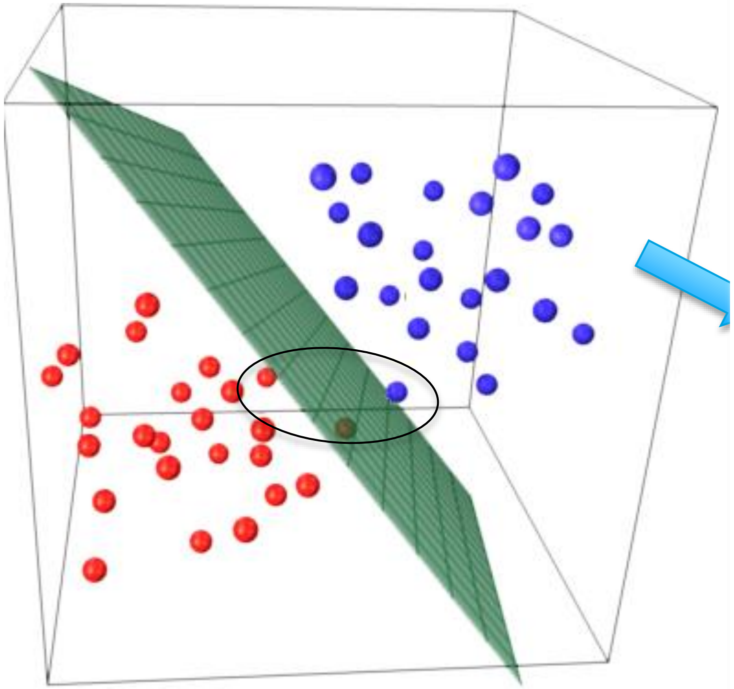
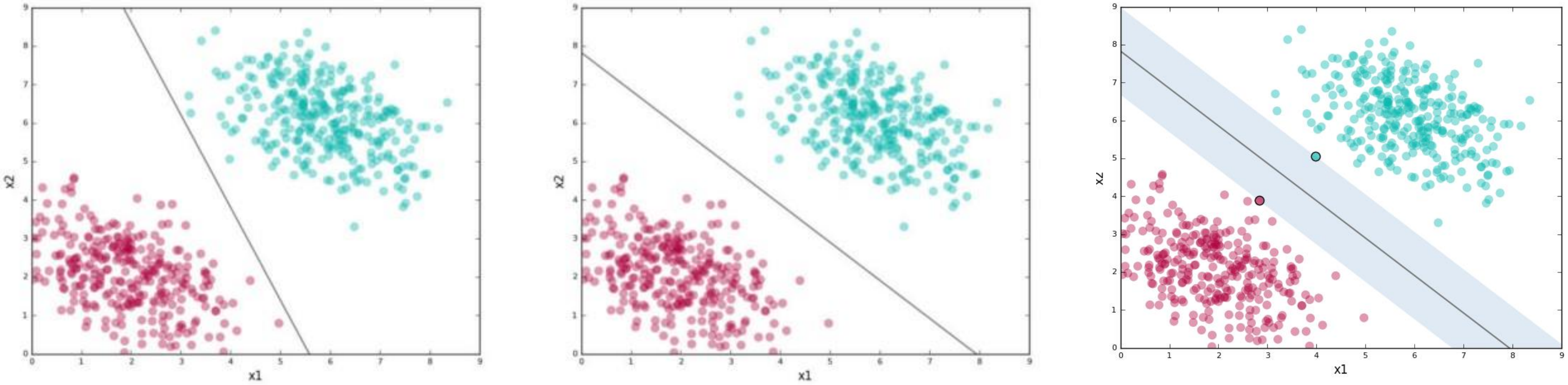
Support Vector Machine

Support Vector Machines

1. Known as maximum-margin hyperplane, find that linear model with max margi. Unlike the liner classifiers, objective is not minimizing sum of squared errors but finding a line/plane that separates two or more groups with maximum margins

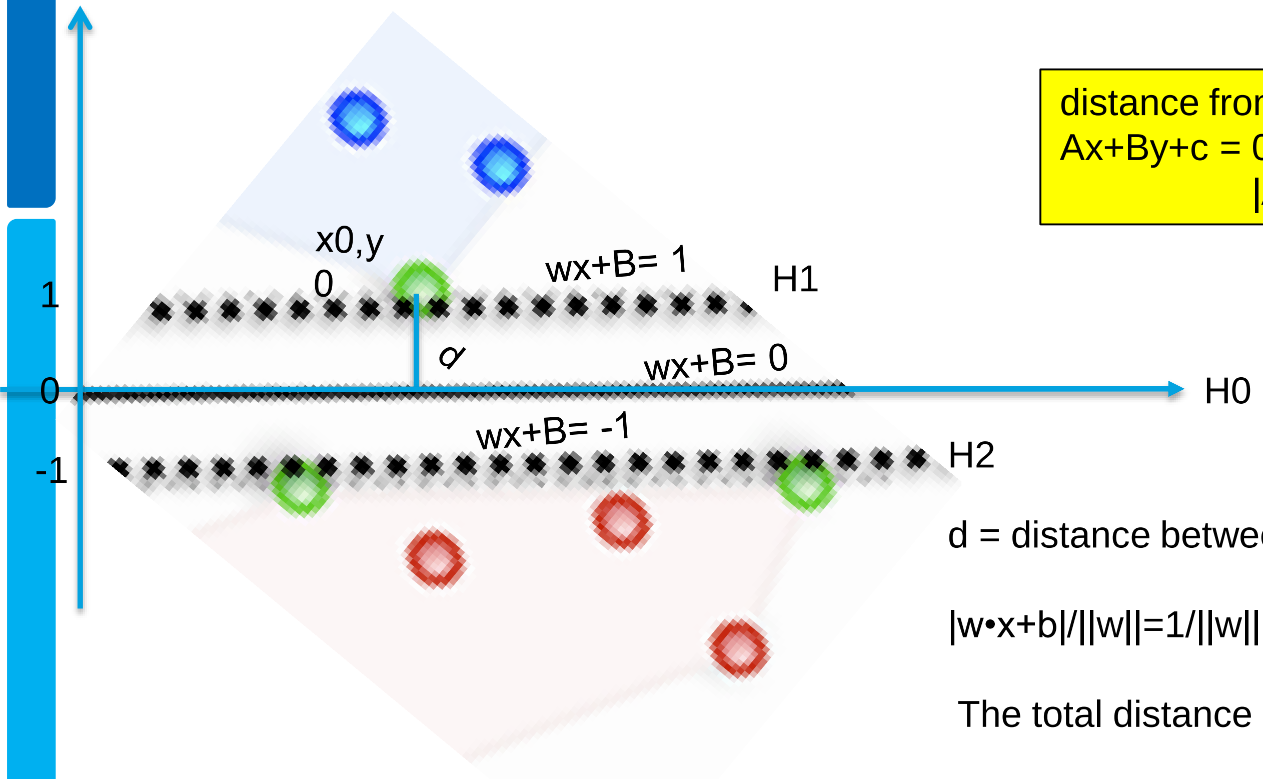


Support Vector Machines



1. First line does separate the two sets but it is too close to both red & green data points
2. Chances are that when this model is put in production, variance in both cluster data may force some data points on wrong side
3. The second line doesn't look so vulnerable to the variance. The two points nearest from different clusters define the margin around the line
4. SVMs try to find the second kind of line where the line is at max distance from both the clusters simultaneously

Support Vector Machines



distance from a point(x0,y0) to a line:
Ax+By+c = 0 is:
|Ax0 +By0 +c|/sqrt(A^2+B^2)

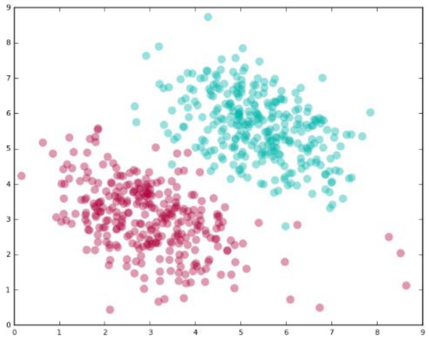
d = distance between H0 and H1 is

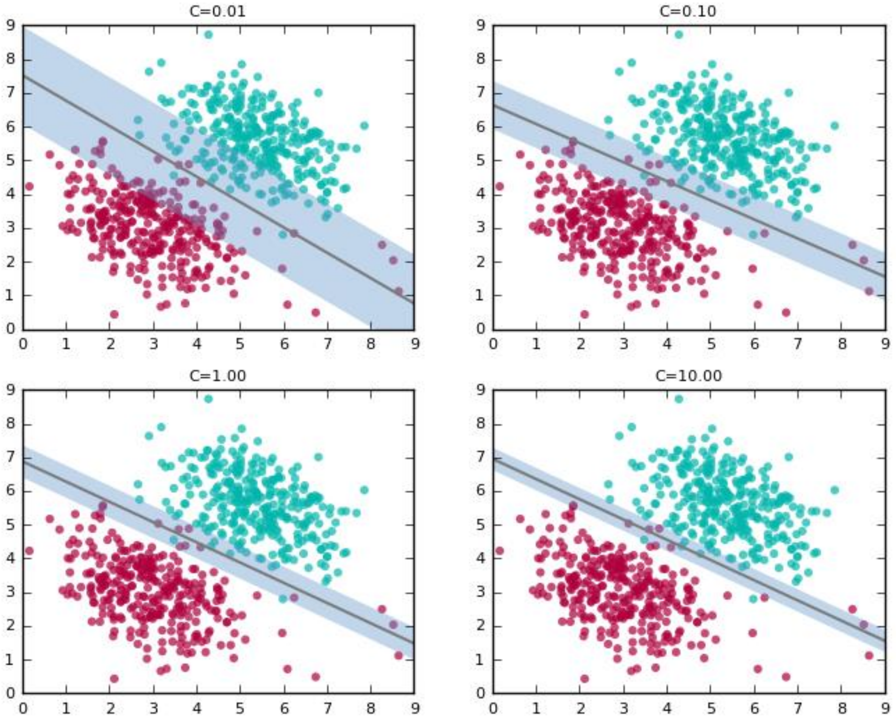
|w•x+b|/||w||=1/||w||,

The total distance between H1 and H2 is thus: 2/||w||

2. Think in terms of multi-dimensional space. SVM algorithm has to find the combination of weights across the dimensions such that the hyperplane has max possible margin around it
3. All the predictor variables have to be numeric and scaled.

Support Vector Machines Allowing Errors





1. Data in real world is typically not linearly separable.
2. There will always be instances that a linear classifier can’t get right
3. SVM provides a complexity parameter, a tradeoff between: wide margin with errors or a tight margin with minimal errors. As C increases, margins become tighter

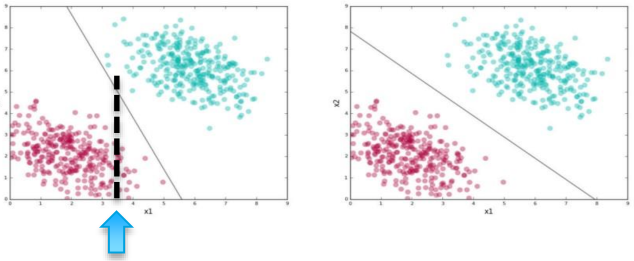
Support Vector Machines Basic Idea

1. Suppose we are given training data $\{(x1, y1),...,(xn, yn) \} \subset X \times R$, where X denotes the space of the input patterns (e.g. X = Rd).
2. Goal is to find a function f(x) that has at most **ε deviation** from the actually obtained targets yi for all the training data, and at the same time it **as flat as possible**
3. In other words, we do not care about errors as long as they are less than ε, but will not accept any deviation larger than this
4. f can take the form **$f(x) = (w, x) + b$ with $w \in X$, $b \in R$**
5. Flatness means that one seeks a small w. One way to ensure this is to minimize the $||w||^2 = (w, w)$.

Support Vector Machines Basic Idea

6. The problem can be represented as convex optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|w\|^2 \\ \text{subject to} & \begin{cases} y_i - \langle w, x_i \rangle - b & \leq & \varepsilon \\ \langle w, x_i \rangle + b - y_i & \leq & \varepsilon \end{cases} \end{array}$$



7. In the first picture, ||w||^2 is not minimized, neither the third constraint. Take the pointer to be x value, yi – (w, xi) – b is < e i.e. diff between green dot and the line but (w, xi) + b –yi i.e. diff between line an red dot is not < e.

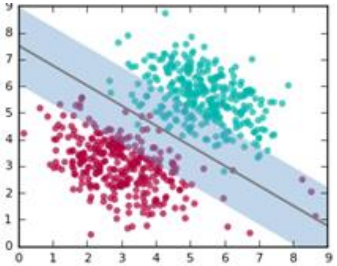
8. In second picture, all three constraints are met

9. Sometimes, it may not be possible to meet the constraint due to data points not being linearly separable so we may want to allow for some errors.

Support Vector Machines Basic Idea

10. We introduce slack variables ξi, ξi∗ to cope with otherwise infeasible constraints of the optimization problem and this is known as soft margin classifier

$$
\begin{array}{ll}
\text{minimize} & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*) \\
\text{subject to} & \left\{\begin{array}{lll} y_i - \langle w, x_i \rangle - b & \leq & \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i & \leq & \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq & 0 \end{array}\right.
\end{array}
$$



11. The epsilon term allows some errors i.e. data points lie within the error margins where error margins is e + epsilon

Machine Learning (Support Vector Machines)

Strengths	Weakness
Very stable as it depends on the support vectors only. Not influenced by any other data point including outliers	Computationally intensive
Can be adapted to classification or numeric prediction problems	Prone to over fitting training data
Capable of modelling relatively more complex patterns than nearly any algorithm	Assumes linear relation between dependent and independent variables
Makes no assumptions about underlying data sets	Generally treated as a blackbox model