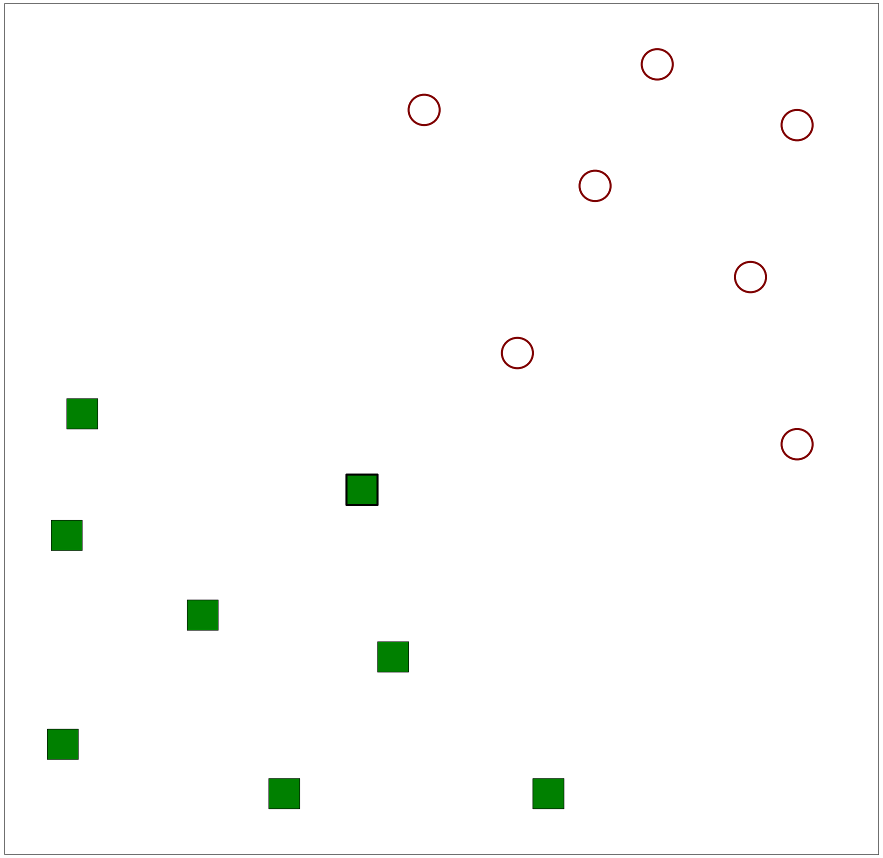
Support Vector Machines

Introduction to Data Mining, 2^{nd} Edition

by

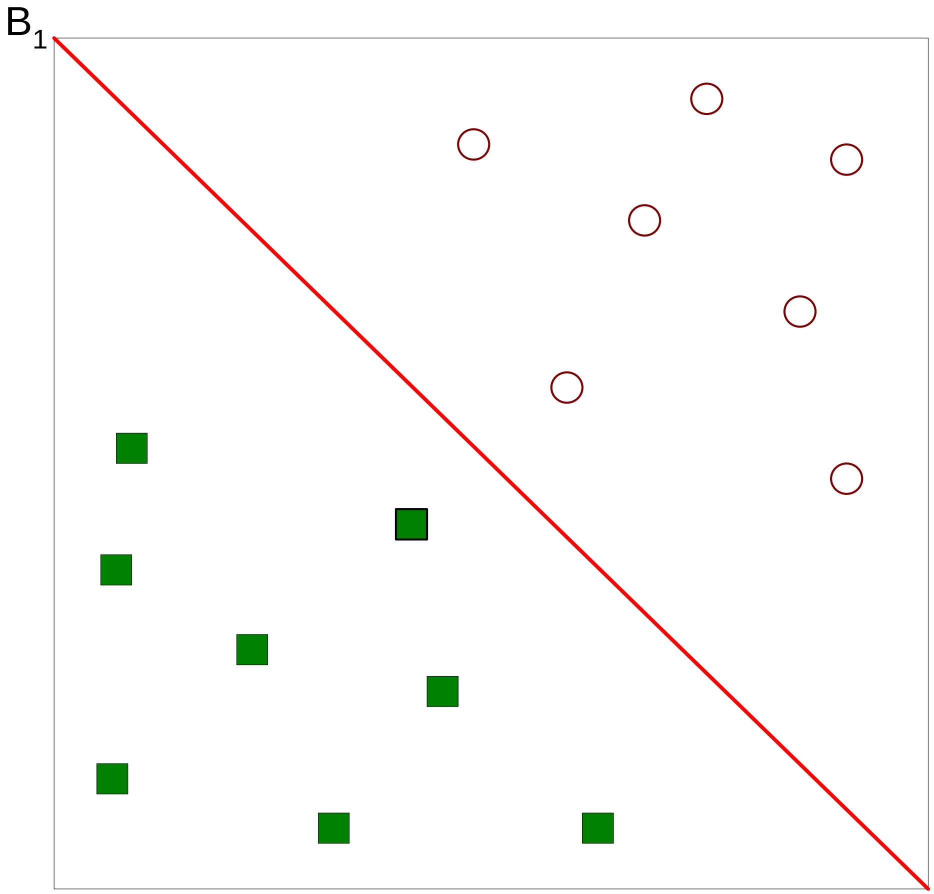
Tan, Steinbach, Karpatne, Kumar

Support Vector Machines



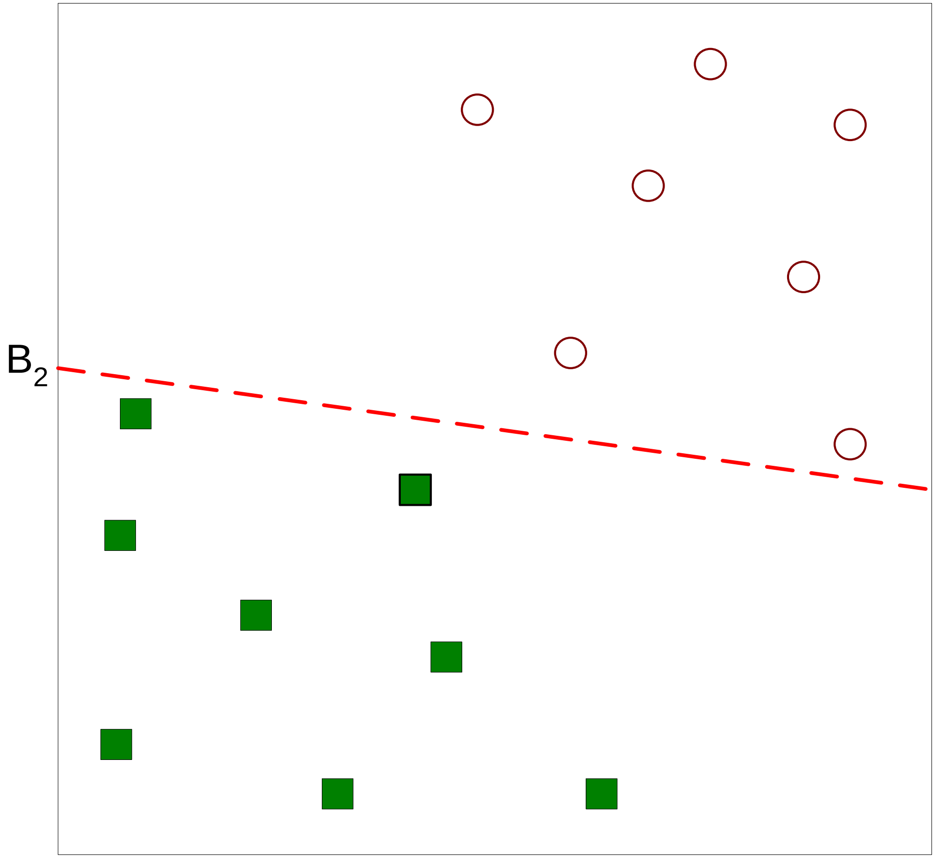
- Find a linear hyperplane (decision boundary) that will separate the data

Support Vector Machines



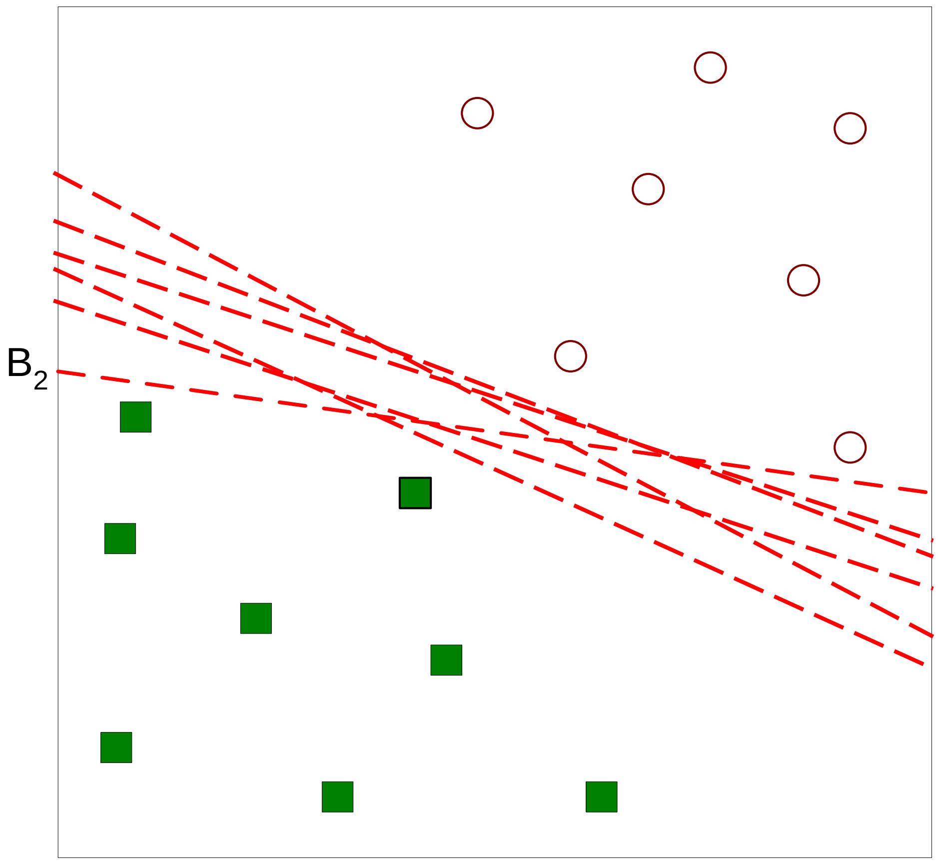
- One Possible Solution

Support Vector Machines



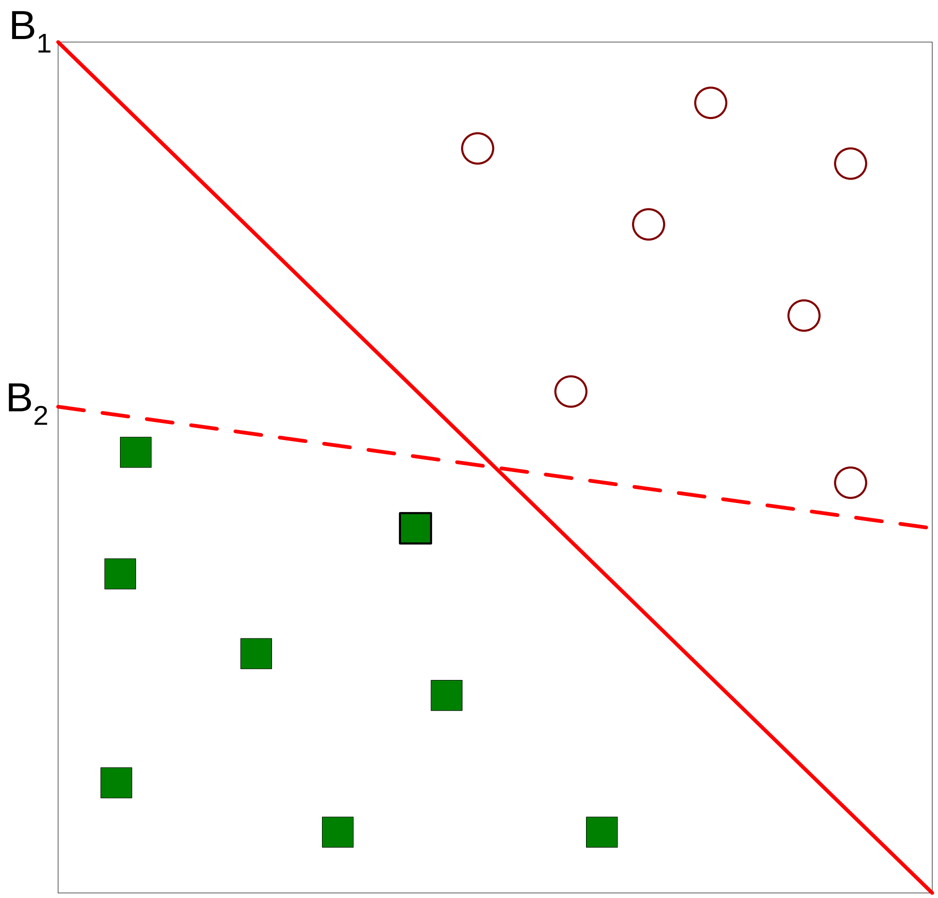
- Another possible solution

Support Vector Machines



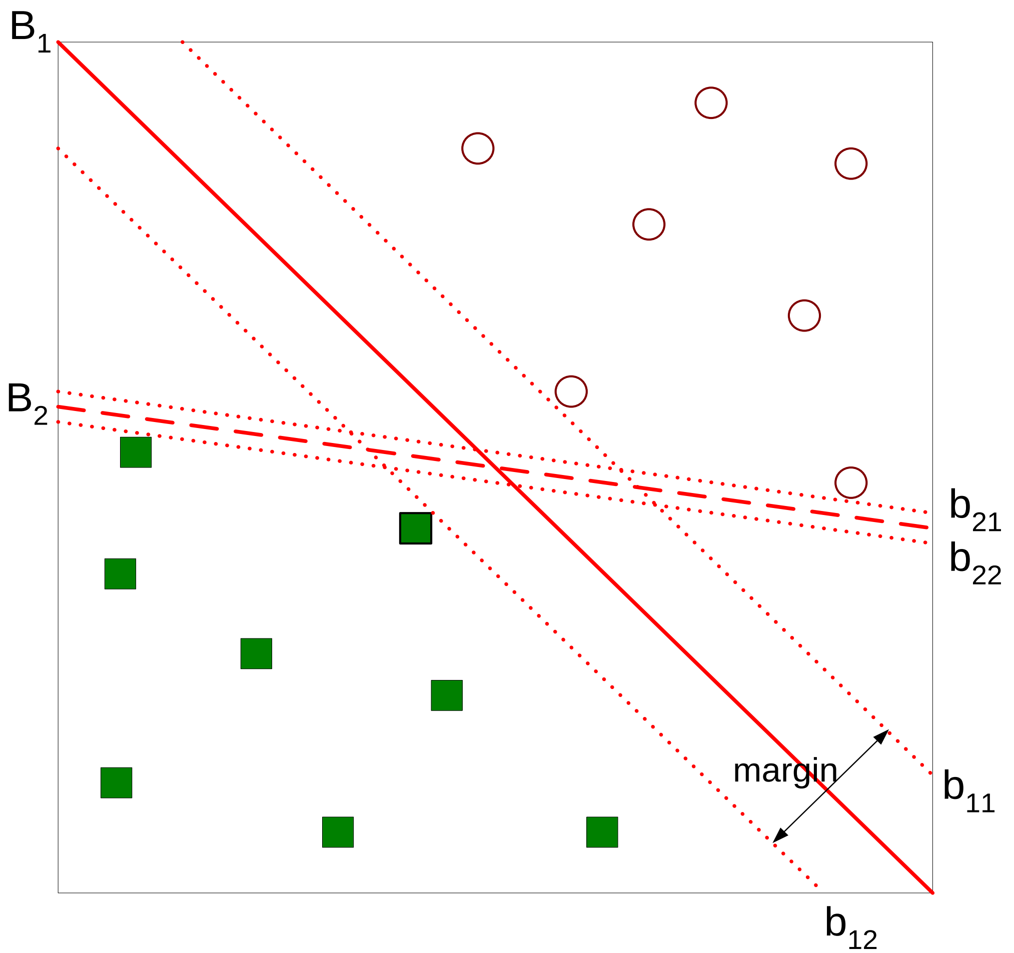
- Other possible solutions

Support Vector Machines



- Which one is better? B1 or B2?
- How do you define better?

Support Vector Machines



- Find hyperplane maximizes the margin => B1 is better than B2

SVM – Linear decision boundary

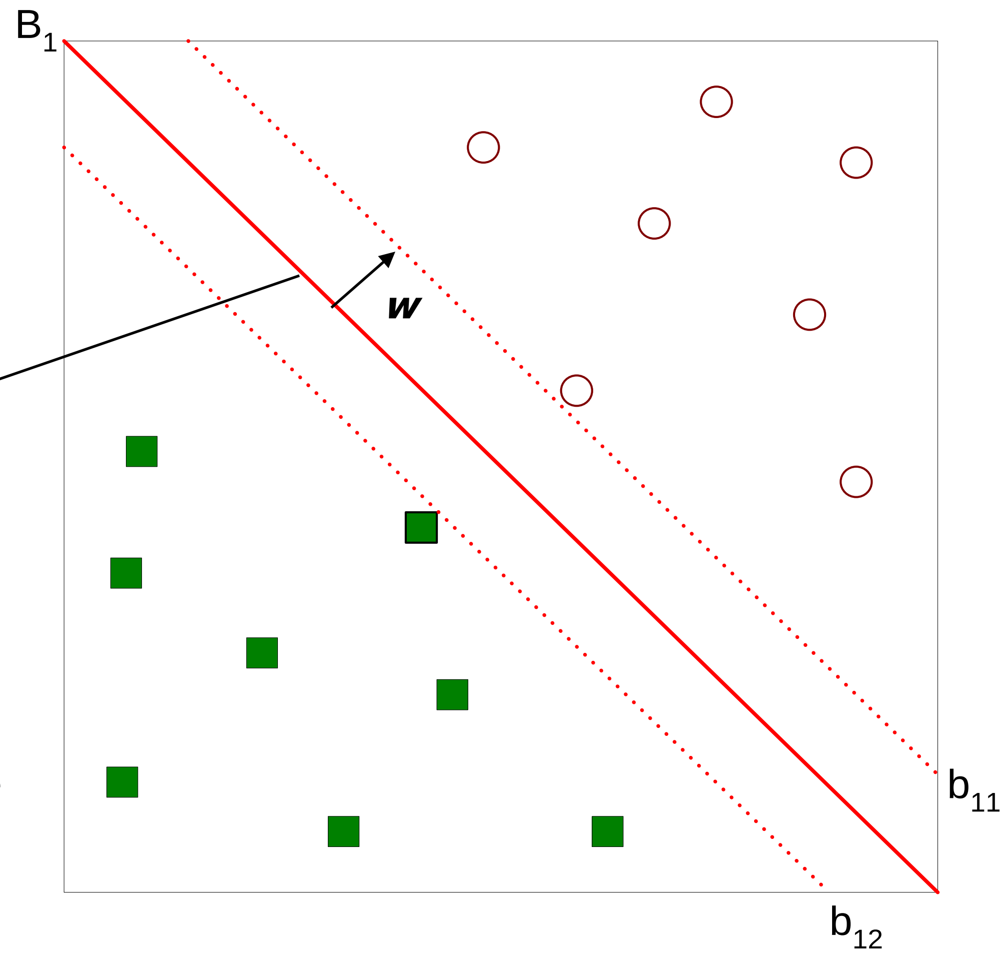
$(x_i, y_i), i = (1, 2, \dots, N)$

$x_i = (x_{i1}, x_{i2}, \dots, x_{id})^T$

$y_i \in= \{-1, 1\}$

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

Decision boundary



w is the normal direction of the hyperplane
b is a form of threshold.

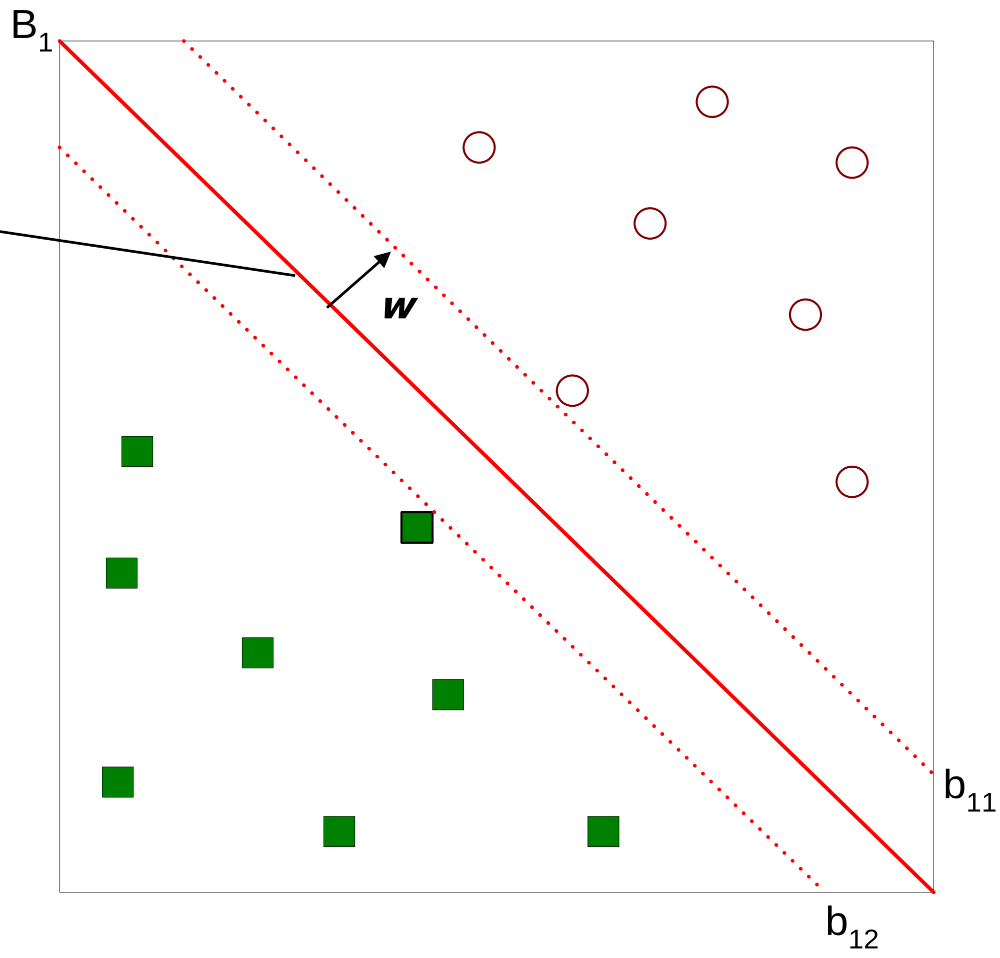
Support Vector Machines

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

Decision boundary

$$\boldsymbol{w} \cdot \boldsymbol{x_a} + \boldsymbol{b} = \boldsymbol{0}$$

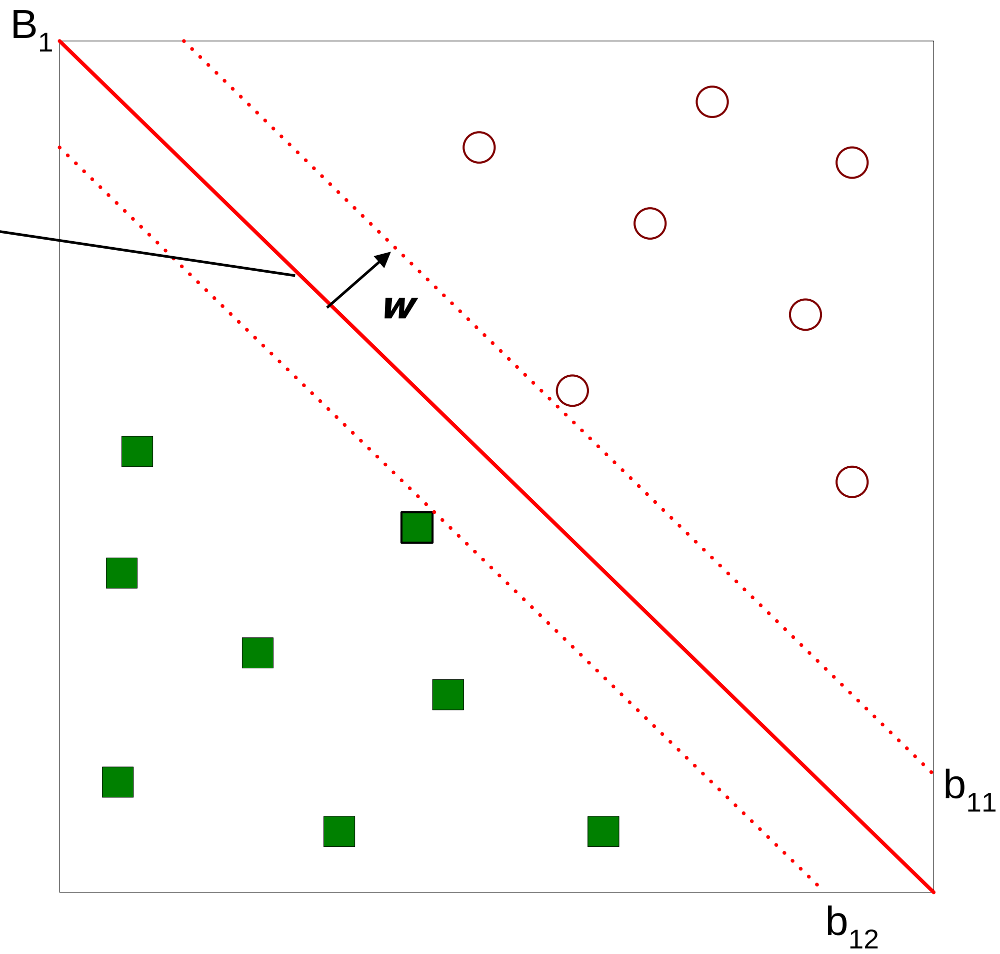
$$\boldsymbol{w} \cdot \boldsymbol{x_b} + \boldsymbol{b} = \boldsymbol{0}$$



Support Vector Machines

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

Decision boundary

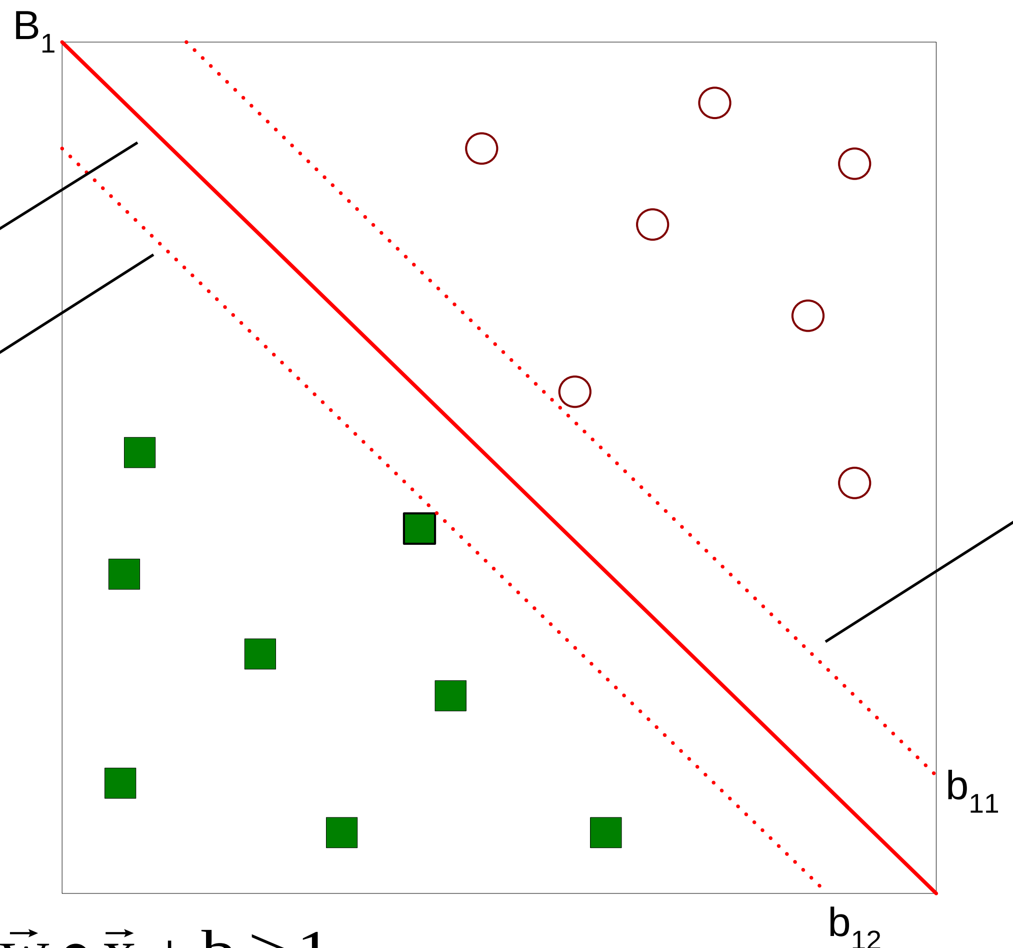


$$\boldsymbol{w} \cdot \boldsymbol{x_a} + \boldsymbol{b} = \boldsymbol{0}$$

$$\boldsymbol{w} \cdot \boldsymbol{x_b} + \boldsymbol{b} = \boldsymbol{0}$$

$$\boldsymbol{w} \cdot (\boldsymbol{x_b} - \boldsymbol{x_a}) = \boldsymbol{0}$$

Support Vector Machines



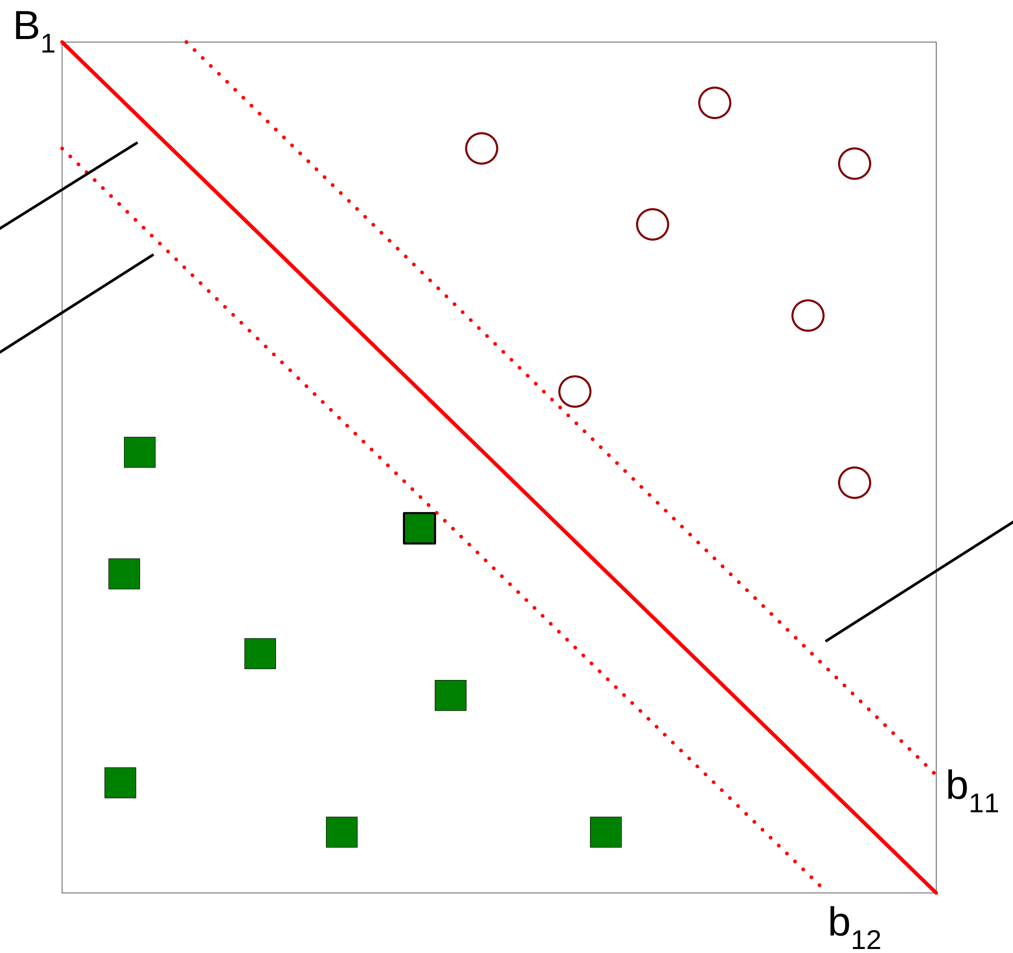
$$\vec{w} \bullet \vec{x} + b = 0$$

$$\vec{w} \bullet \vec{x} + b = -1$$

$$\vec{w} \bullet \vec{x} + b = +1$$

$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}} + \mathrm{b} \geq 1 \\ -1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}} + \mathrm{b} \leq -1 \end{cases}$$

Support Vector Machines - Margin



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

$$\vec{w} \bullet \vec{x} + b = -1$$

$$\vec{w} \bullet \vec{x} + b = +1$$

$$\boldsymbol{w} \cdot (\boldsymbol{x_1} - \boldsymbol{x_2}) = \boldsymbol{2}$$

$$\left| |\boldsymbol{w}| \right| \cdot \boldsymbol{d} = \boldsymbol{2}$$

$$\boldsymbol{d} = \frac{\boldsymbol{2}}{||\boldsymbol{w}||}$$

Learning Linear SVM

- Linear model:

$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}} + \mathrm{b} \geq 1 \\ -1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}} + \mathrm{b} \leq -1 \end{cases}$$

- Learning the model is equivalent to determining the values of $\vec{w}$ and b
  - How to find $\vec{w}$ and b from training data?

Learning Linear SVM

- Linear model:

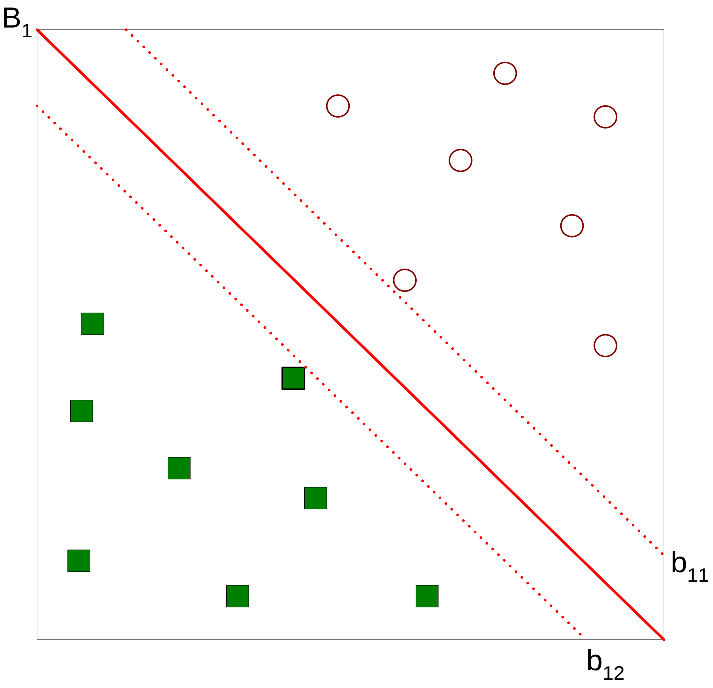
$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x} + b \leq -1 \end{cases}$$

$$\vec{w} \cdot \vec{x} + b = 1$$

$$\vec{w} \cdot \vec{x} + b = -1$$

$$\vec{w} \cdot (x_1 - x_2) = 2$$

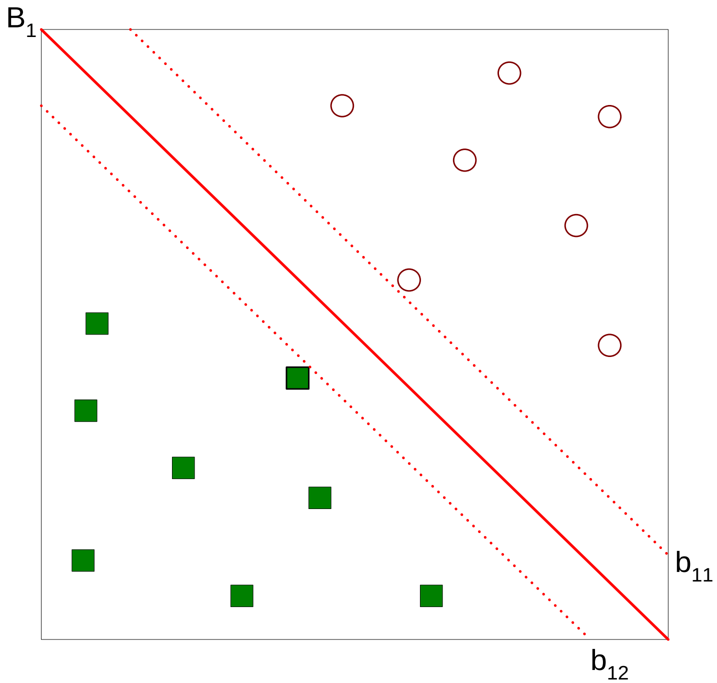
$$\|w\| \times d = 2$$



Learning Linear SVM

- Linear model:

$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x} + b \leq -1 \end{cases}$$



$$\vec{w} \cdot \vec{x} + b = 1$$

$$\vec{w} \cdot \vec{x} + b = -1$$

$$\vec{w} \cdot (x_1 - x_2) = 2$$

$$\|w\| \times d = 2$$

$$\therefore \boldsymbol{d} = \frac{\mathbf{2}}{\|\boldsymbol{w}\|}$$

Learning Linear SVM

- Objective is to maximize: $\text{Margin} = \frac{2}{\|\vec{w}\|}$
  - Which is equivalent to minimizing: $L(\vec{w}) = \frac{\|\vec{w}\|^2}{2}$
  - Subject to the following constraints:

$$y_i = \begin{cases} 1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_\mathrm{i} + \mathrm{b} \geq 1 \\ -1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_\mathrm{i} + \mathrm{b} \leq -1 \end{cases}$$

or

$$y_i(\mathrm{w} \bullet \mathrm{x}_i + b) \geq 1, \qquad i = 1,2,\ldots,N$$

- ◆ This is a constrained optimization problem
  - – Solve it using **Lagrange multiplier method**

Learning Linear SVM

$$\min_{w,\, b} \frac{\|w\|^2}{2}$$

$$subject\ to\quad y_i(w^T x_i + b) \geq 1$$

$$\Downarrow$$

Learning Linear SVM

$$\min_{w,b} \frac{\|w\|^2}{2}$$

$$subject\ to \quad y_i(w^T x_i + b) \geq 1$$

$$\Downarrow$$

$$L_P = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \Lambda_i (y_i(w^T x_i - b) - 1)$$

Learning Linear SVM

$$L_P = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \Lambda_i (y_i (w^T x_i - b) - 1)$$

$$\frac{\partial L_p}{\partial w} = w - \sum_{i=1}^{n} \Lambda_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{n} \Lambda_i y_i x_i$$

$$\frac{\partial L_p}{\partial b} = \sum_{i=1}^{n} \Lambda_i y_i = 0$$

Learning Linear SVM

$$L_P = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \Lambda_i (y_i (w^T x_i - b) - 1)$$

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} \Lambda_i y_i x_i$$

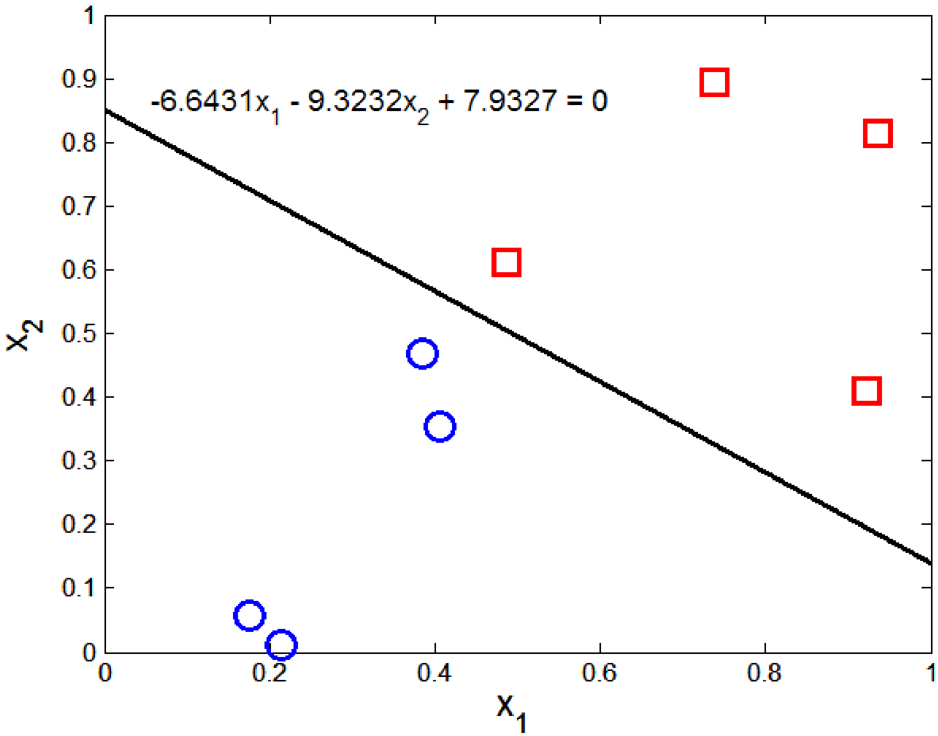
$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow b = \sum_{i=1}^{n} \Lambda_i y_i$$

$$\Lambda_i [y_i (w^T x_i + b) - 1] = 0$$

$$\max_{\Lambda_i} \sum_{i=1}^{n} \Lambda_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \Lambda_i \Lambda_j y_i y_j (x_i \cdot x_j)$$

Subject to $\sum_{i=1}^{n} \Lambda_i y_i = 0, \Lambda_i \geq 0$

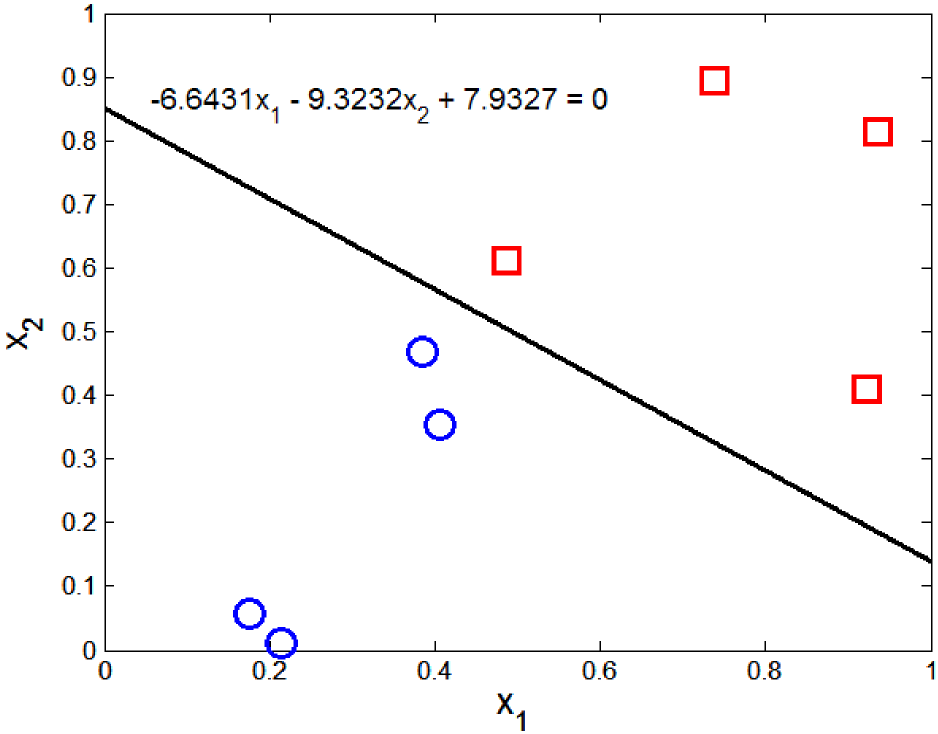
Example of Linear SVM



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

Example of Linear SVM

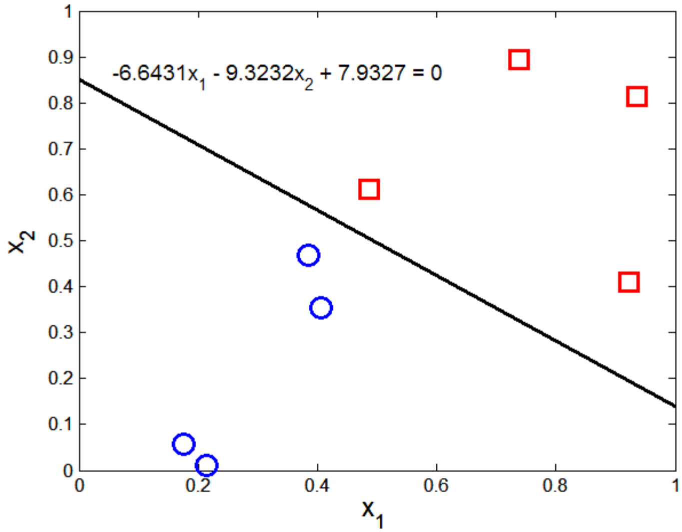


x1	x2	y	λ
0.3858	0.4687	1	65.5261
0.4871	0.611	-1	65.5261
0.9218	0.4103	-1	0
0.7382	0.8936	-1	0
0.1763	0.0579	1	0
0.4057	0.3529	1	0
0.9355	0.8132	-1	0
0.2146	0.0099	1	0

$$w_1 = \sum_i \Lambda_i y_i x_{i1} = 65.5261\ x\ 1\ x\ 0.3858 + 65.5261\ x\ -1\ \ x\ 0.4871 = -6.6431$$

$$w_2 = \sum_i \Lambda_i y_i x_{i2} = 65.5261\ x\ 1\ x\ 0.4687 + 65.5261\ x\ -1\ \ x\ 0.611 = -9.3232$$

Example of Linear SVM



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

$$w_1 = \sum_i \Lambda_i y_i x_{i1} = 65.5261\ x\ 1\ x\ 0.3858 + 65.5261\ x\ -1\ x\ 0.4871 = -6.6431$$

$$w_2 = \sum_i \Lambda_i y_i x_{i2} = 65.5261\ x\ 1\ x\ 0.4687 + 65.5261\ x\ -1\ x\ 0.611 = -9.3232$$

$$b^{(1)} = 1 - w \cdot x_1 = 1 - (-6.64)x\ 0.3858 - (-9.32)0.4687 = 7.93$$

$$b^{(2)} = -1 - w \cdot x_2 = 1 - (-6.64)x\ 0.4871 - (-9.32)0.611 = 7.9289$$

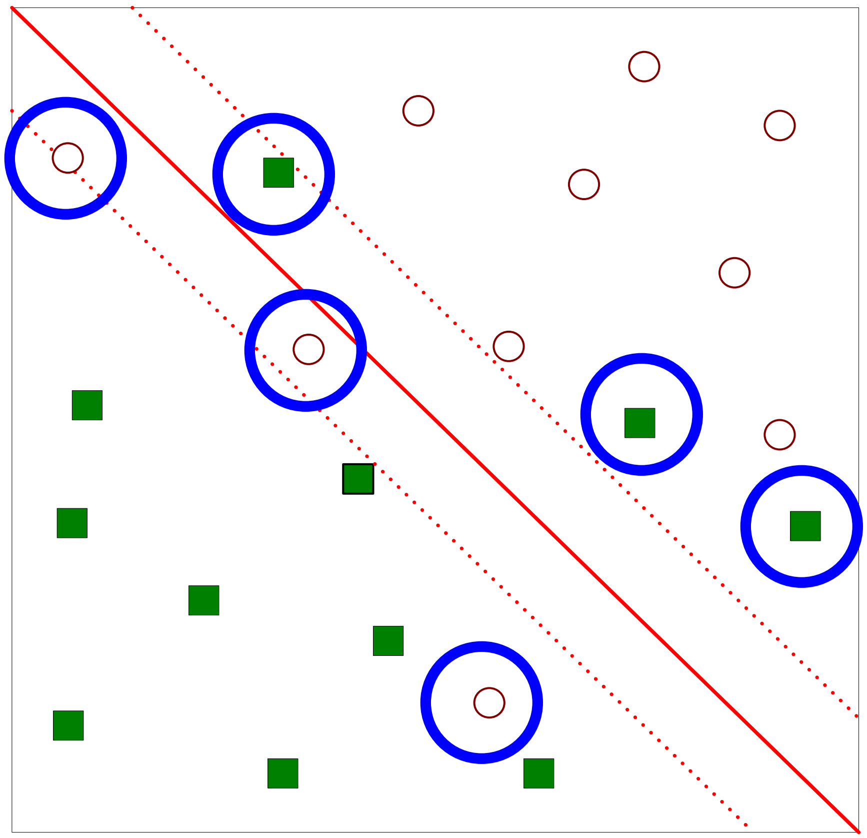
Learning Linear SVM

- 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

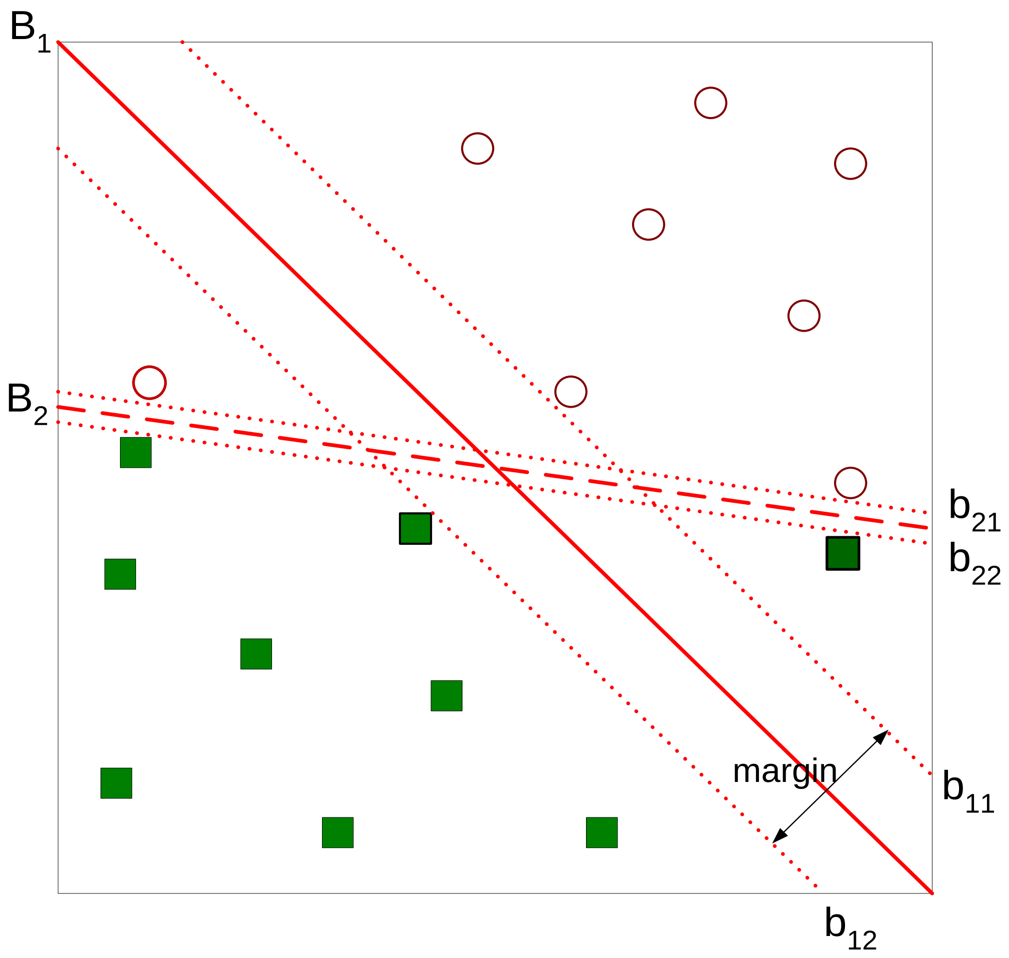
$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_\mathrm{i} + \mathrm{b} \ge 1 \\ -1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_\mathrm{i} + \mathrm{b} \le -1 \end{cases}$$

Support Vector Machines

- What if the problem is not linearly separable?



Support Vector Machines – Soft margin

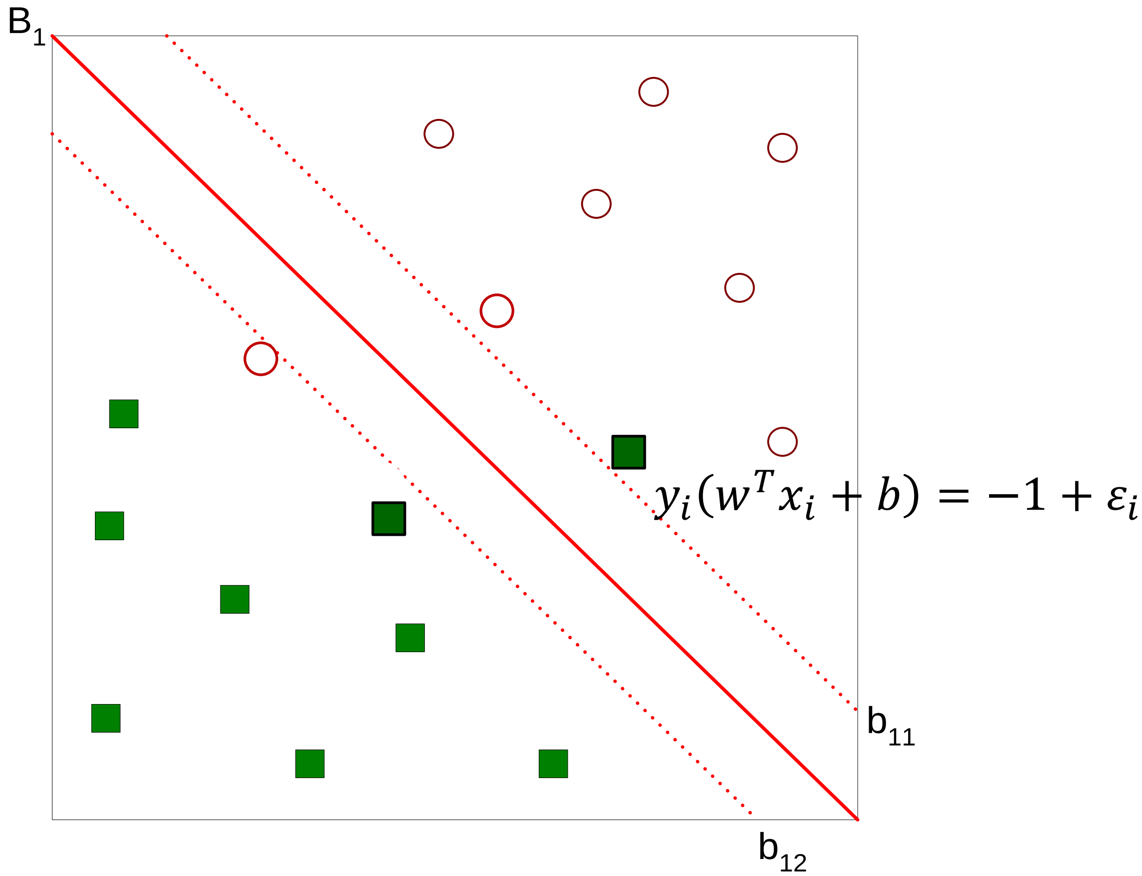


- Find the hyperplane that optimizes both factors

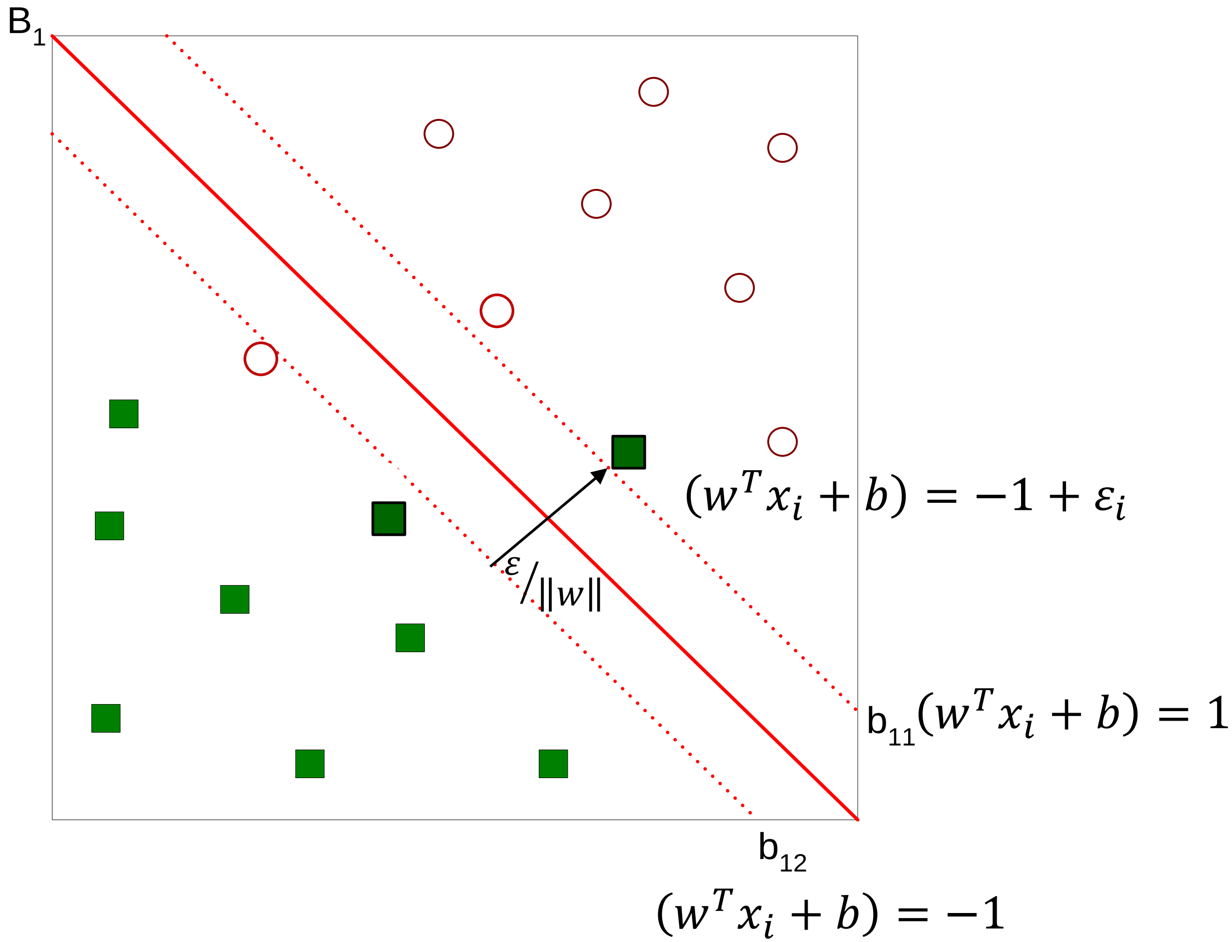
Support Vector Machines – Soft margin

$$y_i(w^T x_i + b) \geq 1 - \varepsilon_i$$

Support Vector Machines – Soft margin



Support Vector Machines – Soft margin



Support Vector Machines – Soft margin

$$\min_{w,b,\varepsilon_i} \frac{\|w\|^2}{2} + C\sum_{i=1}^{n} \varepsilon_i$$

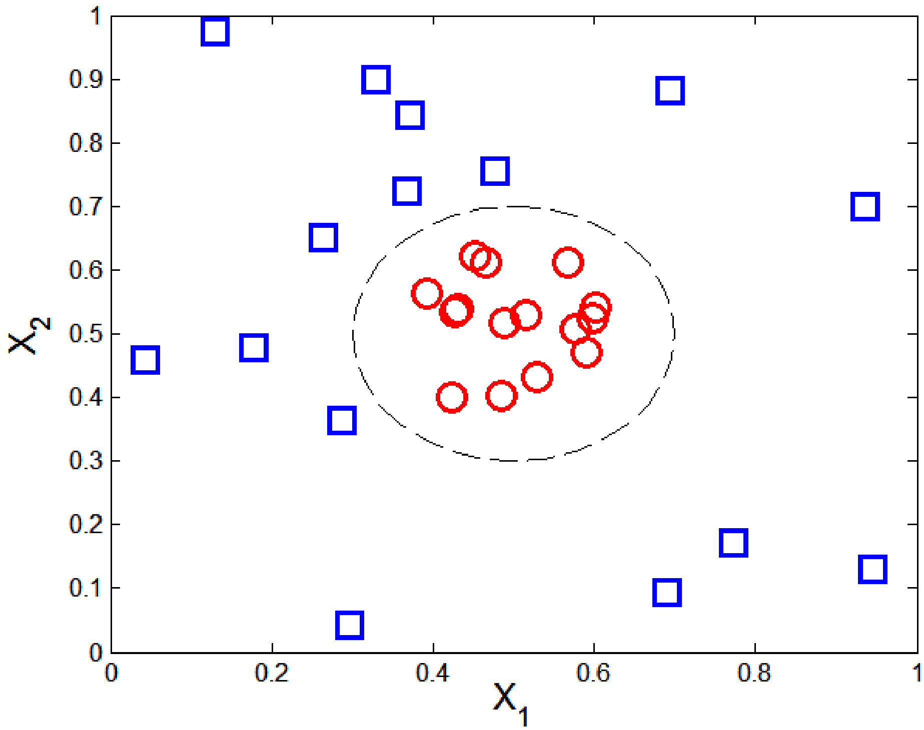
$$\underset{w,b,\varepsilon_i}{subject\ to}\ \ y_i(w^T x_i + b) \geq 1 - \varepsilon_{i,}$$

$$\varepsilon_i \geq 0.$$

NON-LINEAR SVM

Nonlinear Support Vector Machines

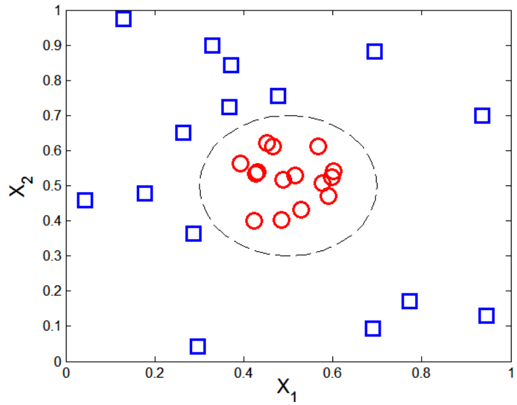
- What if decision boundary is not linear?

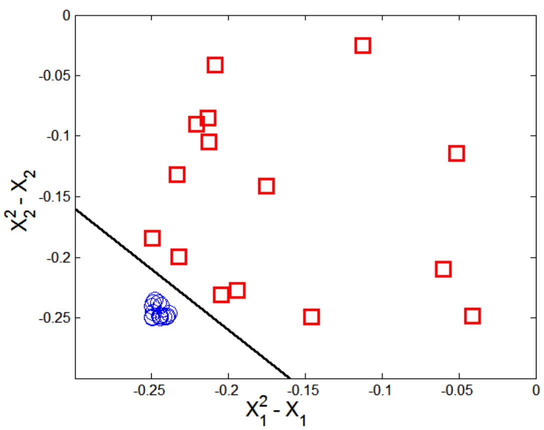


$$y(x_1, x_2) = \begin{cases} 1 & \text{if } \sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2} > 0.2 \\ -1 & \text{otherwise} \end{cases}$$

Attribute transformation

- Transform data into higher dimensional space

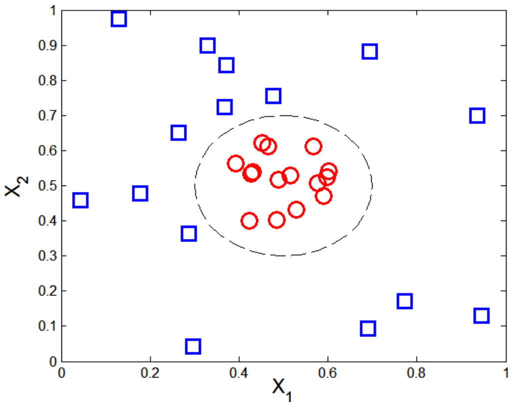


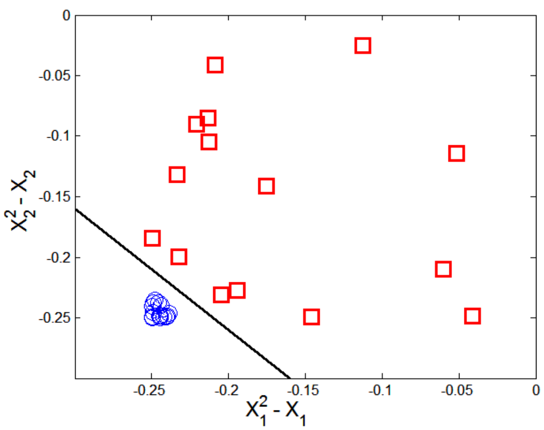


$$y = \begin{cases} 1 & if \sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2} > 0.2 \\ -1 & otherwise \end{cases}$$

Attribute transformation

- Transform data into higher dimensional space



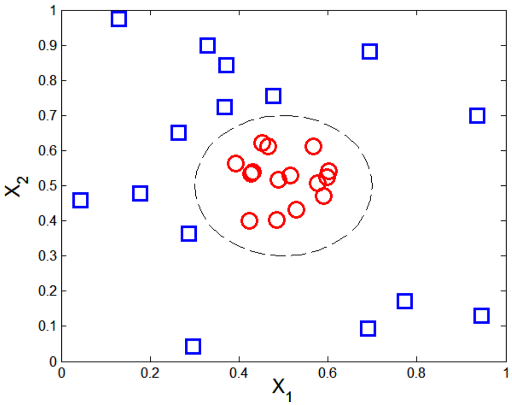


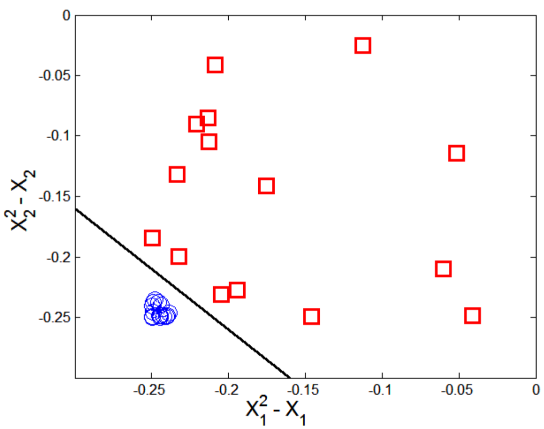
$$y = \begin{cases} 1 & if\ \sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2} > 0.2 \\ -1 & otherwise \end{cases}$$

$$\sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2} = 0.2$$

Attribute transformation

- Transform data into higher dimensional space





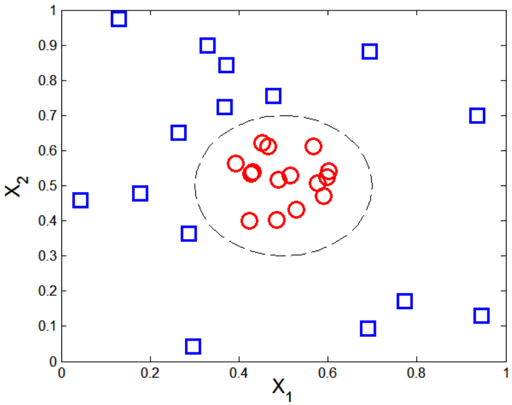
$$y = \begin{cases} 1 & if \sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2} > 0.2 \\ -1 & otherwise \end{cases}$$

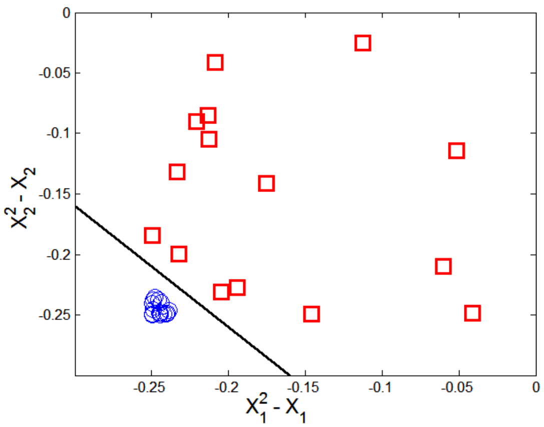
$$\sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2} = 0.2$$

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

Attribute transformation

- Transform data into higher dimensional space





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

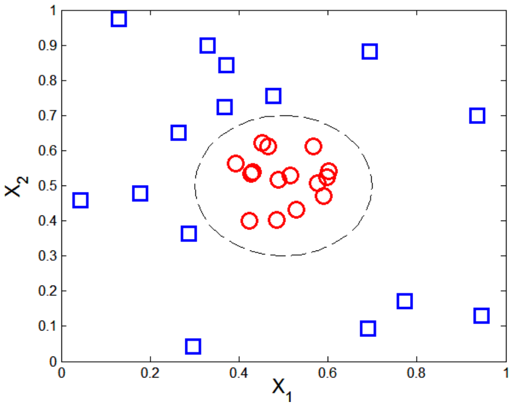
$$\sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2} = 0.2$$

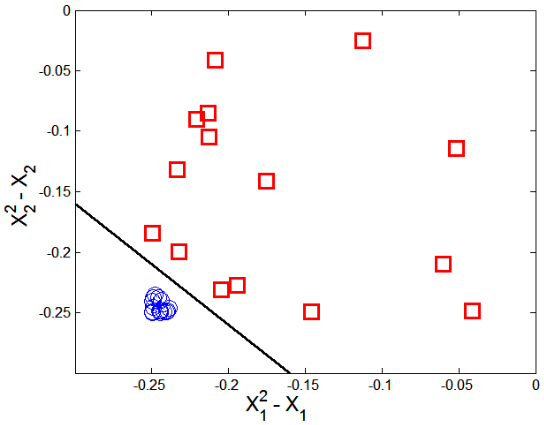
$$x_1^2 - x_1 + x_2^2 - x_2 = -0.46$$

$$\varphi: (x_1, x_2) \rightarrow (x_1^2 - x_1, x_2^2 - x_2)$$

Attribute transformation

- Transform data into higher dimensional space





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

$$\sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2} = 0.2$$

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

$$\varphi: (x_1, x_2) \rightarrow (x_1^2 - x_1, x_2^2 - x_2)$$

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

Learning Nonlinear SVM

- Optimization problem:

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

$$\textit{subject to} \qquad y_i(w \cdot \Phi(x_i) + b) \geq 1, \ \forall\{(x_i, y_i)\}$$

- Which leads to the same set of equations (but involve Φ(x) instead of x)

$$L_D = \sum_{i=1}^{n} \lambda_i - \frac{1}{2} \sum_{i,j} \lambda_i \lambda_j y_i y_j \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$$

$$\mathbf{w} = \sum_i \lambda_i y_i \Phi(\mathbf{x}_i)$$

$$\lambda_i \{ y_i (\sum_j \lambda_j y_j \Phi(\mathbf{x}_j) \cdot \Phi(\mathbf{x}_i) + b) - 1 \} = 0,$$

$$f(\mathbf{z}) = sign(\mathbf{w} \cdot \Phi(\mathbf{z}) + b) = sign(\sum_{i=1}^{n} \lambda_i y_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{z}) + b).$$

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:

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^p$$

$$K(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/(2\sigma^2)}$$

$$K(\mathbf{x}, \mathbf{y}) = \tanh(k\mathbf{x} \cdot \mathbf{y} - \delta)$$

Learning Nonlinear SVM

- 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

- Not all functions can be kernels
  - Must make sure there is a corresponding Φ in some high-dimensional space
  - Mercer's theorem (see textbook)

Characteristics of SVM

- The learning problem is formulated as a convex optimization problem
  - Efficient algorithms are available to find the global minima
  - Many of the other methods use greedy approaches and find locally optimal solutions
  - High computational complexity for building the model

- Robust to noise
- Overfitting is handled by maximizing the margin of the decision boundary,
- SVM can handle irrelevant and redundant better than many other techniques
- The user needs to provide the type of kernel function and cost function
- Difficult to handle missing values

- What about categorical variables?