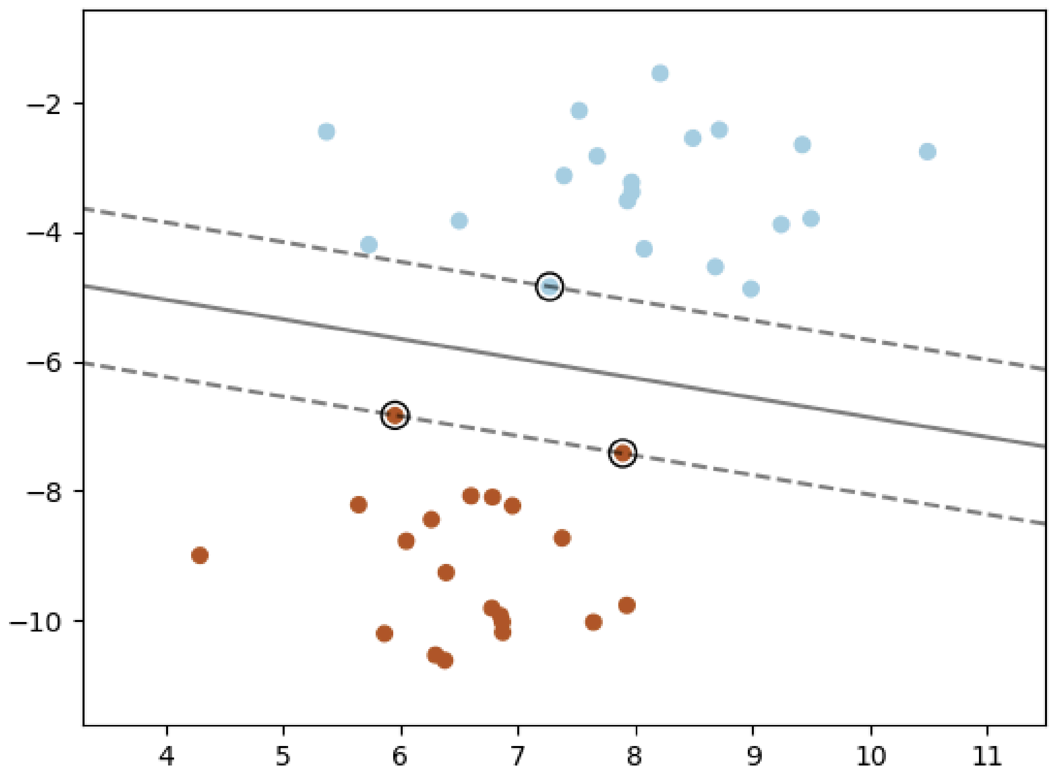
Support Vector Machines (Hard and Soft Margin)

Support Vector Machines (Hard Margin)

- It is a supervised learning algorithm used for classification and outlier detection tasks.
- The idea is to have a decision boundary(**hyperplane**) that separates the data points with a margin
- **Hyperplane**: In SVM, a hyperplane is a decision boundary that separates different classes in the feature space. The goal of SVM is to find the optimal hyperplane that maximizes the margin between the classes.
- **Margin**:It refers to the distance between the supporting hyperplanes and the closest data point from each class.
    - SVM tries to maximize this margin leading to better generalization
- **Support Vector**: These are the data points that are closest to the hyperplane and influence its position and orientation. The optimal hyperplane is determined based on these support vectors.
- This algorithm assumes that the **dataset is linearly separable.**
- For a non linear dataset, we have to apply the **kernel trick** which transforms the features into high dimensional space where it is linearly separable.
  Visualization of a Support Vector Machine
  [



]
Source: Scikit Learn Documentation

Applications :

- **Text Classification**
- **Image Recognition**
- **Bioinformatics**

Mathematical Formulation:

- For a linearly separable dataset, the SVM tries to solve
  - $$\min \quad \frac{1}{2}\|w\|^2$$
  - Subject to :
    $$y_i(w \cdot x_i) \geq 1 \quad \forall i$$
  - Where :
    - w is the weight vector
    - x_i is the i_{th} training data point
    - y_i is the label of x_i . It can be {-1,+1}
    - Assumption: No bias term b included and the decision boundary passes through origin
  - If we want to add a **bias term** b , constraint modifies to:
    $$y_i(w \cdot x_i + b) \geq 1 \quad \forall i$$
    - The decision boundary is $\perp$ to the weight vector w and it equation is given by:
    - $$w^T \cdot x_i = 0$$
- **Lagrange Function**:
  - $$(L)(x, \alpha) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x)$$
  - So , it implies:
    $$(L)(x, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i (1 - (w^T x_i) y_i)$$
- **Dual Problem**:

- - $$\max_{\alpha \geq 0} \quad (\alpha^T - \frac{1}{2}(XY\alpha)^T(XY\alpha))$$
  - X is the design matrix containing the data points as the columns of this matrix
  - Y is the matrix containing the labels
  - α is the matrix containing the coefficients
- Optimal Weight Vector
  - $$w^* = \sum_{i=1}^{n} \alpha_i x_i y_i$$
  - It is a linear combination of the data points x_i whose corresponding α_i >0
  - The data points whose α_i =0 don't contribute to the optimal weight vector w^*
- Only those x_i whose α_i>0 are **Support Vectors**.
- If a **vector lies on the supporting hyperplanes , it doesn't necessarily mean that the vector is a support vector**.

Soft Margin SVM

- **Mathematical Formulation**:
  - **Primal Problem**:

The primal form of the soft margin SVM optimization problem is:

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

Subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i \quad \forall i$$
$$\xi_i \geq 0 \quad \forall i$$

- ξ_i can be thought as the penalty that a data point has to pay if it violates the margin or is misclassified
- ξ_i=0 for a data point that is correctly classified and whose

$$(w^T \cdot x_i)y_i \geq 1$$

- C is a hyperparameter that is used to trade-off between the misclassification and the width of the margin