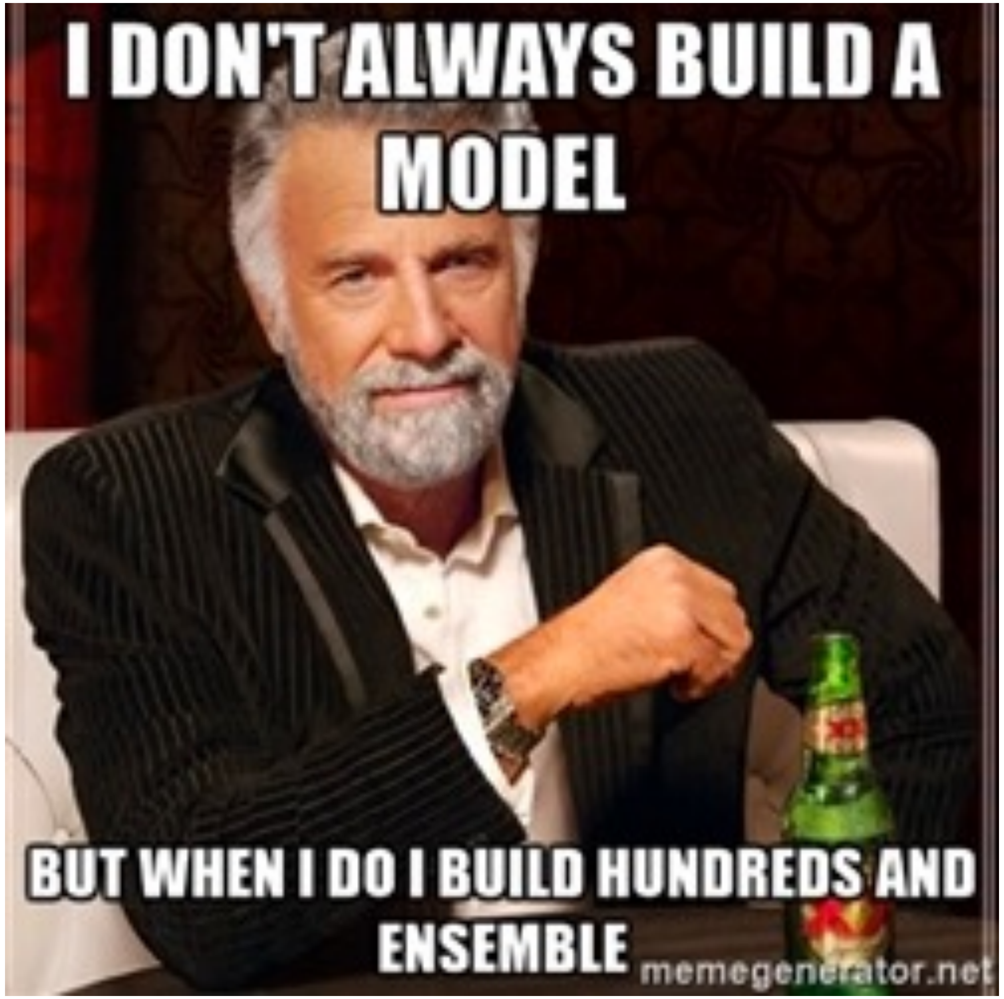
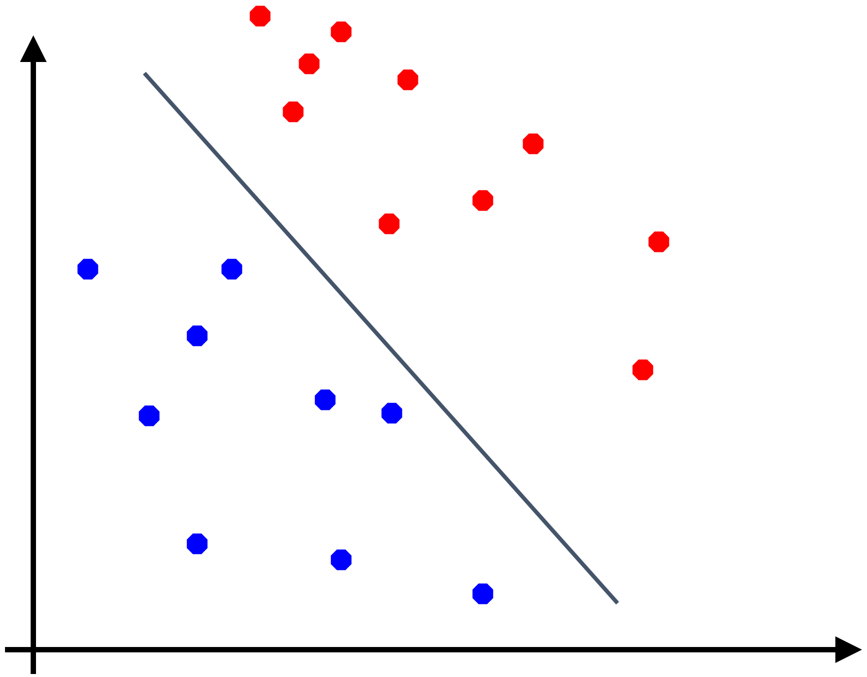

I DON'T ALWAYS BUILD A MODEL
BUT WHEN I DO I BUILD HUNDREDS AND ENSEMBLE
memegenerator.net

Support Vector Machines

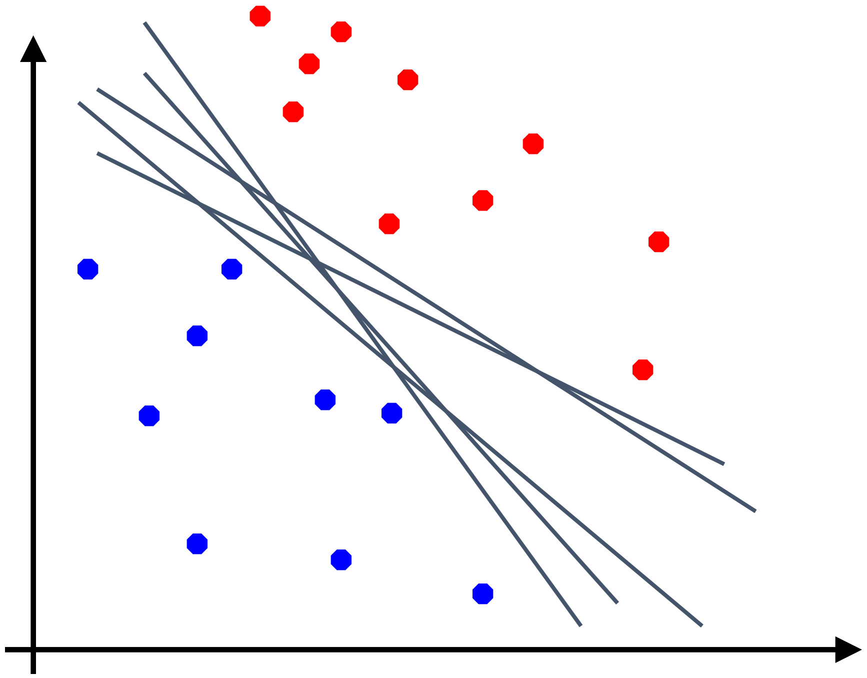
Linear Separators

- Binary classification can be viewed as the task of separating classes in feature space:



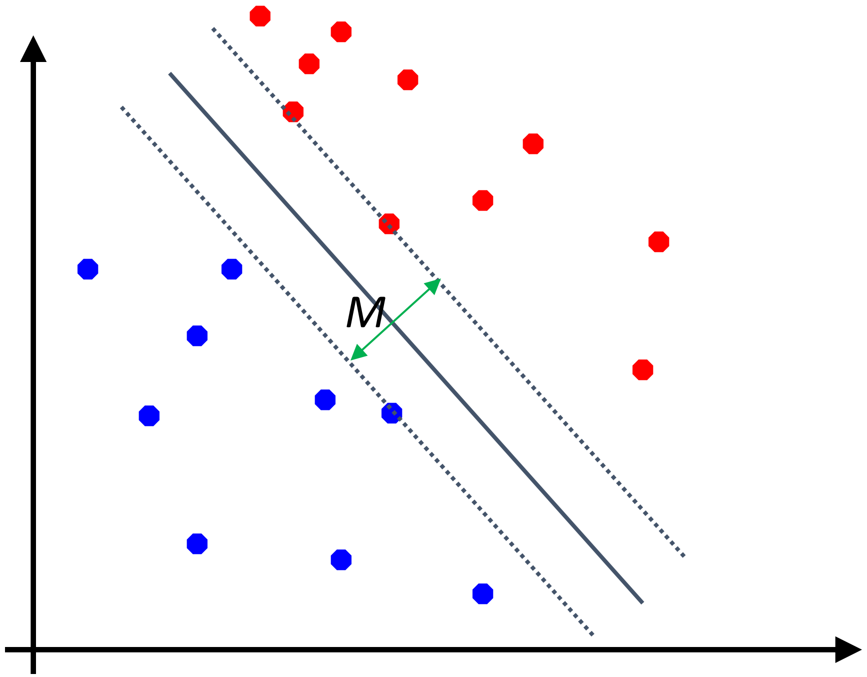
Linear Separators

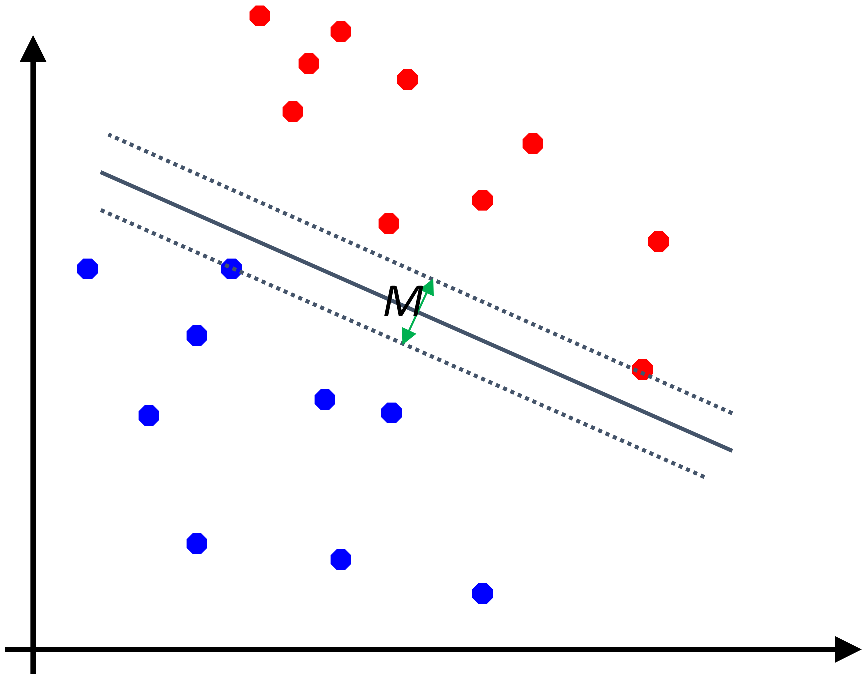
- Which linear separator is optimal?



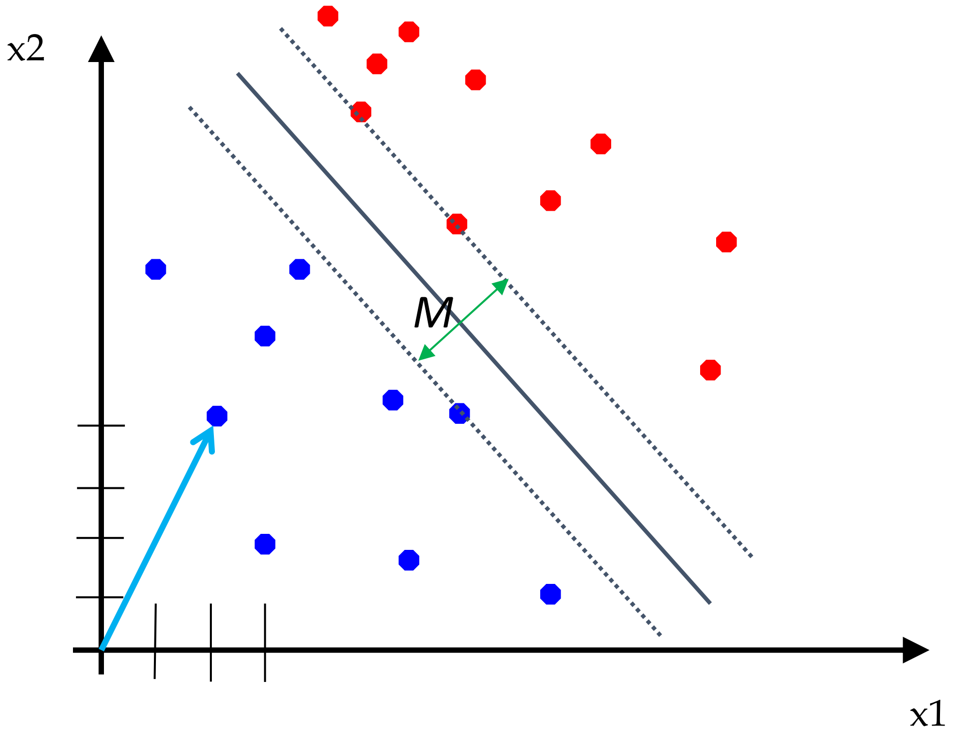
Maximum Margin

- Goal: find the hyperplane with the maximum margin



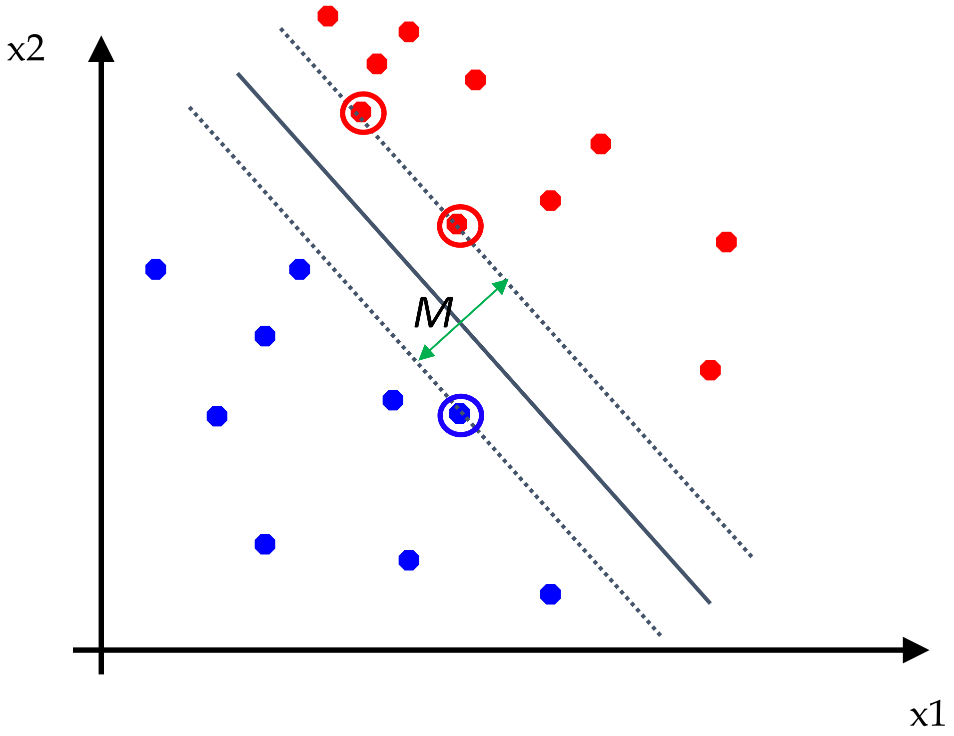


Vectors



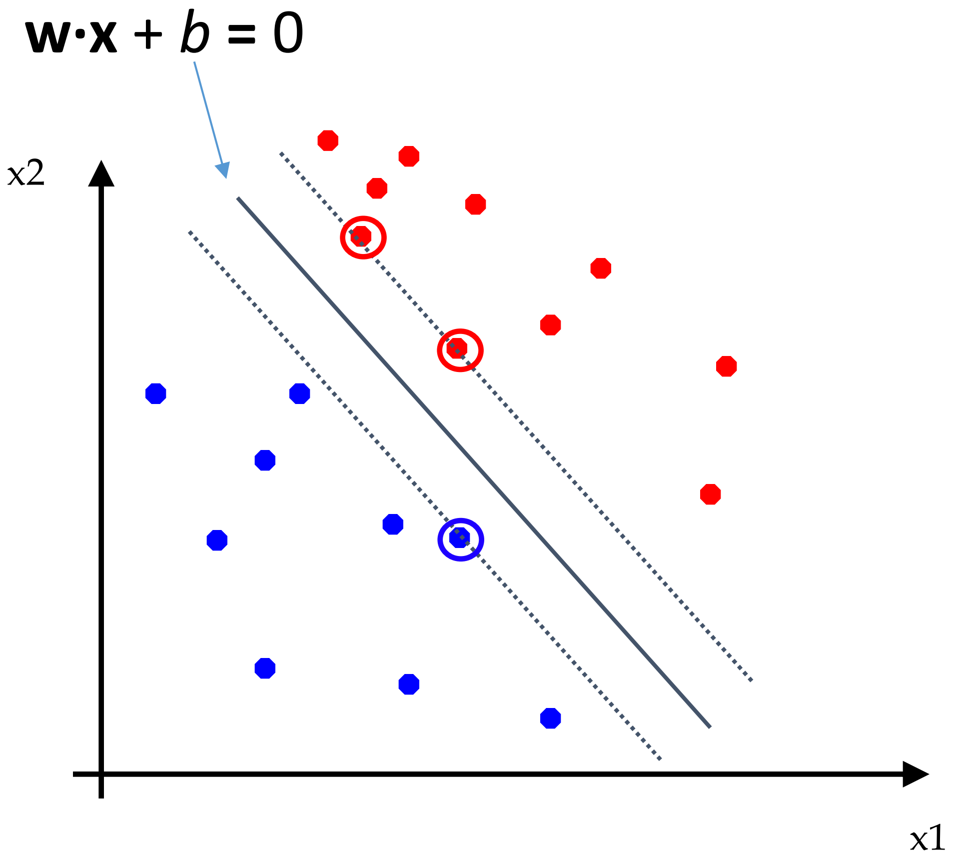
Each training point is denoted by $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i = (x1_i, x2_i, \ldots, xd_i)^T$ for the i^{th} example, and $y_i \in \{-1, 1\}$ denoting its class label.

Support Vectors

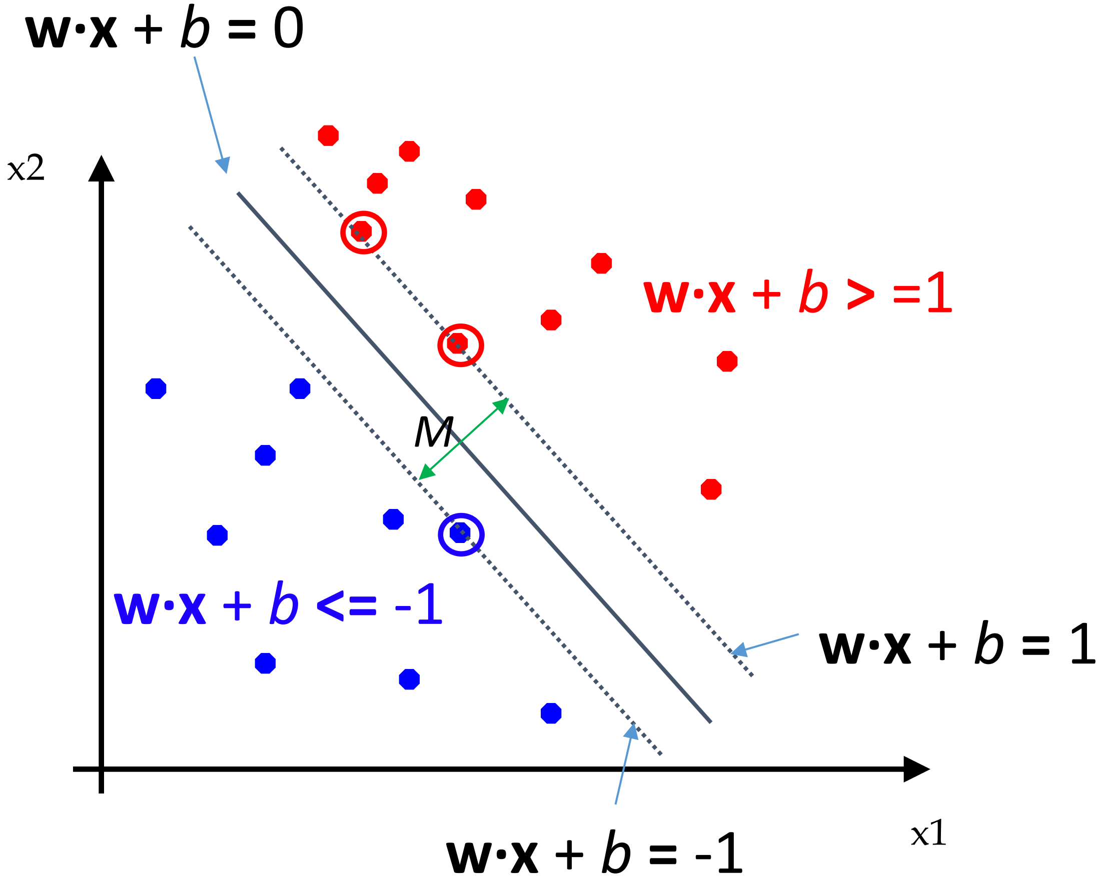


Support vectors are the points that if you moved them, it would change the optimal hyperplane

Hyperplane



Learning



$\mathbf{w}\cdot\mathbf{x_i} + b \;>= 1$, if $y_i = 1$,
$\mathbf{w}\cdot\mathbf{x_i} + b \;<= -1$, if $y_i = -1$

Can be combined to:
$y_i(\mathbf{w}\cdot\mathbf{x_i} + b) >= 1$

Also, want to maximize $M = \frac{2}{||\mathbf{w}||}$

Which is the same as minimizing $||\mathbf{w}||$,

which is the same as minimizing $\frac{||\mathbf{w}||^2}{2}$

Objective Function

- The learning task in SVM can be formalized as the following constrained optimization problem:

$$\min \frac{||\mathbf{w}||^2}{2}$$

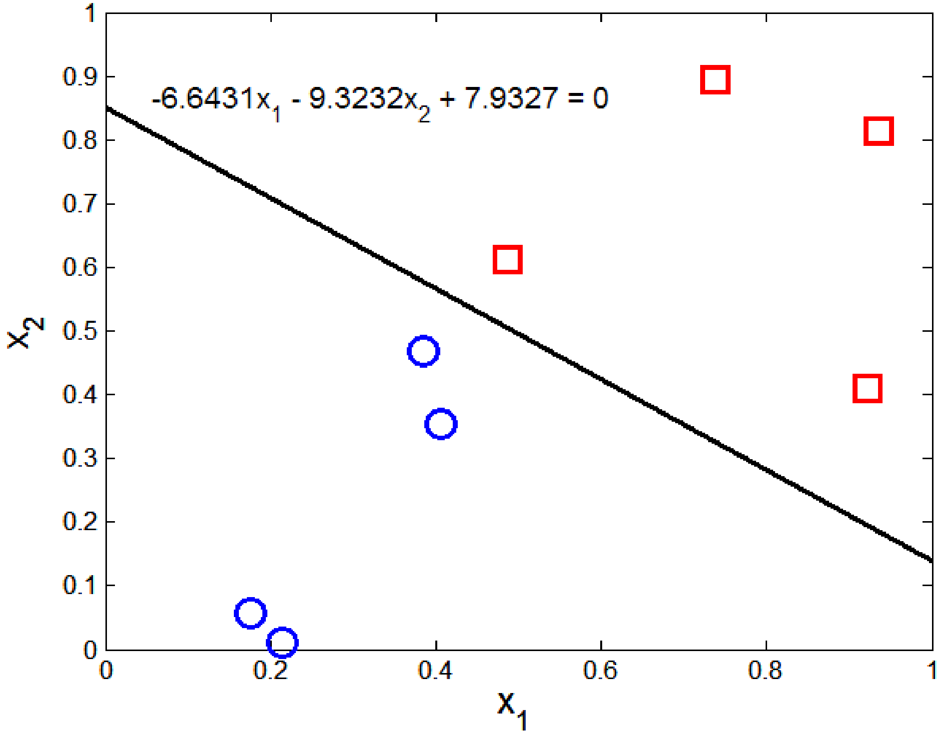
Subject to $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) >= 1$

- This is a convex quadratic minimization problem that can be solved with the Lagrange Multiplier method

http://www.engr.mun.ca/~baxter/Publications/LagrangeForSVMs.pdf

IMPORTANT: Data must be scaled so that all features are on the same scale!

Example of Linear SVM



Support vectors

x1	x2	y	λ
0.3858	0.4687	1	65.5261
0.4871	0.611	-1	65.5261
0.9218	0.4103	-1	0
0.7382	0.8936	-1	0
0.1763	0.0579	1	0
0.4057	0.3529	1	0
0.9355	0.8132	-1	0
0.2146	0.0099	1	0

Learning Linear SVM

- Decision boundary depends only on support vectors
  - If you have data set with same support vectors, decision boundary will not change

- How to classify using SVM once **w** and *b* are found? Given a test record, x_i

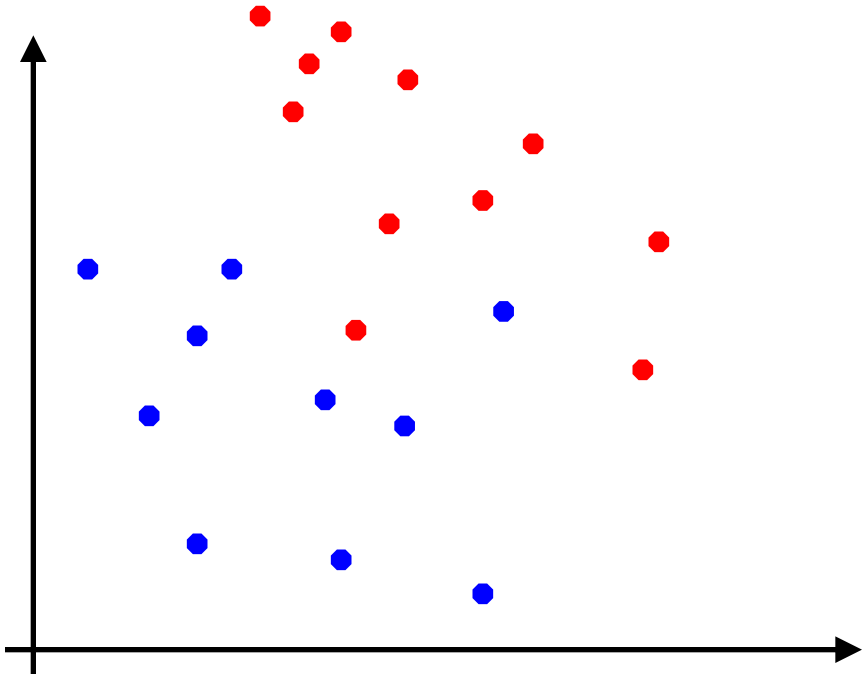
$$f(\mathbf{x}) = sign(\mathbf{w} \cdot \mathbf{x} + b)$$

Multiclass classification

- One vs Rest or One Against All (OvR or OAA)
    - One-vs.-rest[2]:182, 338 (OvR or *one-vs.-all*, OvA or *one-against-all*, OAA) strategy involves training a single classifier per class, with the samples of that class as positive samples and all other samples as negatives. This strategy requires the base classifiers to produce a real-valued confidence score for its decision, rather than just a class label; discrete class labels alone can lead to ambiguities, where multiple classes are predicted for a single sample.
- One vs One (OvO)
    - In the *one-vs.-one* (OvO) reduction, one trains $K (K - 1) / 2$ binary classifiers for a K-way multiclass problem; each receives the samples of a pair of classes from the original training set, and must learn to distinguish these two classes. At prediction time, a voting scheme is applied: all $K (K - 1) / 2$ classifiers are applied to an unseen sample and the class that got the highest number of "+1" predictions gets predicted by the combined classifier.

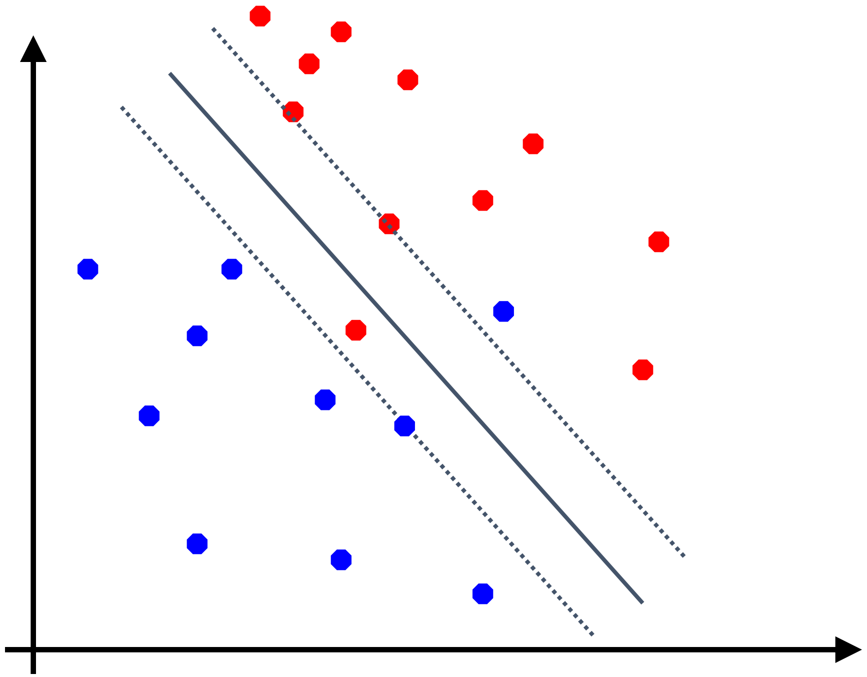
https://en.wikipedia.org/wiki/Multiclass_classification

Non-linearly Separable Data



Soft Margin SVM

- Consider a trade-off between the width of the margin and the number of training errors



$$\min_{\mathbf{w}} \frac{||\mathbf{w}||^2}{2}$$

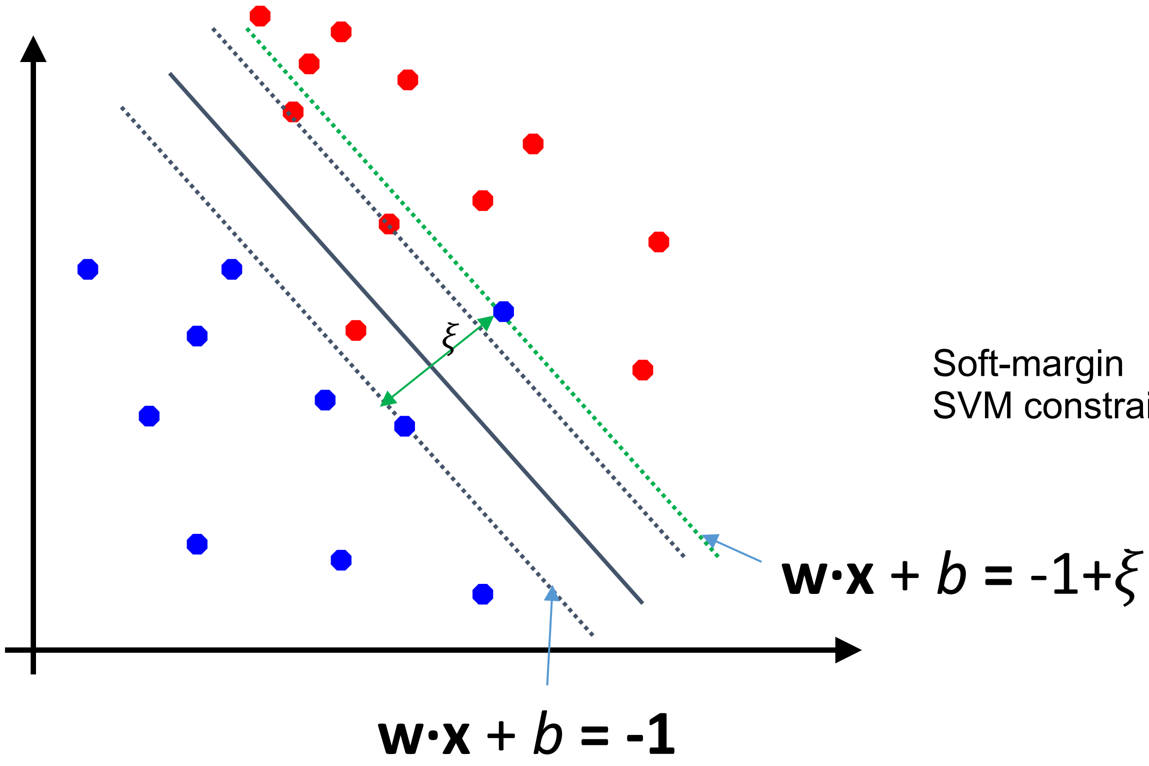
Original constraint

Subject to $y_i(\mathbf{w}\cdot\mathbf{x}_i + b) >= 1$

$\mathbf{w}\cdot\mathbf{x}_i + b >= 1$, if $y_i = 1$,

$\mathbf{w}\cdot\mathbf{x}_i + b <= -1$, if $y_i = -1$

Slack Variables



$$\min_{\mathbf{w}} \frac{\|\mathbf{w}\|^2}{2}$$

Soft-margin SVM constraint

Subject to

$\mathbf{w}\cdot\mathbf{x_i} + b \geq 1-\xi_i$, if $y_i = 1$,
$\mathbf{w}\cdot\mathbf{x_i} + b \leq -1+\xi_i$, if $y_i = -1$

where $\xi_i > 0$

Soft-Margin Objective Function

$$\min \frac{||w||^2}{2} + C\left(\sum_{i=1}^{N} \xi_i\right)$$

Subject to $y_i(\mathbf{w} \cdot \mathbf{x_i} + b) >= 1 - \xi_i$

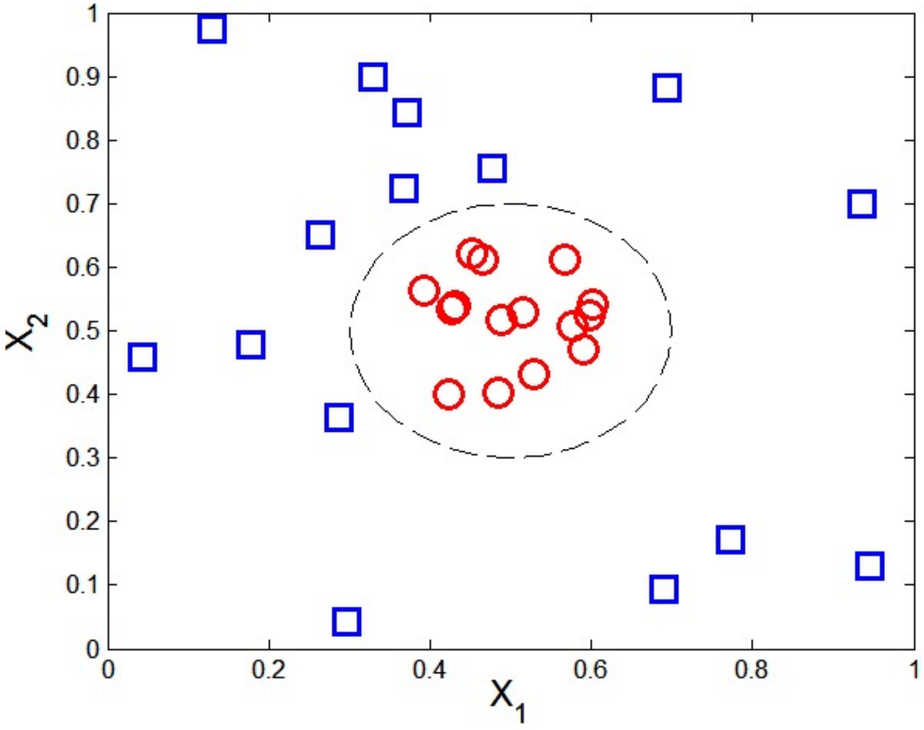
where C is a cost for each misclassification

- This is a convex quadratic minimization problem that can be solved with the Lagrange Multiplier method

IMPORTANT: Data must be scaled so that all features are on the same scale!

Nonlinear Support Vector Machines

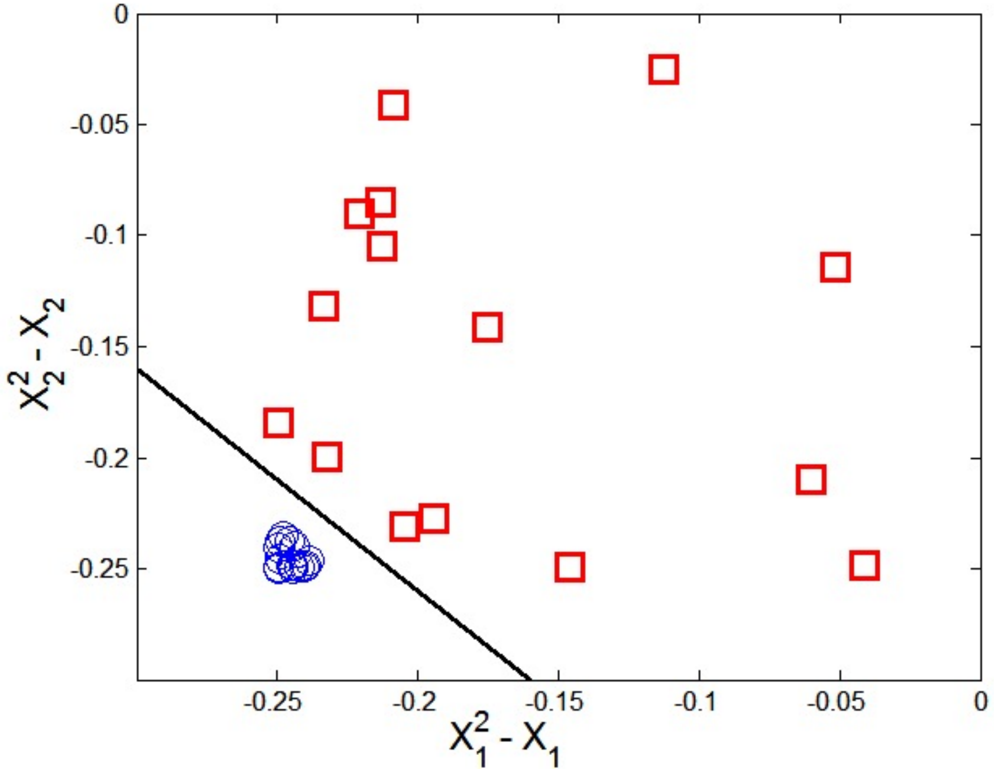
- What if decision



$$y(x_1, x_2) = \begin{cases} 1 & \text{if } \sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2} > 0.2 \\ -1 & \text{otherwise} \end{cases}$$

Nonlinear Support Vector Machines

- Transform the data from its original coordinate space, $\mathbf{x}$, to a new space, $\Phi(\boldsymbol{x})$



$$x_1^2 - x_1 + x_2^2 - x_2 = -0.46.$$

$$\Phi : (x_1, x_2) \longrightarrow (x_1^2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, 1).$$

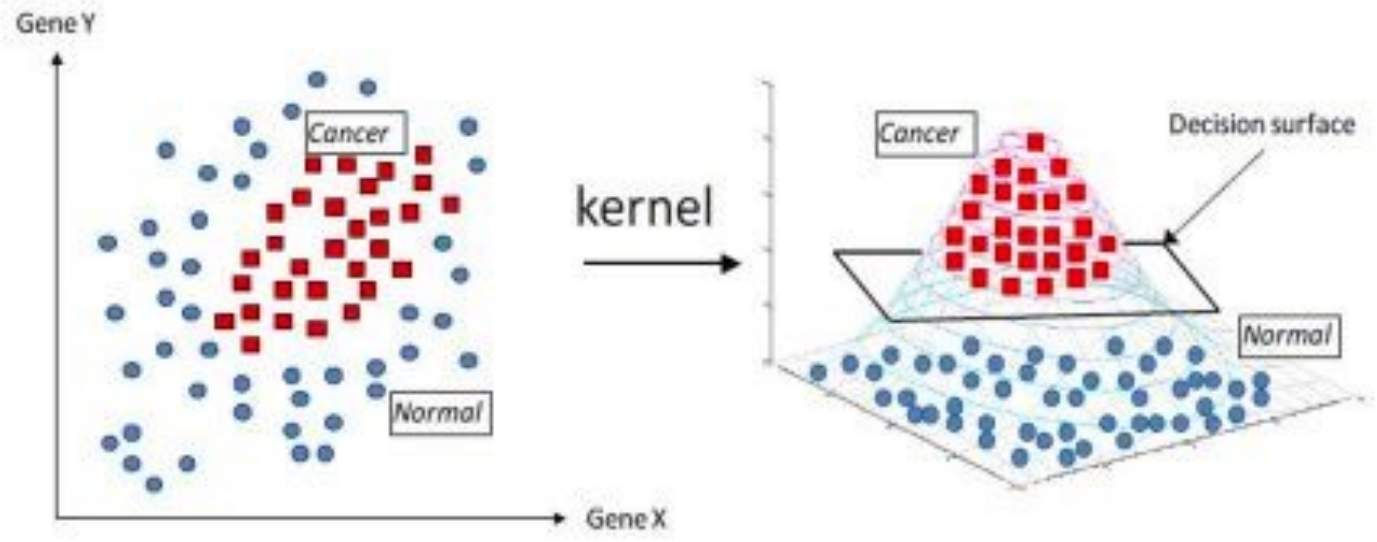
$$w_4 x_1^2 + w_3 x_2^2 + w_2 \sqrt{2} x_1 + w_1 \sqrt{2} x_2 + w_0 = 0.$$

Decision boundary:

$$\vec{w} \bullet \Phi(\vec{x}) + b = 0$$

Kernel Methods

- Transform the data to a higher dimensional space, so that it can be linearly separated



[Kernel functions]

Learning Nonlinear SVM

- Kernel Trick:
  - $\Phi(x_i) \bullet \Phi(x_j) = K(x_i, x_j)$
  - $K(x_i, x_j)$ is a kernel function (expressed in terms of the coordinates in the original space)
    - Examples:

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^p$$
$$K(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/(2\sigma^2)}$$
$$K(\mathbf{x}, \mathbf{y}) = \tanh(k\mathbf{x} \cdot \mathbf{y} - \delta)$$

- Advantages of using kernel:
  - Don't have to know the mapping function Φ
  - Computing dot product $\Phi(x_i) \bullet \Phi(x_j)$ in the original space avoids curse of dimensionality

Characteristics of SVMs

- SVM is a convex optimization problem, which means there are known algorithms to solve it and find the global minimum. Other classification algorithms (like decision trees) use a greedy strategy and therefore may not arrive at the globally optimal solution.
- Data needs to be scaled so that all feature are on the same scale
- Can be extended to multi-class problems via multi-class partitioning
- Not susceptible to the curse of dimensionality
- Selecting the right kernel function and cost can be difficult
- One of the most widely used classification algorithms