MAXIMUM MARGIN CLASSIFIER

- Multiple ways of separating training data
- Which one is the best?
  - Smallest generalization error
- SVM uses margin
  - The smallest distance between the decision boundary and any of the samples
  - Find the decision boundary that maximizes the margin
PROBLEM STATEMENT

- Decision boundary
  \[ w^T x + b = 0 \]
- Margin
  \[ (w^T x_i + b) y_i / \|w\| > c / \|w\| \]
- The optimization problem
  \[ \max_{w,b} \frac{1}{\|w\|} \]
  \[ \text{s.t.} \]
  \[ y_i (w^T x_i + b) \geq 1, \quad \forall i \]
- At least 2 active constraints when the margin is maximized
Lagrangian Duality

- The dual problem

\[
\max_{\alpha} J(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i^T x_j)
\]

s.t. \( \alpha_i \geq 0, \quad i = 1, \ldots, k \)

\[
\sum_{i=1}^{m} \alpha_i y_i = 0.
\]

- How did we get it?
  - Write the Lagrangian
    \[
    \mathcal{L}(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{m} \alpha_i [y_i (w^T x_i + b) - 1]
    \]
    s.t. \( \alpha_i \geq 0, \quad i = 1, \ldots, k \)
  - Take the derivative
    \[
    \nabla_w \mathcal{L}(w, b, \alpha) = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0,
    \]
    \[
    \nabla_b \mathcal{L}(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i y_i = 0,
    \]
  - Substitute in the Lagrangian
Properties of Lagrangian Duality

- Weak duality always holds:
  - \( d^* \leq p^* \)

- Strong duality holds
  - If there exists a saddle point
  - Or
    - If the primal problem is convex,
    - And some constraint qualification holds (e.g. Slater’s condition)

- If there exists some saddle point, then the saddle point satisfies the KKT conditions

- If \( w^*, \alpha^*, \) and \( \beta^* \) satisfy KKT, it is the solution to the primal and the dual problems
GEOMETRIC INTERPRETATION

\[ \lambda u + t = g(\lambda) \]

\[ \lambda^* u + t = g(\lambda^*) \]

\[ \lambda_1 u + t = g(\lambda_1) \]
Support Vectors

- After training, we only need the support vectors

\[ \alpha_i g_i(w) = 0, \quad i = 1, \ldots, m \]

Call the training data points whose \( \alpha_i \)'s are nonzero the support vectors (SV)
SOFT MARGIN HYPERPLANE

- Allow error in classification
- Penalize the error that increases with the distance from the boundary

\[
\min_{w,b} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i
\]

s.t. \( y_i (w^T x_i + b) \geq 1 - \xi_i, \quad \forall i \)
\( \xi_i \geq 0, \quad \forall i \)

- For misclassified point \( x_i \), \( \xi_i > 1 \)
- For correctly classified point that lies inside the margin \( x_i \), \( 0 < \xi_i \leq 1 \)
- For misclassified points \( x_i \) that lies outside the margin \( \xi_i = 0 \)
**LOSS FUNCTION**

- **Hard margin**
  - Infinite error for misclassified data point
  - Zero for correctly classified data point

- **Soft Margin**
  - Zero for data points at the right side of the margin
  - Increases linearly as it crosses the boundary
  - Sensitive to outliers
THE KERNEL TRICK

- Maps data to high dimensional space
  \[ \phi(x_i) \]
- But still maintains low computation complexity
  \[ K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \]
- Handles non-linearly separable data
  \[
  \max_\alpha \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \\
  \text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, k \\
  \sum_{i=1}^{m} \alpha_i y_i = 0.
  \]
- Symmetry
- Positive-semidefinite