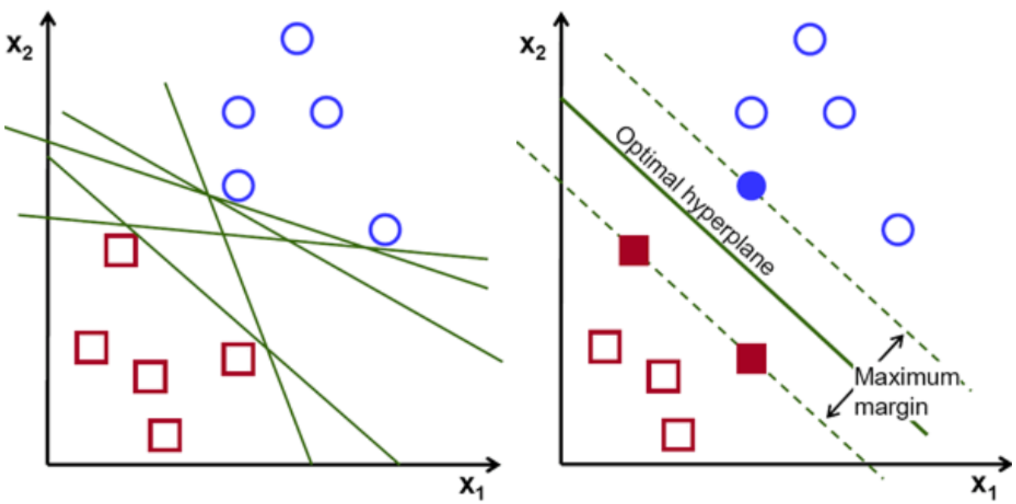
Topic 7.8 Support Vector Machines

- Simplest form:
  - Two input attributes, *A*1 and *A*2, such that the 2-D data are **linearly separable** (linear), that is, a straight line can be drawn to separate all the samples into two classes: +1 / ‘Buys computer = yes’ and -1 / ‘Buys computer = no’.
  - There may be an infinite number of separating lines; we want to find the “best” one.
  - If our data were 3-D (i.e., with three attributes), we would want to find the best separating **plane**. Generalizing to *n* dimensions, we would want to find the best **hyperplane**.
  - The term hyperplane is used for 2-D data also.



✓An SVM approaches the problem by searching for the **maximum margin hyperplane**.

- A separating hyperplane can be written as

$$\boldsymbol{W} \bullet \boldsymbol{X} + b = 0,$$

where $\boldsymbol{W}$ is a weight vector, namely, $\boldsymbol{W} = (w_1, w_2, \ldots, w_n)$; n is the number of attributes; and b is a scalar, often referred to as a bias.

- Let's consider 2-D samples, e.g., $\boldsymbol{X} = (x_1, x_2)$, where x_1 and x_2 are the values of attributes $A1$ and $A2$, respectively, for $\boldsymbol{X}$.

- If we think of b as an additional weight, w_0, we can rewrite above equation as $w_0 + w_1x_1 + w_2x_2 = 0$.
- Thus, any point that lies above the separating hyperplane satisfies

$$w_0 + w_1x_1 + w_2x_2 > 0.$$

- Similarly, any point that lies below the separating hyperplane satisfies

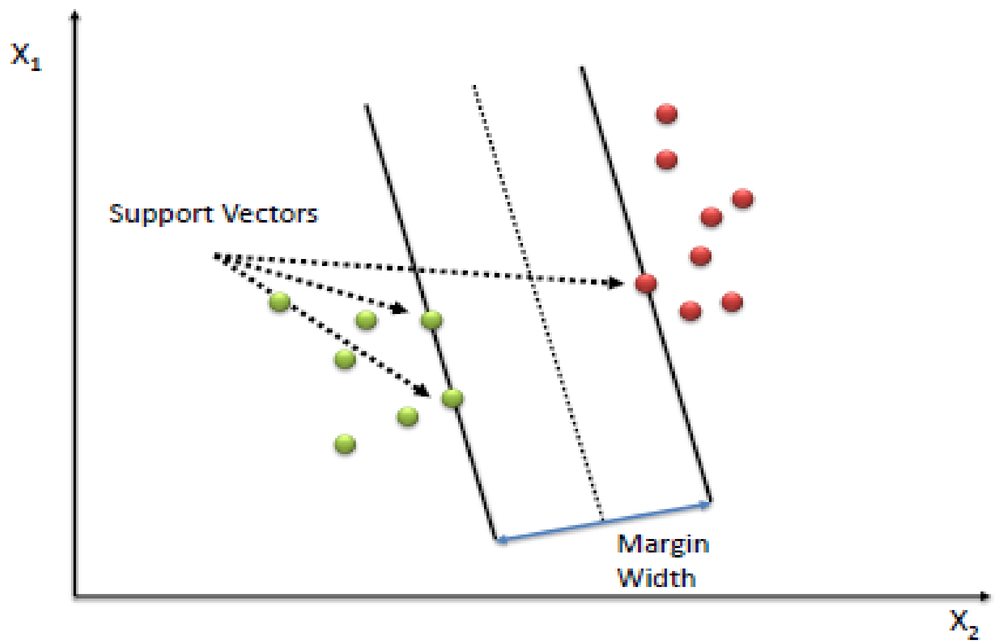
$$w_0 + w_1x_1 + w_2x_2 < 0.$$

- The weights can be adjusted so that the hyperplanes defining the "sides" of the margin can be written, in accordance with special loss function computations, as

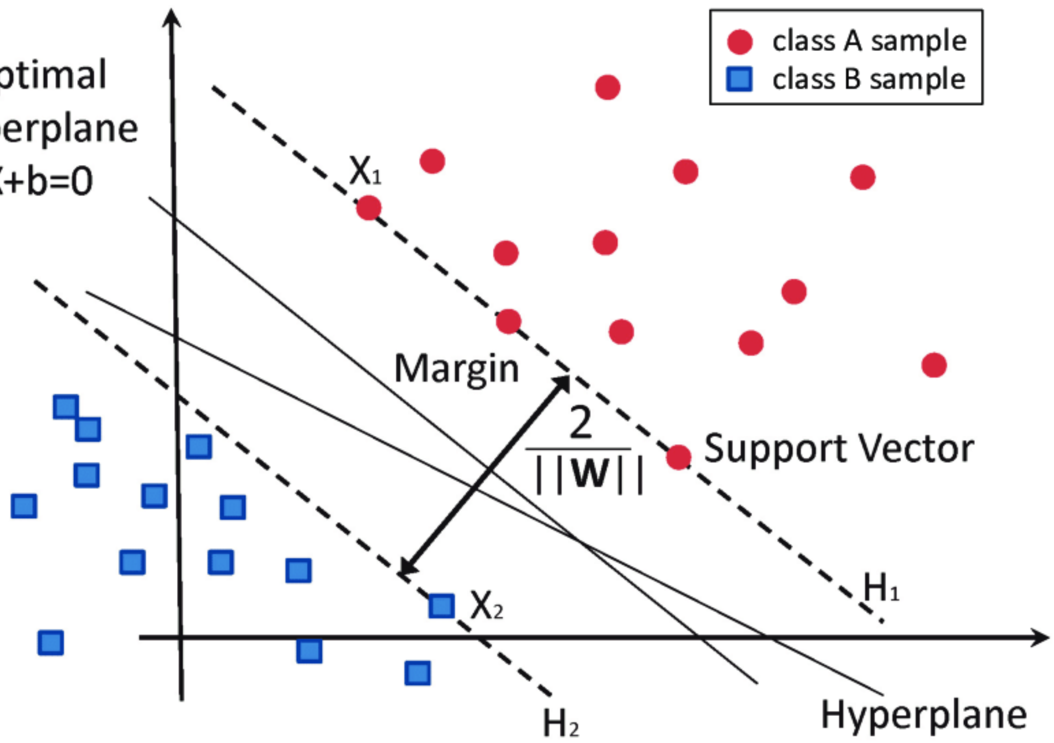
  $H1$: $w_0 + w_1x_1 + w_2x_2 \geq +1$ for *class* +1,

  $H2$: $w_0 + w_1x_1 + w_2x_2 \leq -1$ for *class* -1.

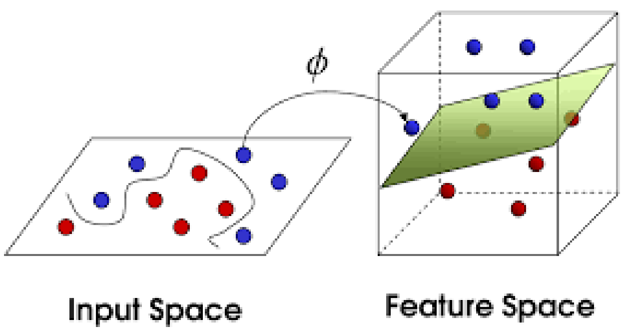
- Training samples that fall on hyperplanes $H1$ or $H2$ (i.e., the "sides" defining the margin) are called **support vectors**.

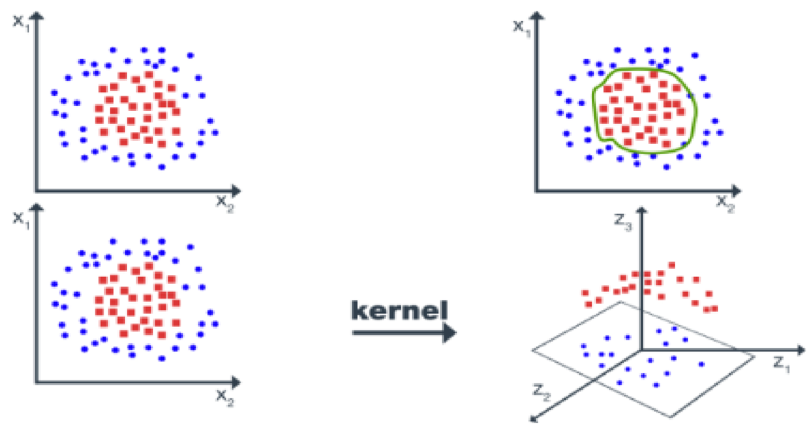


- The distance from the separating hyperplane to any point on *H*1 is $1/||W||$, where $||W||$ is the Euclidean norm of *W*. If $W = (w_1, w_2, \dots, w_n)$, then it is $(W \bullet W)^{1/2} = (w_1^2+w_2^2+\dots+w_n^2)^{1/2}$.
- By definition, this is equal to the distance from any point on *H*2 to the separating hyperplane. Therefore, the maximal margin is $2/||W||$.



- ✓ SVM: Supervised learning model for both classification and regression
- ✓ It is a method for the classification of both linear and nonlinear data.
- ✓ Uses a nonlinear mapping to transform the original training data into a higher dimension (**kernel trick**)





❑ Source of all pictures: Internet

- ✓ Data not separable in original space can be found easily separable in higher dimensions.
- ✓ And, higher dimensional linear separator may be actually nonlinear in the original space.

✓ They are flexible, can represent complex functions, but are, generally, like neural networks, resistant to overfitting.

✓ Key insight: Some examples are more important to select separator.

✓ Handwritten digit recognition, object recognition, speaker identification, protein classification for cancer detection, etc. are widely done using SVMs.