Support vector machines

Dominik Wisniewski
Wojciech Wawrzyniak

Outline

1. A brief history of SVM.
2. What is SVM and how does it work?
3. How would you classify this data?
4. Are all the separating lines equally good?
5. An example of small and large margins.
6. Transforming the Data.
7. Learning.
8. Support vectors.
9. Kernel functions.
10. Predicting the classification.
11. References.

A brief history of SVM.

- SVMs were introduced by Boser, Guyon, Vapnik in 1992.
- SVMs have become popular because of their success in handwritten digit recognition.
- SVMs are now important and active field of all Machine Learning research and are regarded as an main example of “kernel methods”.

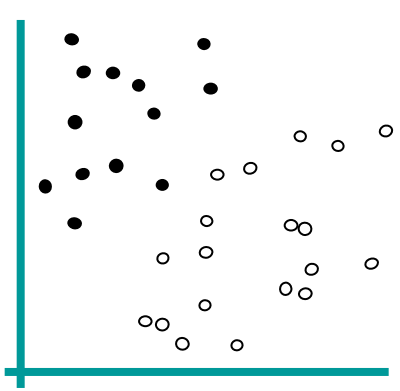
∈

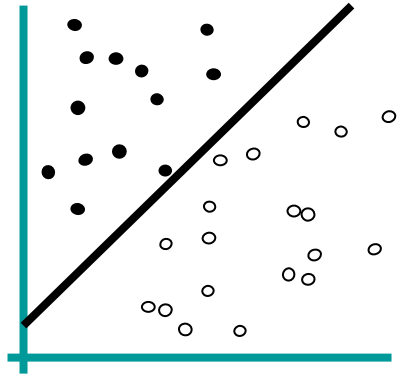
What is SVM and how does it work.

- Family of machine-learning algorithms that are used for mathematical and engineering problems including for example handwriting digit recognition, object recognition, speaker identification, face detections in images and target detection.
- Task: Assume we are given a set S of points $x_i \in R^n$ with i = 1,2,..., N. Each point x_i belongs to either of two classes and thus is given a label $y_i \in \{-1,1\}$. The goal is to establish the equation of a hyperplane that divides S leaving all the points of the same class on the same side.
- SVM performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories.

How would you classify this data?

- Let's consider the objects on illustration on the left. We can see that the objects belong to two different classes. The separating line (2 – dimentional hyperplane) on the second picture is a decision plane which divides the objects into two subsets such that in each subset all elements are similar.



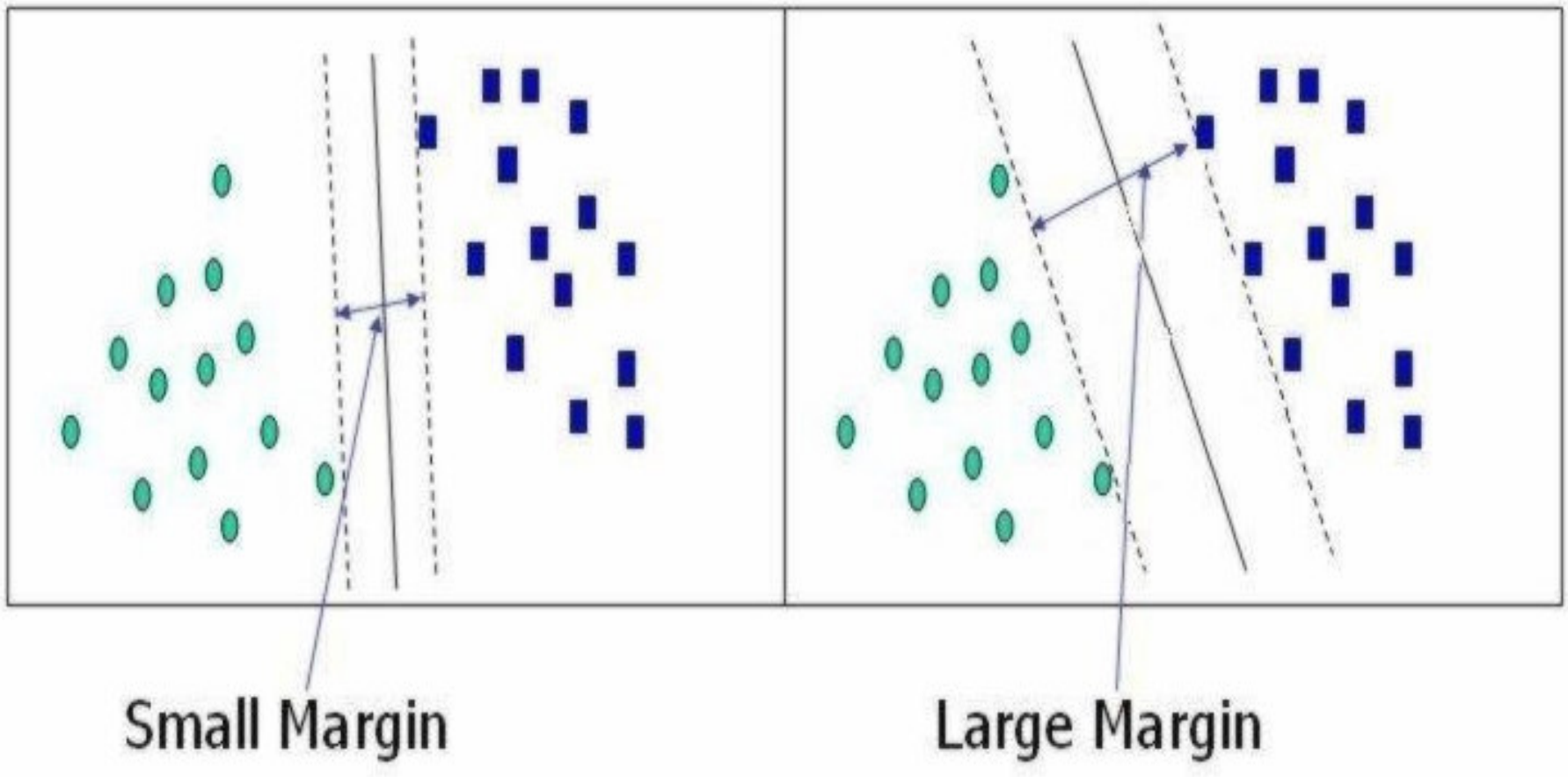


Note: There are a lot of possible separating lines for a given set of objects. Are all the separating lines (decision boundaries = decision planes) equally good?

Are all the separating lines equally good?

- Among the possible hyperplanes, we select the one where the distance of the hyperplane from the closest data points (the “margin”) is as large as possible. An intuitive justification for this criterion is: suppose the training data are good, in the sense that every possible test vector is within some radius r of a training vector. Then, if the chosen hyperplane is at least r from any training vector it will correctly separate all the test data. By making the hyperplane as far as possible from any data, r is allowed to be correspondingly large. The desired hyperplane (that maximizes the margin) is also the bisector of the line between the closest points on the convex hulls of the two data sets.

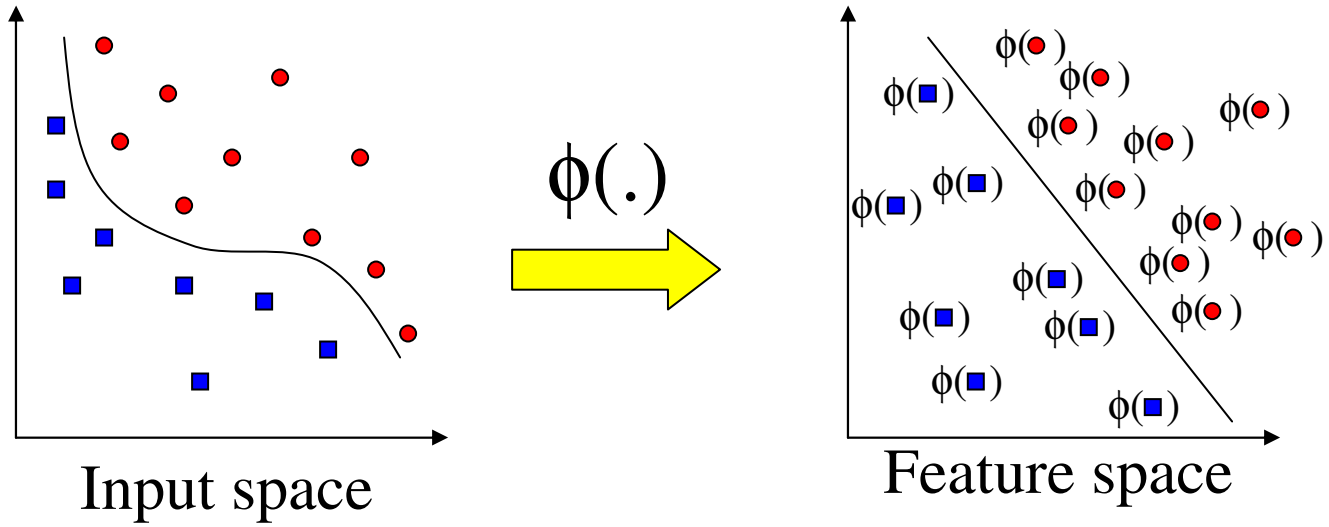
An example of small and large margins.



Transforming the Data

- The mathematical equation which describes the separating boundary between two classes should be simple.
- This is why we map the data of input space into feature space. The mapping (rearranging) involves increasing dimension of the feature space.
- The data points are mapped from the input space to a new feature space before they are used for training or for classification.
- After transforming the Data and after learning we look for an answer by examing simpler feature space.

$$x = (x_1, \ldots\ldots, x_n) \mapsto \phi(x) = (\phi_1(x), \ldots\ldots, \phi_N(x))$$



Learning

- Learning can be regarded as finding the maximum margin separating hiperplane between two classes of points. Suppose that a pair (w,b) defines a hyperplane which has the following equation:

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

- Let $\{x_1, ..., x_m\}$ be our data set and let $y_i \in \{1,-1\}$ be the class label of x_i.
- The decision boundary should classify all points correctly i.e. the following equations have to be satisfied:

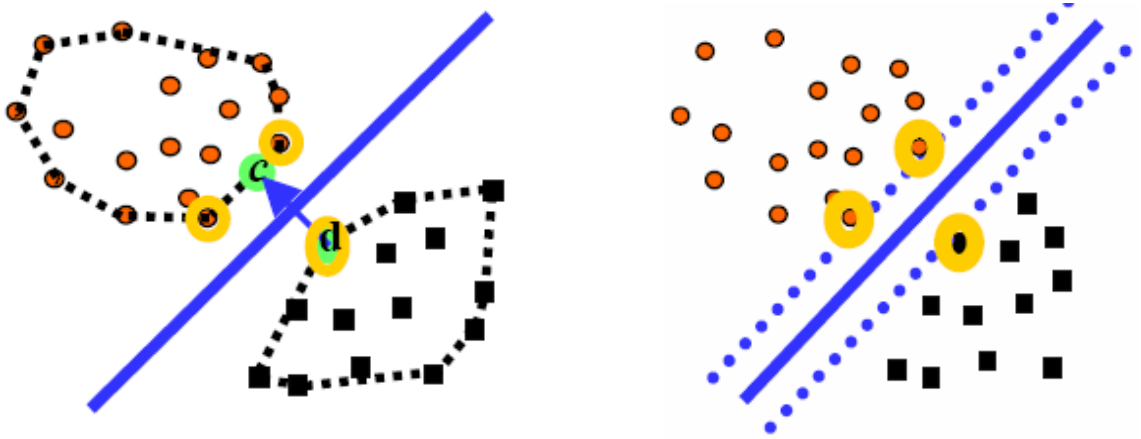
$$\begin{aligned} \mathbf{w} \cdot \mathbf{x}_i + b &\geq 1 \quad \text{if } y_i = 1 \\ \mathbf{w} \cdot \mathbf{x}_i + b &\leq -1 \quad \text{if } y_i = -1 \end{aligned} \quad \Rightarrow \quad y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$$

- Among all hyperplanes separating the data, there exists a unique one yielding the maximum margin of separation between the classes which can be determined in the following way;

$$\max_{\mathbf{w},b} \min\{\|\mathbf{x} - \mathbf{x}_i\| : \mathbf{x} \in \mathbb{R}^N, (\mathbf{w} \cdot \mathbf{x}) + b = 0, i = 1, \ldots, m\}$$

Support vectors (1)

- Let's notice that not all of the training points are important when choosing the hyperplane. To make it clear let's consider the following example:



- Let's make two covex hulls for the two seperate classes of pionts. It's is clear that the rear points are not important for choosing the decison boundary. At the above pictures the points which are relevant are marked by yellow colour. The points are called *Support Vectros.*

Support vectors (2)

- All training points have associated coefficients with them. The coefficients express the strength with which that points are embedded in the final decision function for any given test points. For all Support Vectors, which are the points that lie closest to the separating hyperplane, the coefficients are greater than 0. For the rest of the points the corresponding coefficients are equal to zero.
- The following equation describes the dependency between the training points and the decision boundary:

$$w = \sum_i \alpha_i y_i \phi(\mathbf{X}_i)$$

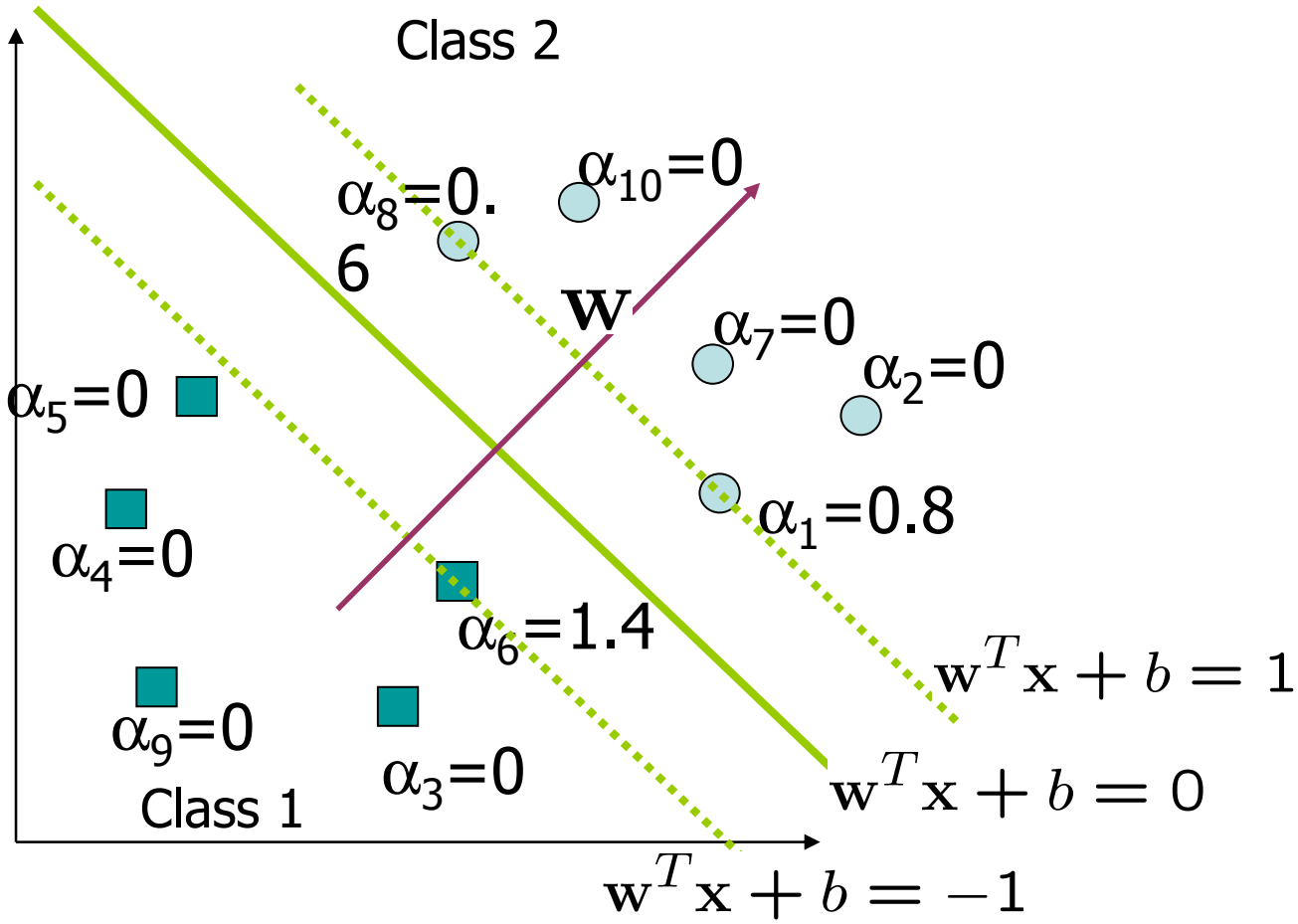
,where α_i are positive real numbers – the coefficients.

The coefficients need to satisfy the following conditions:

$$\sum_i \alpha_i - \sum_{ij} \alpha_i \alpha_j y_i y_j \langle \phi(\mathbf{X}_i), \phi(\mathbf{X}_j) \rangle \qquad \sum \alpha_i y_i = 0, \alpha_i > 0.$$

The coefficients have to be chosen to maximize the first equation.

Support vectors (3)



Kernel functions

- A Kernel is a function K, such that for all

$$x_1, x_2 \in X \qquad K(x_1, x_2) = \langle \phi(x_1) \cdot \phi(x_2) \rangle$$

- It computes the similarity of two data points in the feature space using dot product.
- The selection of an appropriate kernel function is important, since the kernel function defines the feature space in which the training set examples will be classified.
- The kernel expresses prior knowledge about the phenomenon being modeled, encoded as a similarity measure between two vectors.
- A support vector machine can locate a separating hyperplane in the feature space and classify points in that space without even representing the space explicitly, simply by defining a kernel function, that plays the role of the dot product in the feature space.

Predicting the classification

- Let X be a test point. The Support Vector Machine will predict the classification of the test point X using the following formula:

$$f(\mathbf{X}) = \mathbf{sign}\left(\langle \mathbf{w}, \phi(\mathbf{X}) \rangle - \mathbf{b}\right)$$

- The function returns 1 or -1 depends on which class the X point belongs to.

$\langle w \cdot \varphi(X) \rangle$ - this is a dot product of vector w and vector form the origin to the point $\varphi(X)$.

b - this is a shift of the hyperplane from the origin of the coordinate system.

References

- http://www.idi.ntnu.no/emner/it3704/lectures/papers/Bennett_2000_Support.pdf
- http://aya.technion.ac.il/karniel/CMCC/SVM-tutorial.pdf
- http://www.ecs.soton.ac.uk/~srg/publications/pdf/SVM.pdf
- http://en.wikipedia.org/wiki/Support_vector_machine

Exercises

- Explain what is the difference between soft margin and regular margin.
- Give two examples of kernel function.