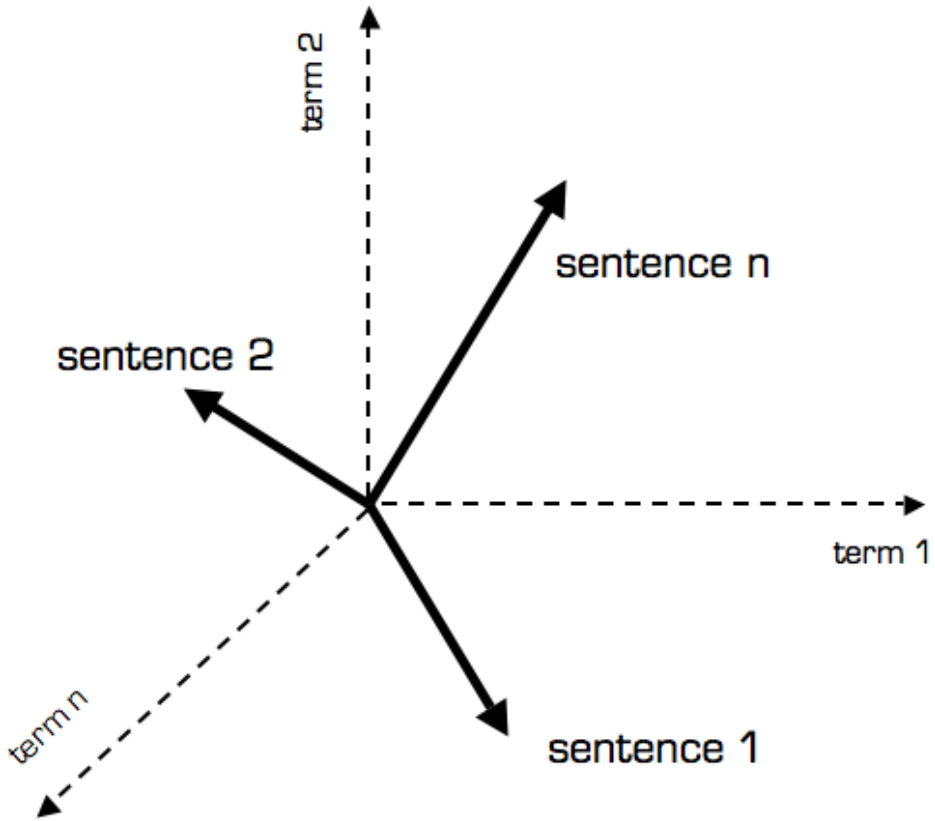
1. Linearly separable data
2. Non linearly separable data

Vector Space

Suppose you have to classify whether a document is *relevant* or not. We can think to use *terms* as features to divide properly the data. We will use a *t-dimensional* vector space to represent our documents.

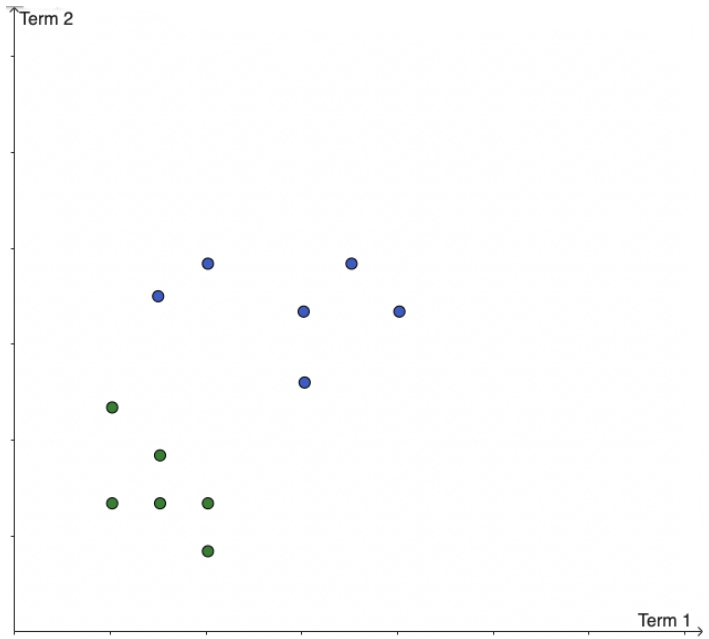
Vector Space

Suppose you have to classify whether a document is *relevant* or not. We can think to use *terms* as features to divide properly the data. We will use a *t-dimensional* vector space to represent our documents.



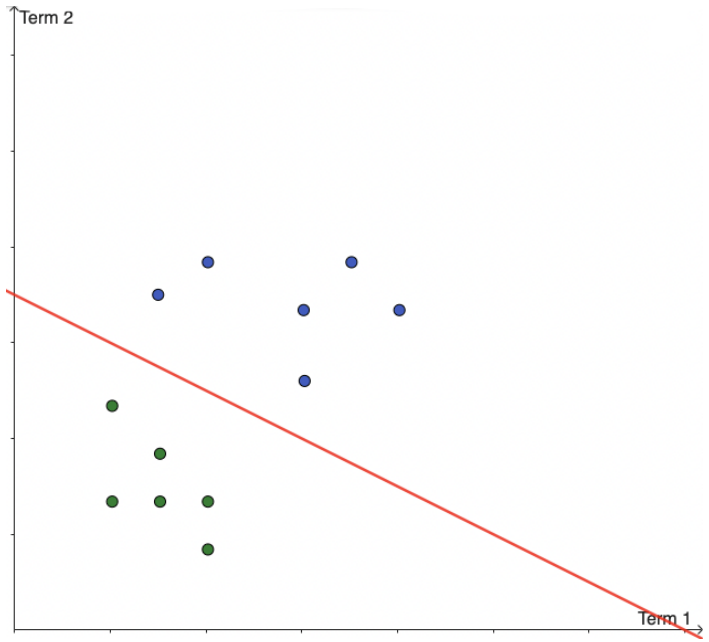
Classification

For simplicity, we can reason using a 2D Space.
Data can be separated using a *decision boundary*, which is an *hyperplane*.



Classification

For simplicity, we can reason using a 2D Space.
Data can be separated using a *decision boundary*, which is an *hyperplane*.



Support Vectors

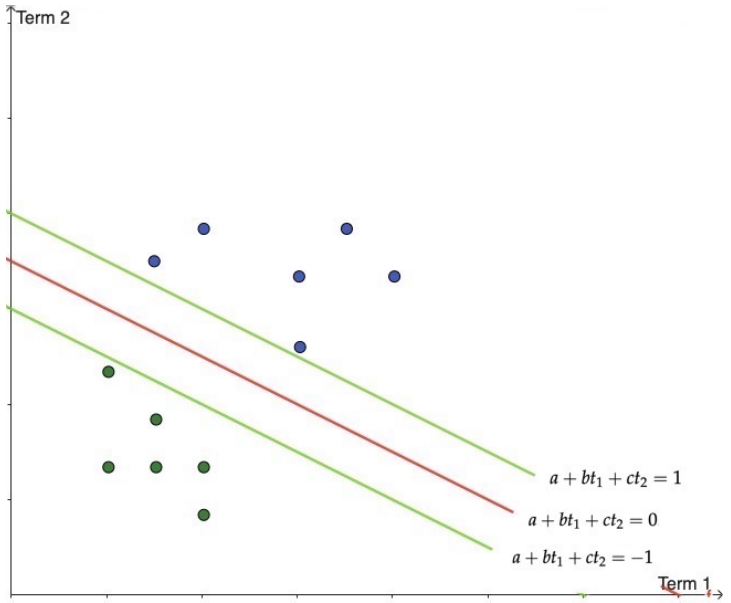
In the previous example the *decision boundary* is a line, represented by the equation $a + bt_1 + ct_2 = 0$.
We can introduce two parallels hyperplanes (lines) to the decision boundary, called *support vectors* whose equations are

$$\begin{cases} a + bt_1 + ct_2 = 1 \\ a + bt_1 + ct_2 = -1 \end{cases} \tag{1}$$

Support Vectors

In the previous example the *decision boundary* is a line, represented by the equation $a + bt_1 + ct_2 = 0$.
We can introduce two parallels hyperplanes (lines) to the decision boundary, called *support vectors* whose equations are

$$\begin{cases} a + bt_1 + ct_2 = 1 \\ a + bt_1 + ct_2 = -1 \end{cases} \tag{1}$$



Margin

For convenience, we can rewrite equations using *vectorization* notation.

$$\begin{cases} b^T t + a = 0 \\ b^T t + a = \pm 1 \end{cases} \tag{2}$$

Margin

For convenience, we can rewrite equations using *vectorization* notation.

$$\begin{cases} b^T t + a = 0 \\ b^T t + a = \pm 1 \end{cases} \tag{2}$$

Intuitively, we can define the *margin* of the two support vectors as the distance between them.

Margin

For convenience, we can rewrite equations using *vectorization* notation.

$$\begin{cases} b^T t + a = 0 \\ b^T t + a = \pm 1 \end{cases} \tag{2}$$

Intuitively, we can define the *margin* of the two support vectors as the distance between them.

We can consider two points x_1, x_2 that lie respectively on the two support vectors, their distance is $\lambda \|b\| = \frac{2}{\sqrt{b^T b}}$.

SVM are used to find the *maximum margin linear classifier*, thus we want to maximize the margin.

Cost Function

Remembering we want to classify documents, our goal is to find specific b s.t. given a document x belonging to class y the decision boundary behave the following:

$$\begin{cases} b^T x + a \geq 1 & if\ y = 1 \\ b^T x + a \leq -1 & if\ y = -1 \end{cases} \tag{3}$$

Cost Function

Remembering we want to classify documents, our goal is to find specific b s.t. given a document x belonging to class y the decision boundary behave the following:

$$\begin{cases} b^T x + a \geq 1 & if\ y = 1 \\ b^T x + a \leq -1 & if\ y = -1 \end{cases} \tag{3}$$

We can define the *cost function* as a system of equations.

$$\begin{cases} \min_{b,a} \frac{\sqrt{b^T b}}{2} \\ subject\ to \quad y_i(b^T x_i + a) \geq 1 \quad \forall x_i \end{cases} \tag{4}$$

Soft-Margin

In real world problem it is not likely to get an exactly separate line dividing the data within the space. It would be better for the smooth boundary to ignore few data points than be curved or go in loops, around the outliers.

Soft-Margin

In real world problem it is not likely to get an exactly separate line dividing the data within the space. It would be better for the smooth boundary to ignore few data points than be curved or go in loops, around the outliers.

So, we will use *slack variables* to introduce a penalty for each misclassified point.

Soft-Margin

In real world problem it is not likely to get an exactly separate line dividing the data within the space. It would be better for the smooth boundary to ignore few data points than be curved or go in loops, around the outliers.

So, we will use *slack variables* to introduce a penalty for each misclassified point.

The new cost function will be

$$\begin{cases} \min_{b,a} \frac{\sqrt{b^T b}}{2} + C \sum_i \xi_i \\ \textit{subject to} \quad y_i(b^T x_i + a) \geq 1 - \xi_i \quad \textit{and } \xi_i > 0 \quad \forall x_i \end{cases} \tag{5}$$

Visualization

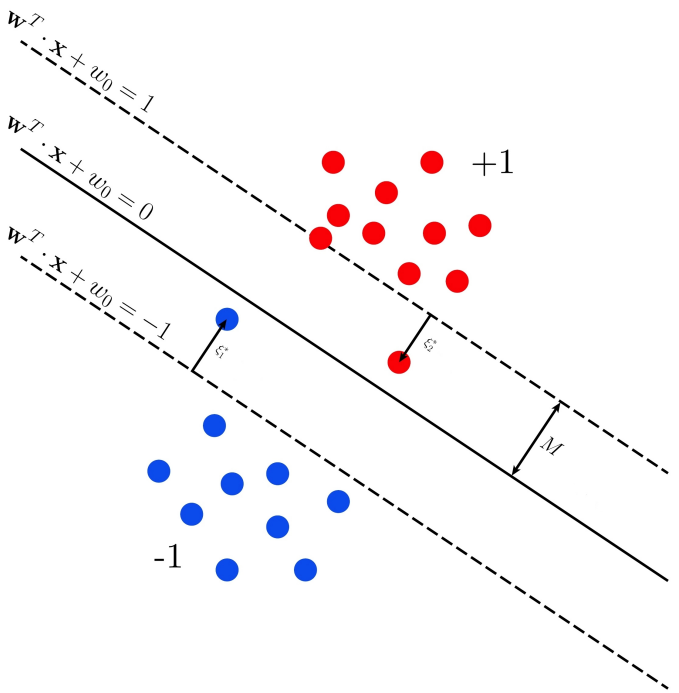
The larger is C the stricter the classification is, since a larger C will give more evidence to slack variables.

$$\begin{cases} \min_{b,a} \frac{\sqrt{b^T b}}{2} + C \sum_i \xi_i \\ \textit{subject to} \quad y_i(b^T x_i + a) \geq 1 - \xi_i \quad \textit{and } \xi_i > 0 \quad \forall x_i \end{cases} \tag{6}$$

Visualization

The larger is C the stricter the classification is, since a larger C will give more evidence to slack variables.

$$\begin{cases} \min_{b,a} \frac{\sqrt{b^T b}}{2} + C \sum_i \xi_i \\ \textit{subject to} \quad y_i(b^T x_i + a) \geq 1 - \xi_i \quad \textit{and } \xi_i > 0 \quad \forall x_i \end{cases} \tag{6}$$

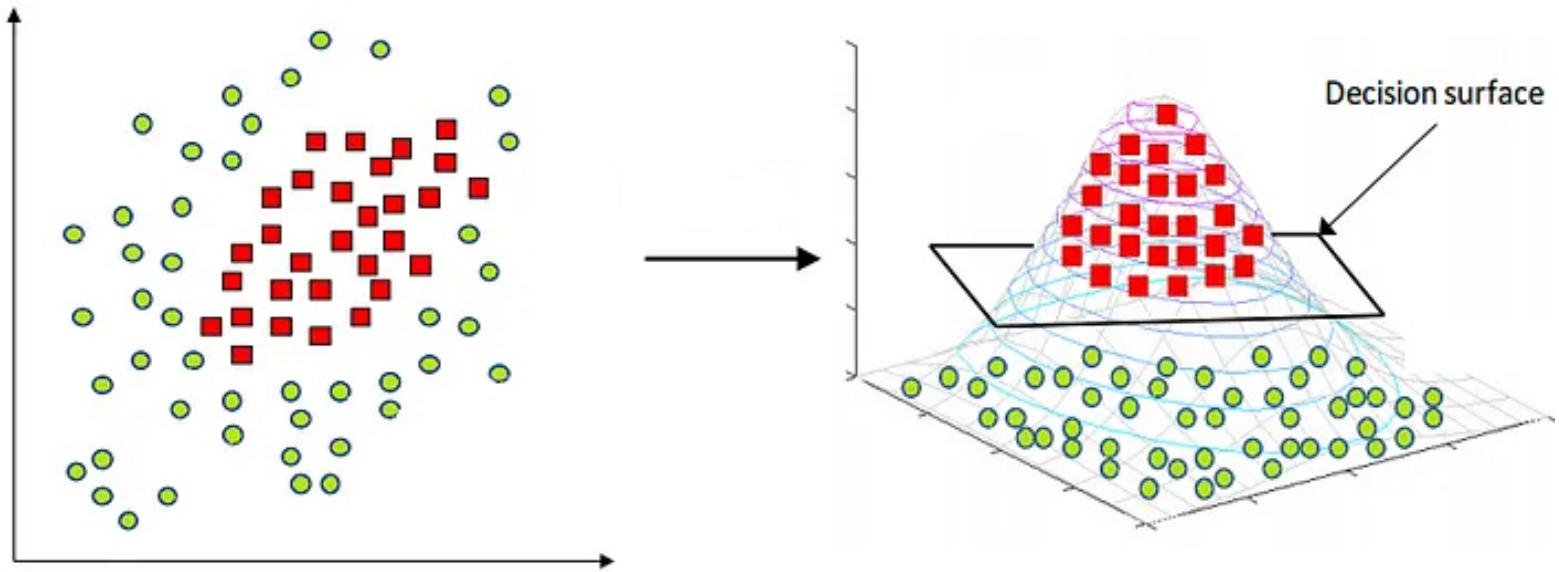


Non linearly separable data

What if our data is not linearly separable?

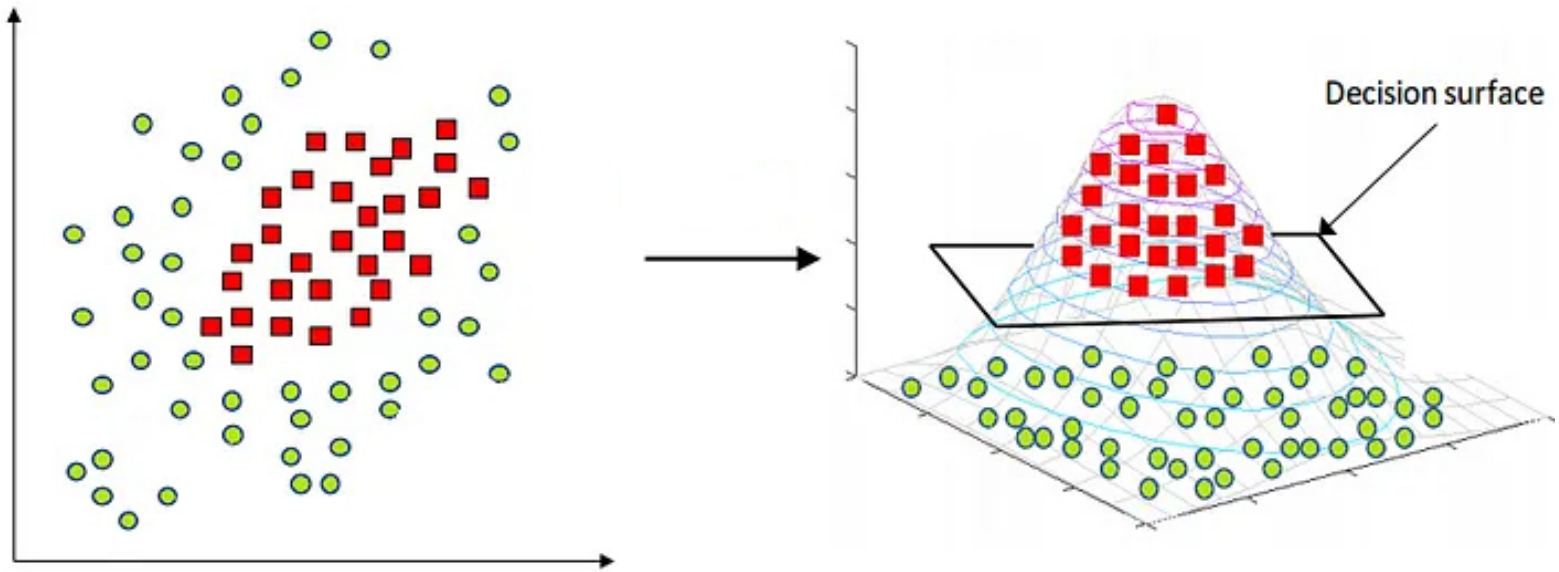
Non linearly separable data

What if our data is not linearly separable? It should be better to reason in a bigger dimensional space.



Non linearly separable data

What if our data is not linearly separable? It should be better to reason in a bigger dimensional space.



But augmenting dimensions costs a lot…

Kernel Trick

Here *Kernel Trick* comes handy.

Kernel Trick

Here *Kernel Trick* comes handy.

Consider the function $\phi : \mathbb{R}^{\not\vdash} \mapsto \mathbb{R}^{\not\vdash\not\vdash}$

Kernel Trick

Here *Kernel Trick* comes handy.

Consider the function $\phi : \mathbb{R}^{\not\vdash} \mapsto \mathbb{R}^{\not\vdash\not\vdash}$

It allows us to operate in the original feature space without computing the coordinates of the data in a higher dimensional space.