KAZAKH-BRITISH TECHNICAL UNIVERSITY

Introduction to Machine Learning
Week 11

Olivier JAYLET

School of Information Technology and Engineering

Support Vector Machines (SVM)

- Mathematical Strength: Regarded as one of the most mathematically robust statistical learning methods.
- Comparison with Other Classifiers:
    - Competes well with other statistical learning classifiers.
    - Kernels are $N \times N$, leading to scalability issues in large datasets.

Support Vector Classifier

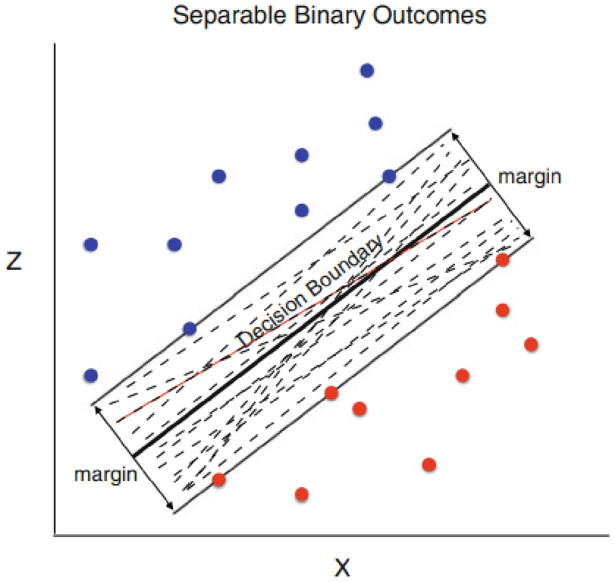
- Binary Response Variable:
  - The response (target variable) is coded as 1 or -1.
  - The function $f(x)$ is written linearly as:

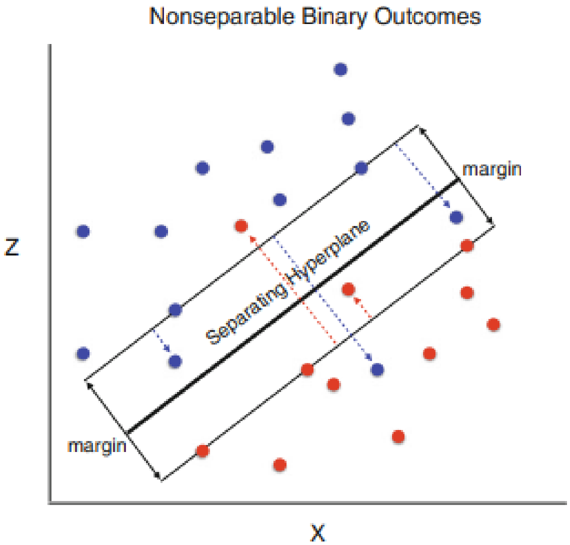
$$f(x) = \beta_0 + x^T \beta \tag{1}$$

- Key Points:
  - $f(x)$ gives a numeric output for prediction.
  - Observations:
    - If $f(x) > 0$, assign label 1.
    - If $f(x) < 0$, assign label -1.
  - This approach does not fit logits, probabilities, or proportions (unlike logistic regression).

Support Vector Classifier (SVC)

- A Support Vector Classifier (SVC) finds an optimal hyperplane or decision boundary that maximally separates data points of different classes in a feature space
- Objective:
  - Minimize mismatch between labels predicted by $f(x)$ and the actual binary labels (1 or -1).
  - Ensure the method generalizes accurately to new data as well as current data.


Separable Binary Outcomes
margin
Decision Boundary
margin
Z
X


Nonseparable Binary Outcomes
margin
Separating Hyperplane
margin
Z
X

Classification Using a Linear Hyperplane

- The separating hyperplane is defined using the linear combination:

$$f(x) = \beta_0 + x^T \beta = 0 \tag{2}$$

- Classification rule:

$$G(x) = \text{sign}(\beta_0 + x^T \beta) \tag{3}$$

  - If $f(x) > 0$: Classified as $+1$.
  - If $f(x) < 0$: Classified as -1.
- Observations:
  - The value 0 lies halfway between -1 and 1.
  - $f(x)$ can be used to compute the signed distance of a point from the separating hyperplane.
  - This helps determine whether a point is correctly classified and, if not, how far it is from the correct side.

Maximizing the Margin for Linear Classification

- For the separable case, the objective is to find β and β_0 to maximize the margin.
- Let M represent the distance between the separating hyperplane and the margin boundary. The optimization problem can be written as:

$$\max_{\beta,\beta_0} M \quad \text{subject to } \|\beta\| = 1 \tag{4}$$

- The constraints ensure that every correctly classified observation satisfies:

$$y_i(\beta_0 + x_i^T\beta) \geq M, \quad i = 1, \ldots, N \tag{5}$$

- Notes:
  - For ease of computation, the regression coefficients (β) are standardized to have a unit length (i.e., $\|\beta\| = 1$).

Alternative Formulation for Margin Maximization

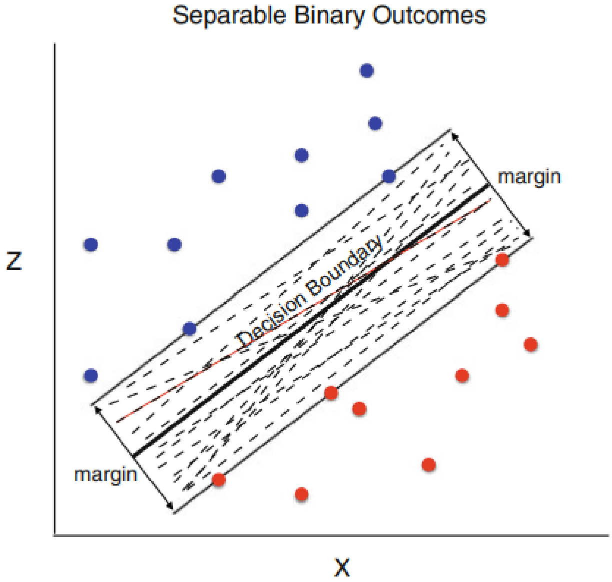
- Equivalent Formulation: Instead of maximizing the margin M, an equivalent and more mathematically convenient approach is used:

$$\min_{\beta,\beta_0} \|\beta\| \tag{6}$$

subject to:

$$y_i(\beta_0 + x_i^T \beta) \geq 1, \quad i = 1, \ldots, N \tag{7}$$

- Notes:
  - Since $M = \frac{1}{\|\beta\|}$, minimizing $\|\beta\|$ is equivalent to maximizing M.
  - This approach simplifies the optimization problem.
  - The alternative formulation does not affect the underlying optimization problem or the final solution.


Separable Binary Outcomes
margin
Decision Boundary
Z
margin
X

Nonseparable Case: Introducing Slack Variables

- For the nonseparable case, minor violations of the margin (buffer zone) may occur.
- Introduce slack variables $\xi = (\xi_1, \xi_2, \dots, \xi_N)$ with $\xi_i \geq 0$:
  - $\xi_i = 0$: Observation is on the correct side of the margin.
  - $\xi_i > 0$: Observation crosses into or through the margin.
- The constraint is revised as:

$$y_i(\beta_0 + x_i^T \beta) \geq M(1 - \xi_i), \quad \forall\, i \tag{8}$$

- Additional constraints:

$$\sum_{i=1}^{N} \xi_i \leq W, \quad \xi_i \geq 0 \tag{9}$$

- W: Quantifies the tolerance for misclassifications.

Canonical Formulation for Nonseparable Case

- An equivalent and commonly used formulation for the nonseparable case is:

$$\min_{\beta,\beta_0} \|\beta\| \tag{10}$$

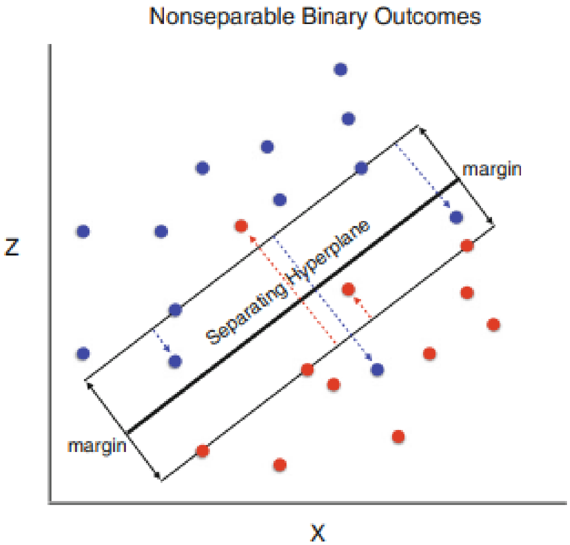
subject to:

$$y_i(\beta_0 + x_i^T \beta) \geq 1 - \xi_i, \quad i = 1, \ldots, N \tag{11}$$

with:

$$\xi_i \geq 0, \quad \sum_{i=1}^{N} \xi_i \leq W \tag{12}$$

- Notes:
  - For larger ξ_i, points are allowed to violate the margin more, relaxing the linear constraint.


Nonseparable Binary Outcomes
margin
Separating Hyperplane
margin
Z
X