Support Vector Machines

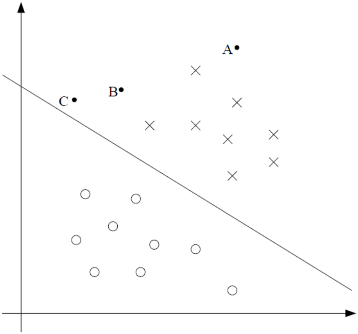
ML Instruction Team, Fall 2022

CE Department
Sharif University of Technology

Ali Sharifi-Zarchi
Behrooz Azarkhalili
Alireza Gargoori Motlagh

Intuition: Margins

- Separating Hyperplane



- Our confidence about the prediction of classes of A, B and C relies on their distance from decision boundary.
- We try to find the optimal hyperplane that separates the classes in the feature space.

Hyperplane

- **Hyperplane**: A hyperplane in *p* dimensions is a flat affine subspace of dimension *p-1*:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0$$

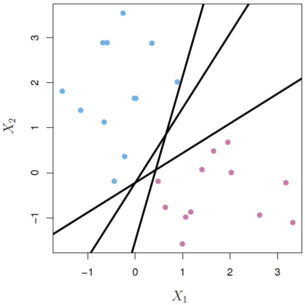
- The vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ is called the normal vector – it points in a direction orthogonal to the surface of the defined hyperplane.

- If $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = \beta^T X + \beta_0$, then $f(X)$ divides the *p*-dimensional feature space into two half-spaces ($f(X) > 0$ for one side and $f(X) < 0$ for the other side).

- So if we code $Y^{(i)} \in \{\pm 1\}$, then $\forall i$

$$Y^{(i)} f(X^{(i)}) > 0$$

Maximal Margin Classifier



- **Maximal(Optimal) Separating Hyperplane**: The separating hyperplane with the biggest margin between the classes.

$$
\begin{aligned}
&\max_{\beta_0,\beta_1,\ldots,\beta_p,\ M} M \\
&\texttt{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 = 1 \\
&\qquad y^{(i)}(\beta^T x^{(i)} + \beta_0) > M \quad \forall i \in \{1, 2, \ldots, N\}
\end{aligned}
\tag{1}
$$

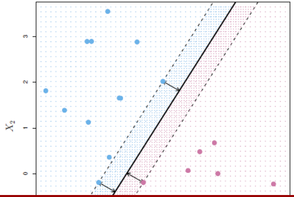
Maximal Margin Classifier: Quadratic Program

- Eq.(1) can be rephrased as a convex quadratic problem and be solved efficiently using QP solvers.
- (Euclidean) distance between two hyperplanes

$$\mathcal{H}_1 = \{x | \beta^T x + \beta_0 = 1\} \qquad \mathcal{H}_2 = \{x | \beta^T x + \beta_0 = -1\}$$

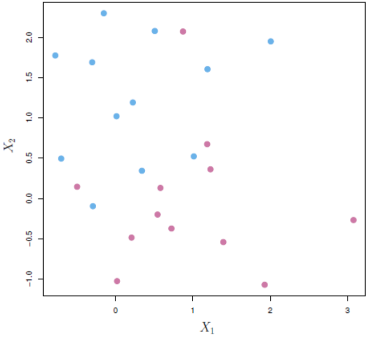
is **dist**$(\mathcal{H}_1, \mathcal{H}_2) = 2/||\beta||_2$

$$\begin{aligned} \min_{\beta, \beta_0} \quad & \frac{1}{2}||\beta||_2^2 \\ \text{s.t.} \quad & y^{(i)}(\beta^T x^{(i)} + \beta_0) \geq 1 \quad \forall i \in 1, 2, \ldots, N \end{aligned} \tag{2}$$



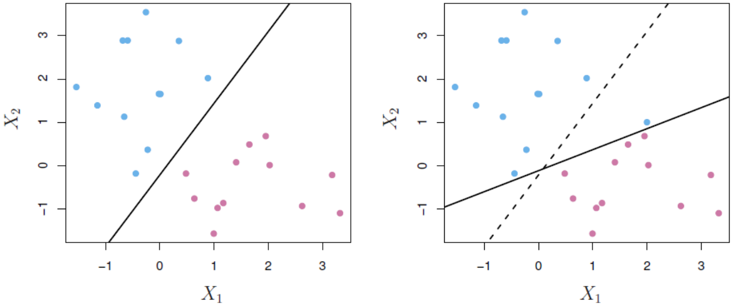
Non-linear Separable Data

- In most cases however, the data are not linearly separable unless $N < p$.



Noisy Data

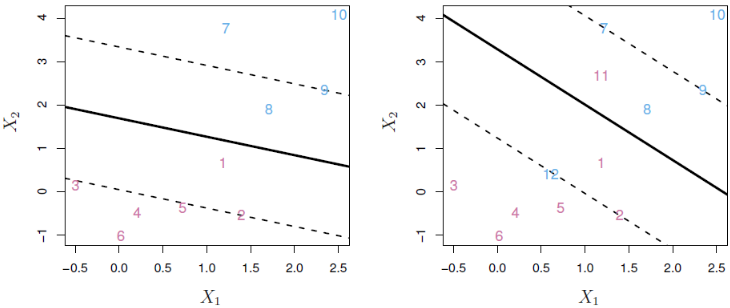
- Sometimes the data are linearly separable, but noisy. This can lead to a poor solution for the maximal margin classifier. Also, hard-margin classifier is sensitive to outliers.



- The *support vector classifier* maximizes a *soft* margin.

Support Vector Classifier(Soft Margin Classifier)

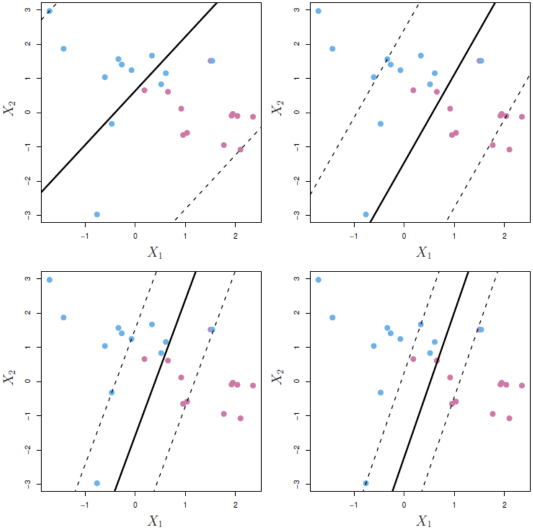
- Allowing some samples to violate the margin, with *slack variables*, in a controlled manner.



$$\begin{aligned} \min_{\beta,\beta_0,\xi} \quad & \frac{1}{2}||\beta||_2^2 + C\sum_{i=1}^{N}\xi_i \\ \texttt{s.t.} \quad & y^{(i)}(\beta^T x^{(i)} + \beta_0) \geq 1 - \xi_i \\ & \xi_i \geq 0 \quad \forall i \in \{1, 2, \ldots, N\} \end{aligned} \tag{3}$$

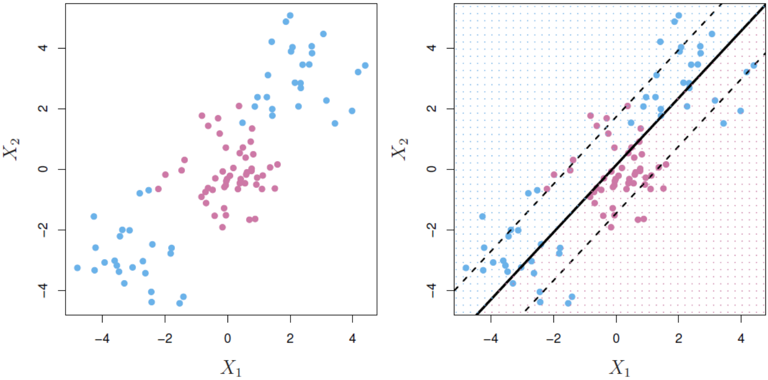
Effect of Regularization Parameter

- C is a regularization parameter that controls the bias-variance trade-off of the support vector classifier.



The Need for Non-Linear Boundary

- Linear boundary can fail in many cases, regardless of the value of C.



Final Notes

Thank You!

Any Question?