Kourosh Davoudi
kourosh@uoit.ca

Classification:
Support Vector Machines (SVM)

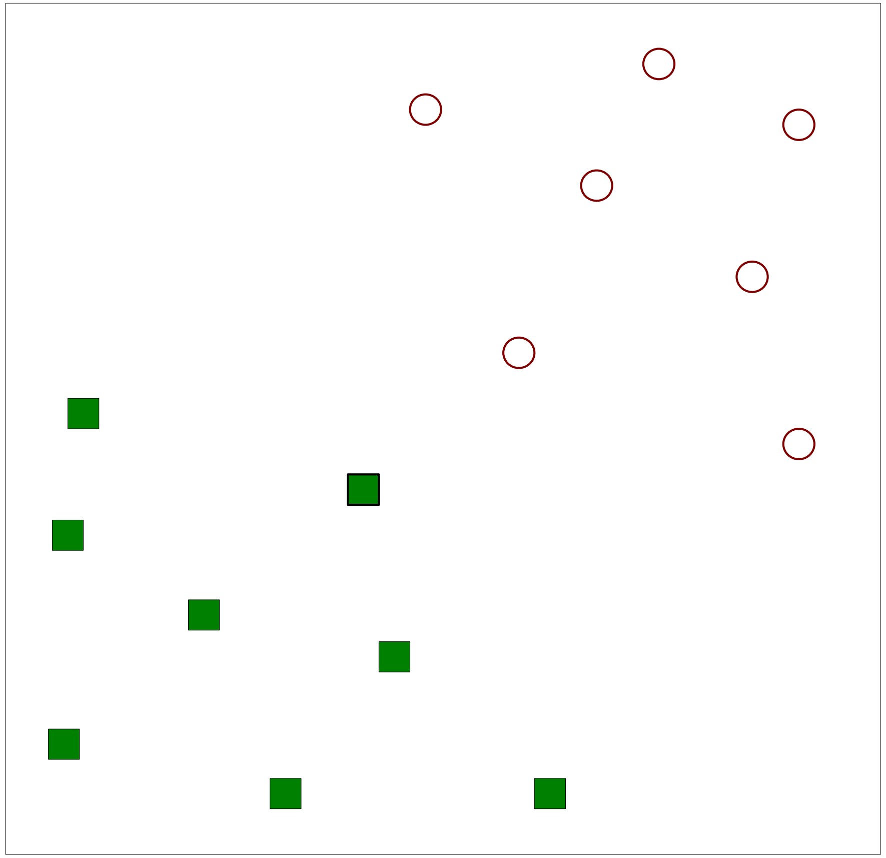
CSCI 4150U: Data Mining

Learning Outcome

- What is the Nearest Neighbor Classifier?
  - Learn the ideas
  - Know the issues
- What is the Naïve Bayes classifier
  - Learn the main ideas
  - Explain are the issues and considerations
- What is Bayesian Belief Network?
- What are the Support Vector Machines?
  - Understand the main ideas
- What are ensemble approaches?
  - Learn the ideas and different approaches

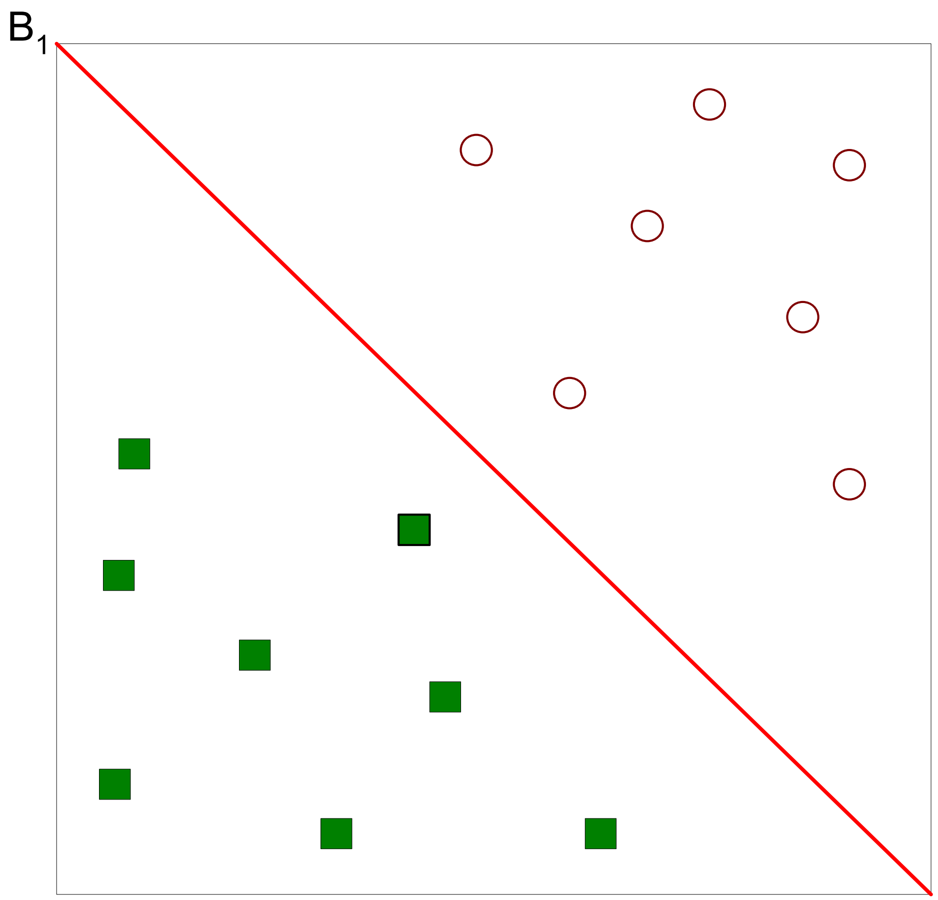
Support Vector Machines

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



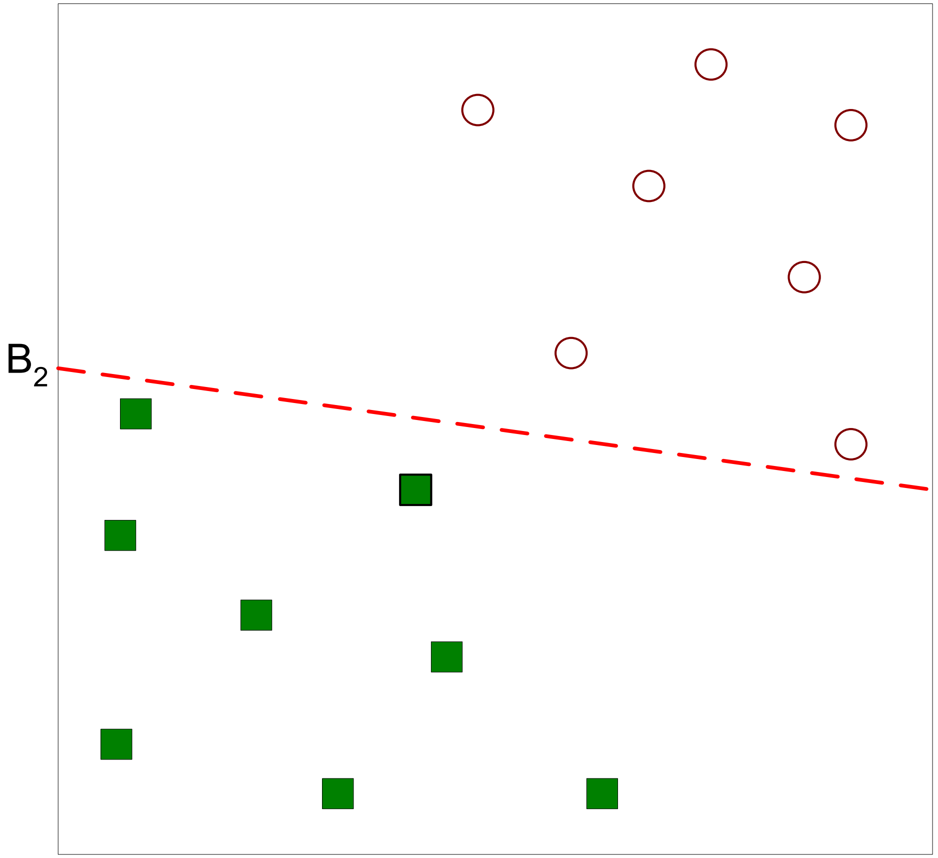
Support Vector Machines

- One Possible Solution



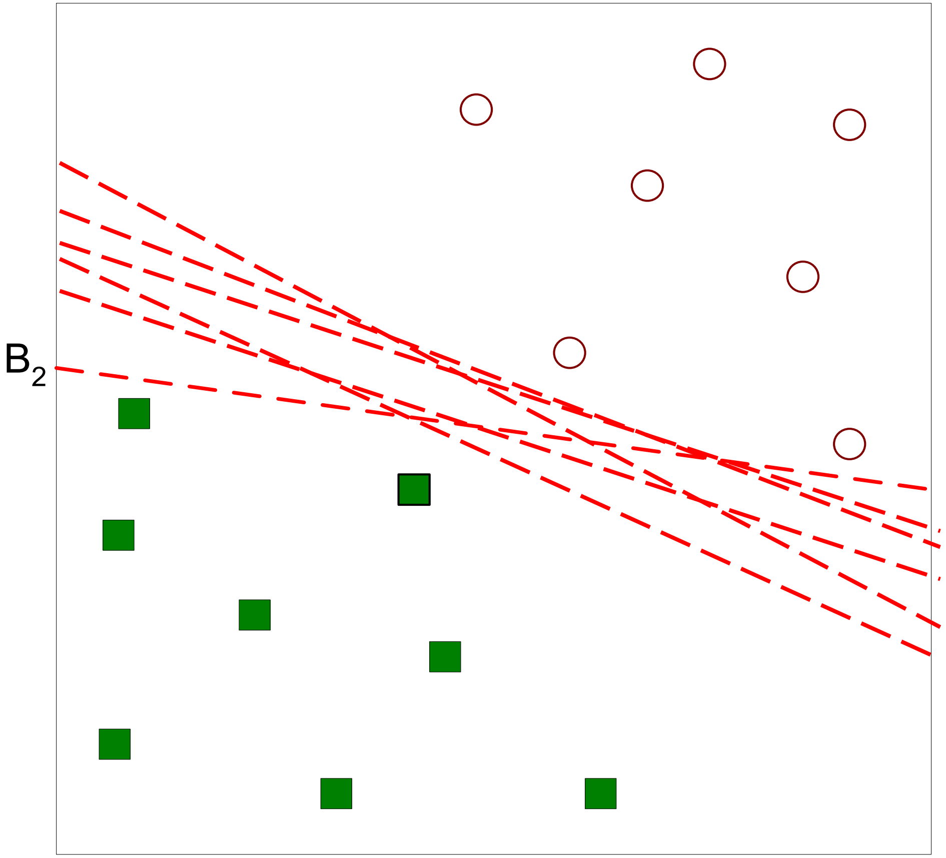
Support Vector Machines

- Another possible solution



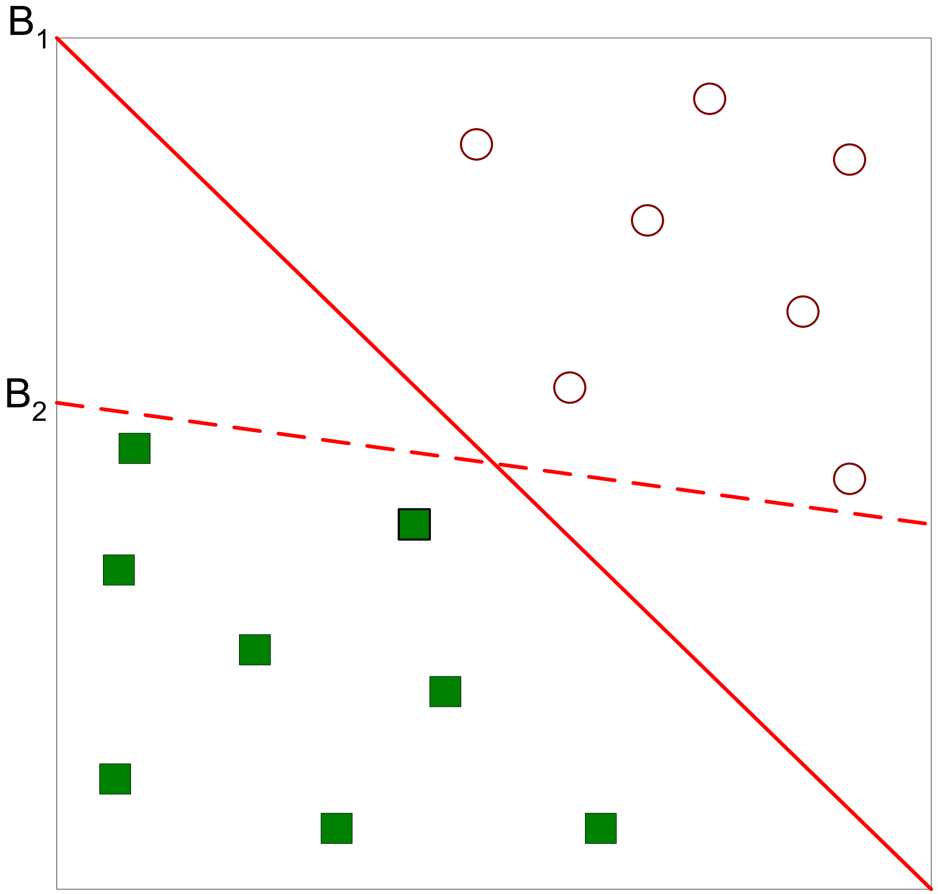
Support Vector Machines

- Other possible solutions



Support Vector Machines

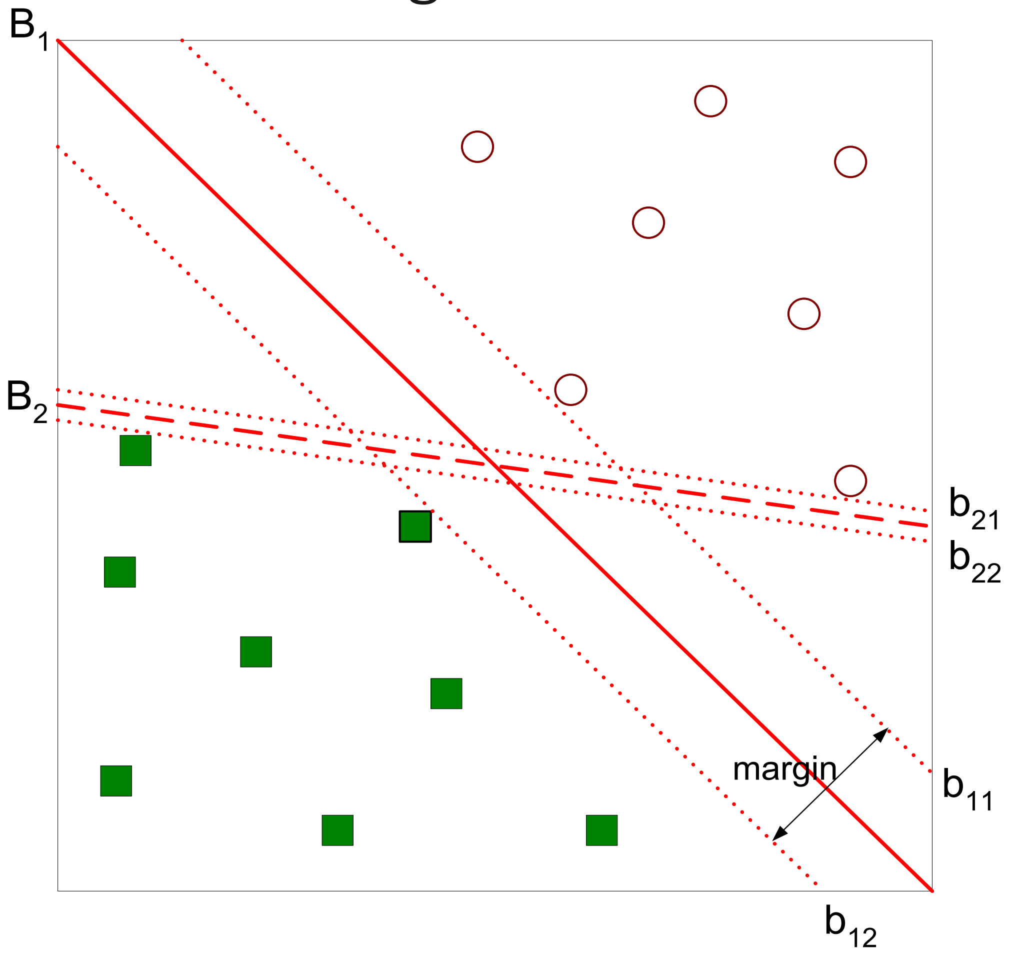
- Which one is better? B1 or B2?



How do you define better?

Support Vector Machines

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

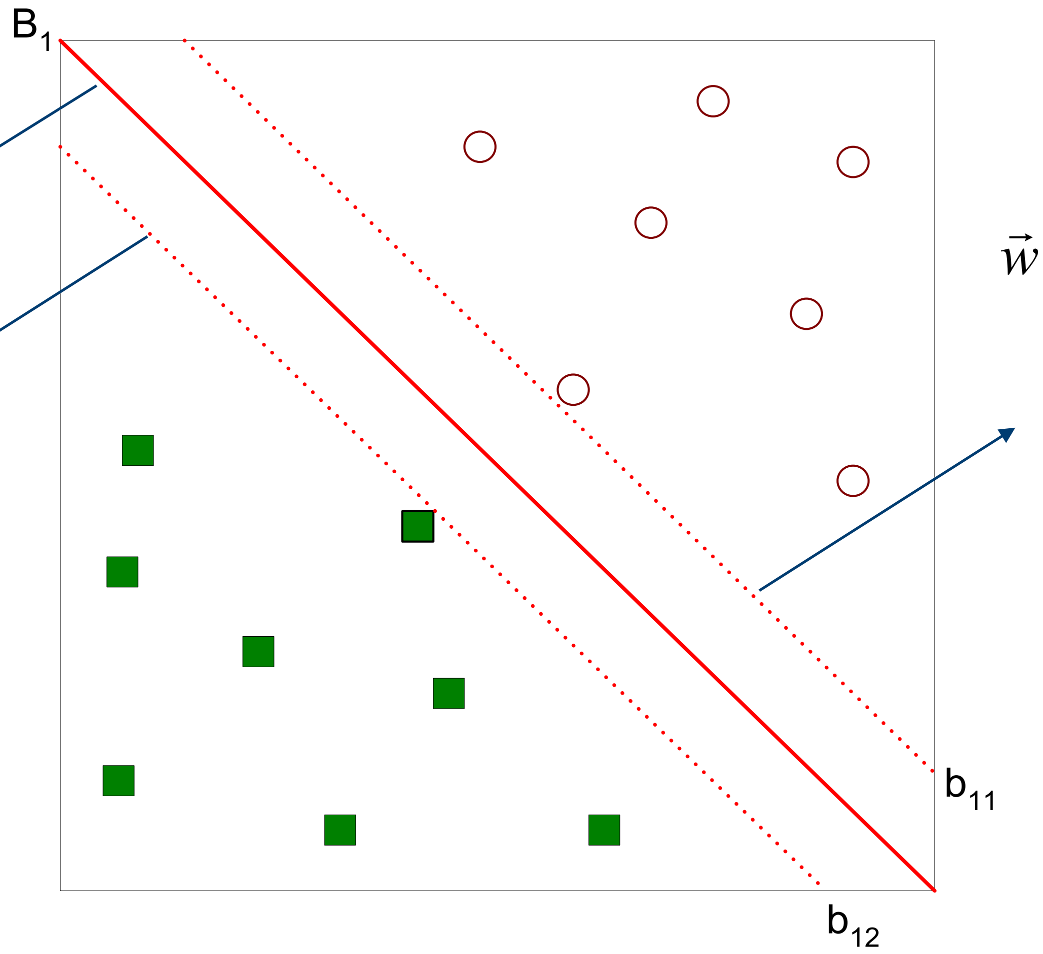


Support Vector Machines

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

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

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



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

$$\text{Margin} = \frac{2}{\|\vec{w}\|}$$

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?

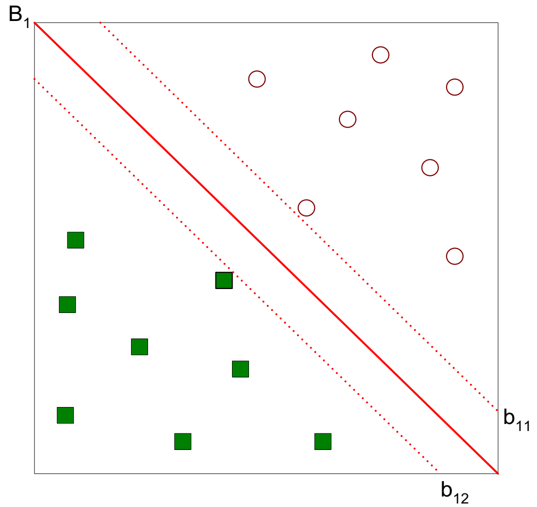
Learning Linear SVM

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

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

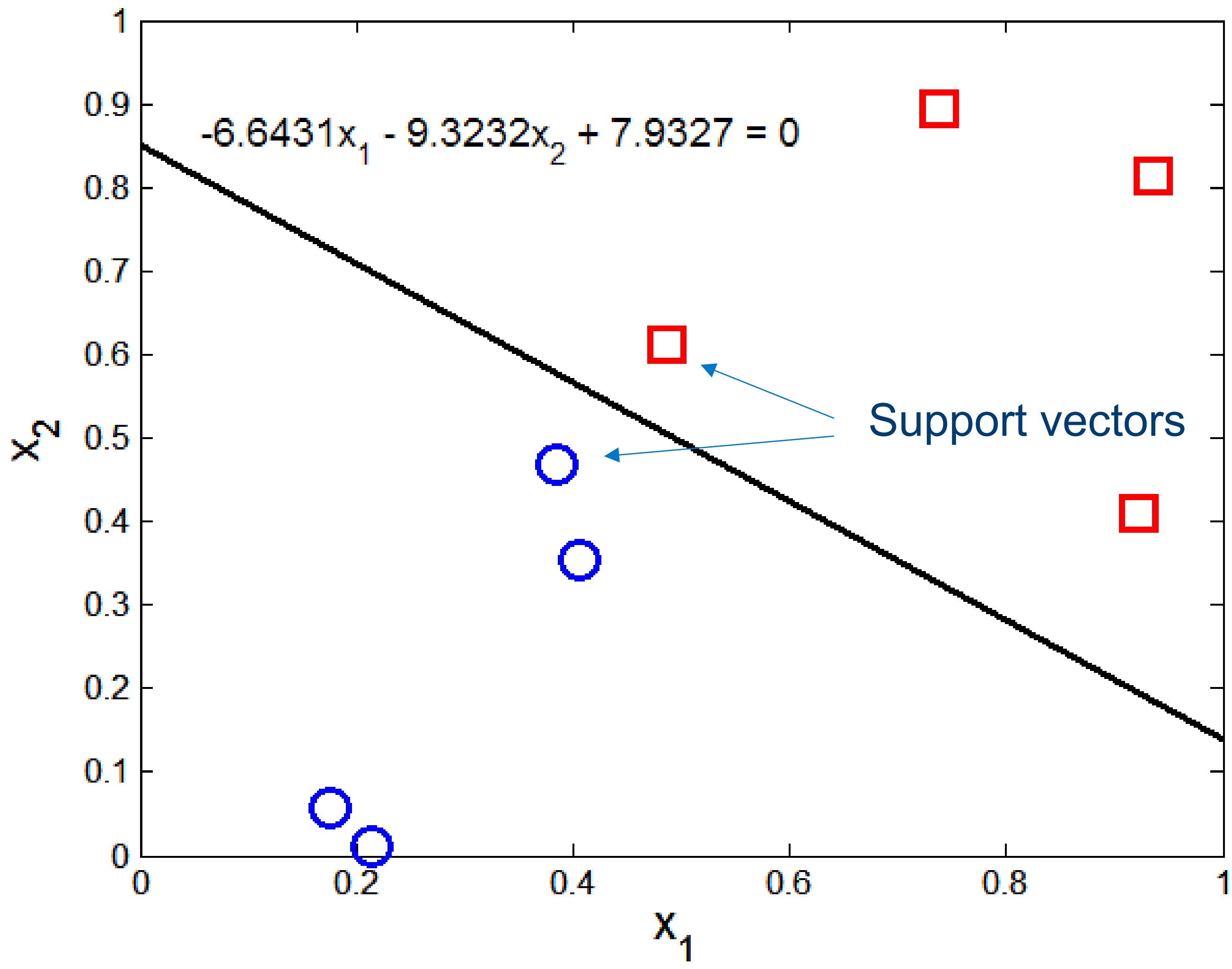
or

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



- This is a constrained optimization problem

Example of Linear SVM



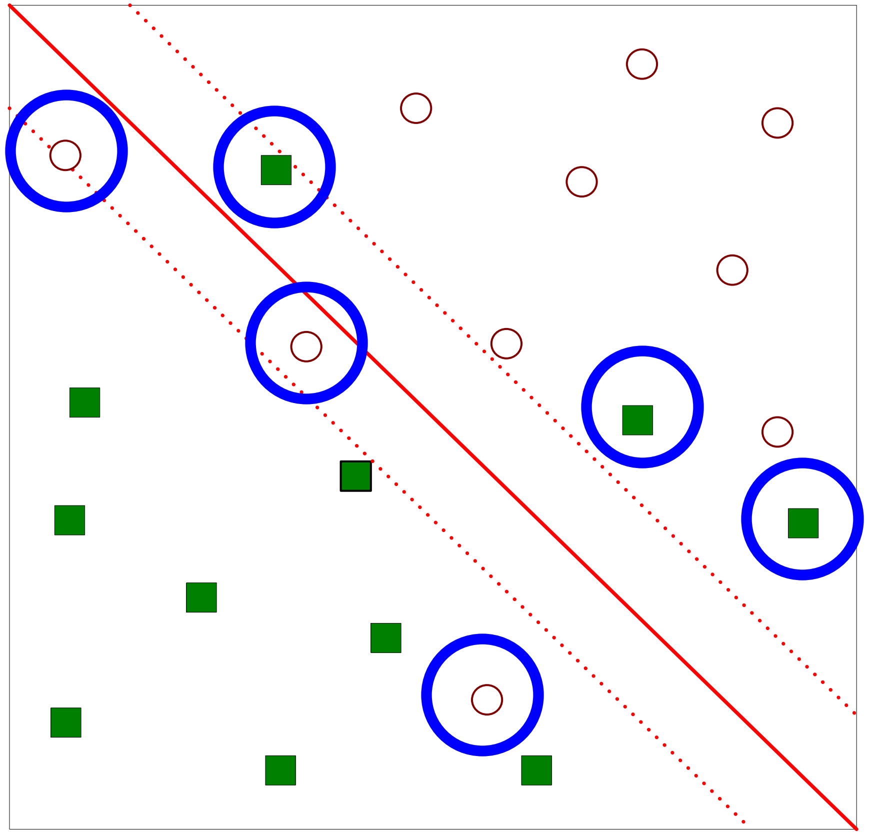
Learning Linear SVM

- 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

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

Support Vector Machines

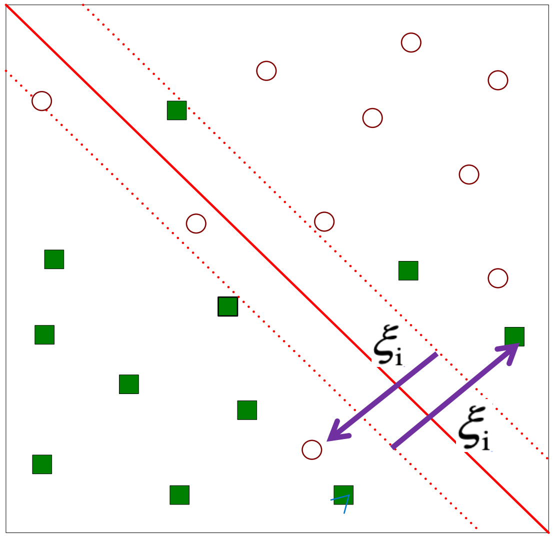
- What if the problem is not linearly separable?



Support Vector Machines

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

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



- What if the problem is not linearly separable?
  - Introduce slack variables
    - Need to minimize:

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

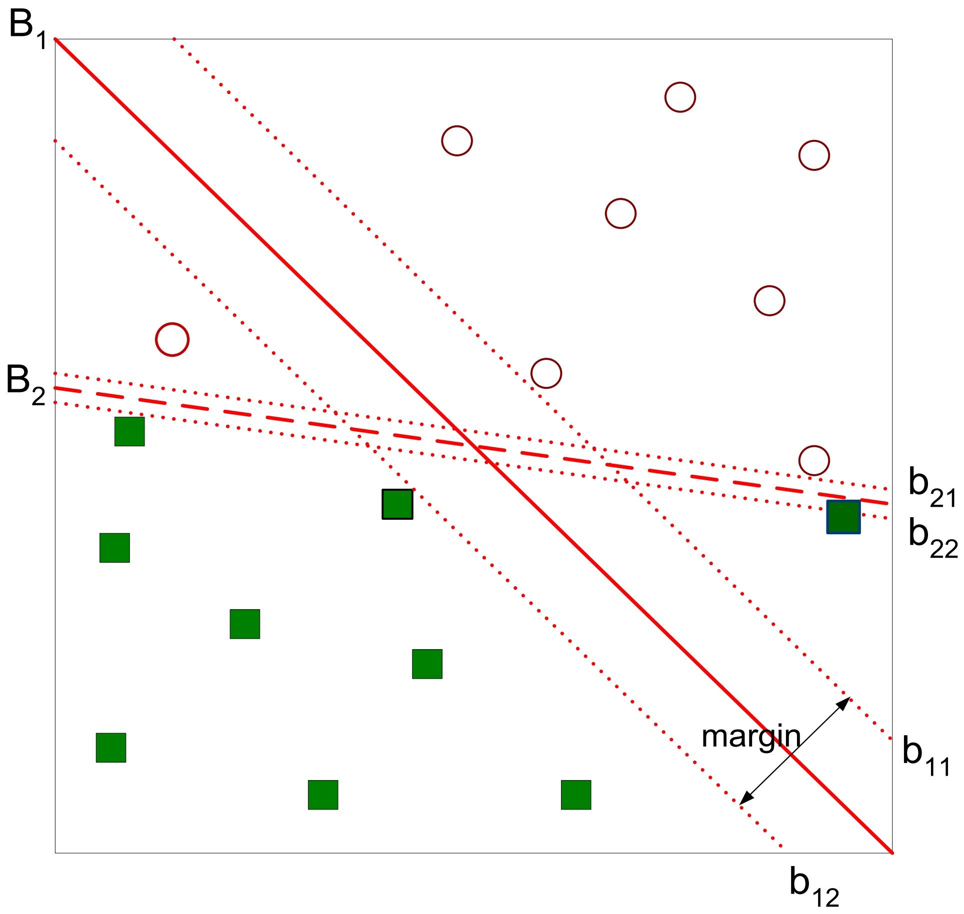
    - Subject to:

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

    - If hyperparameter k is usually is 1 or 2

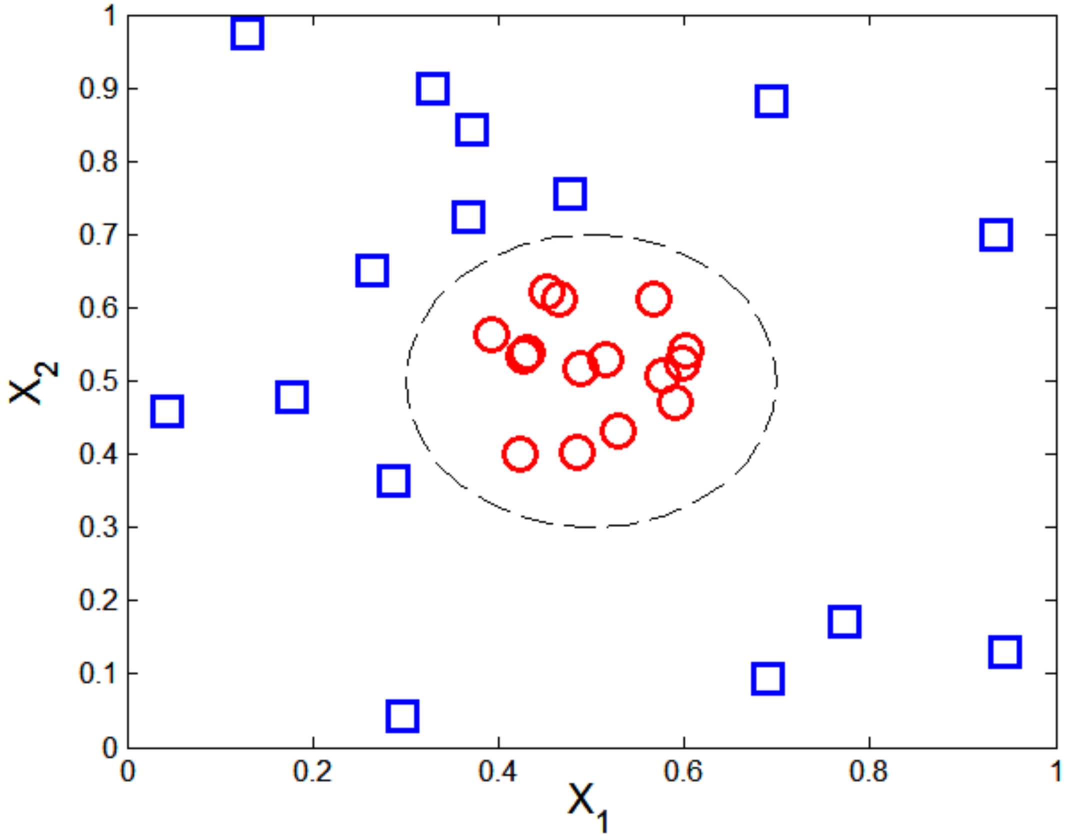
Support Vector Machines

- Find the hyperplane that optimizes both factors



Nonlinear Support Vector Machines

- What if decision boundary is not linear?



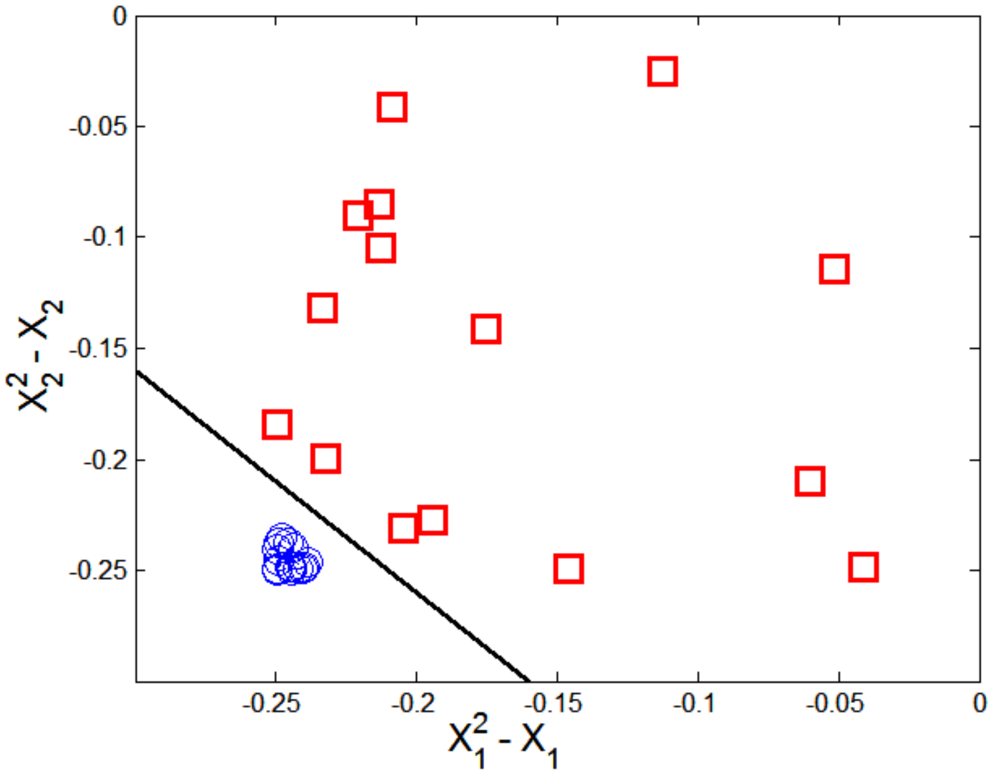
$$y(x_1, x_2) = \begin{cases} 1 & \text{if } \sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2} > 0.2 \\ -1 & \text{otherwise} \end{cases}$$

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

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

Nonlinear Support Vector Machines

- Transform data into higher dimensional space



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

$$\phi(x_1, x_2) = (x_1^2 - x_1, x_2^2 - x_2)$$

Decision boundary:

$$w \,.\, \varphi(x\,) + b = 0$$

Nonlinear Support Vector Machines

- Kernel Trick
  - Instead of transformation function ϕ we specify the kernel function K:

$$K(x_i, x_j) = \phi(x_i).\,\phi(x_j)$$

Why? Because most computations involve dot product:

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

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

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

https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

Characteristics of SVM

- The learning problem is formulated as a convex optimization problem
  - Efficient algorithms are available to find the global minima
- 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
- It is possible to handle instances which are not linearly separable by a technique is called kernel trick.