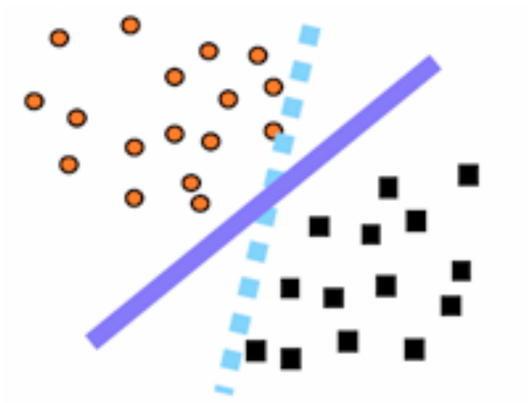
Support Vector Machines

Nan Li

2011.11.21

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

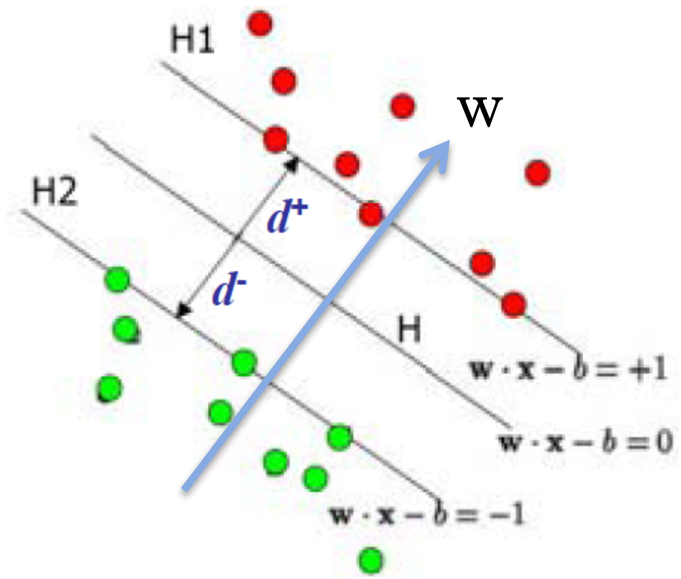
$$w^T x + b = 0$$

- Margin

$$(w^T x_i + b) y_i / \|w\| > c / \|\mathbf{w}\|$$

- The optimization problem

$$\max_{w,b} \quad \frac{1}{\|w\|}$$
$$\text{s.t} \quad y_i (w^T x_i + b) \geq 1, \quad \forall i$$



- At least 2 active constraints when the margin is maximized

LAGRANGIAN DUALITY

- The dual problem

$$\max_{\alpha} \mathcal{J}(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j)$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, k$$

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

- How did we get it?
  - Write the Lagrangian

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{i=1}^{m} \alpha_i \left[y_i (w^T x_i + b) - 1 \right]$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, k$$

  - Take the derivative

$$\nabla_w \mathcal{L}(w, b, \alpha) = w - \sum_{i=1}^{m} \alpha_i y_i x_i = 0,$$

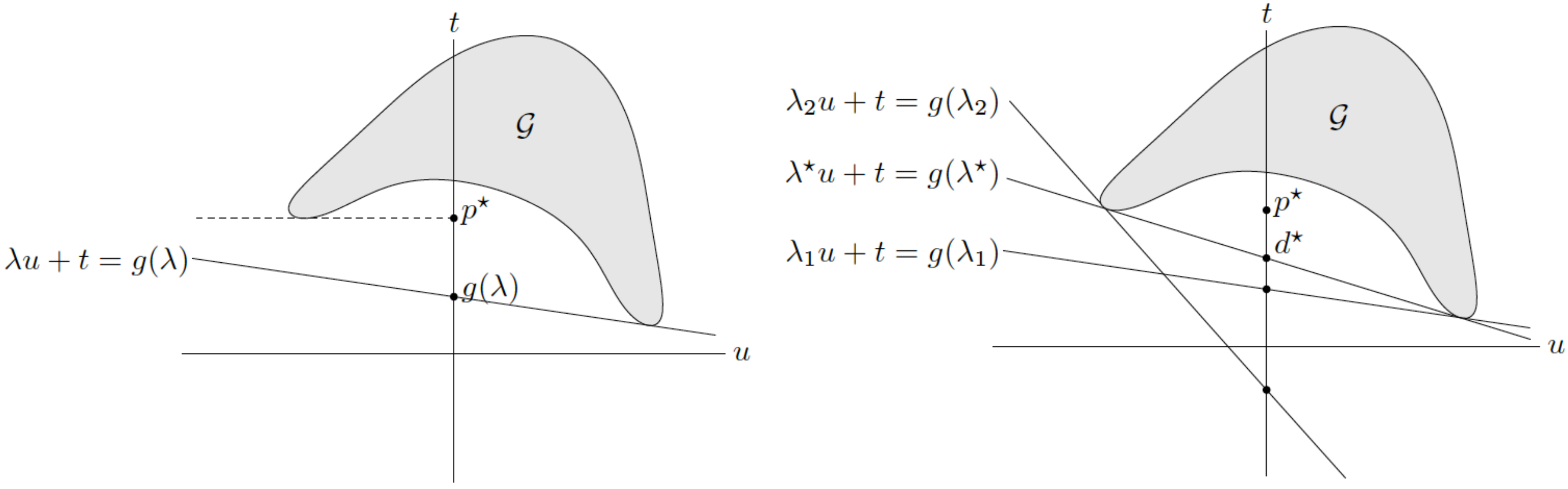
$$\nabla_b \mathcal{L}(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i y_i = 0,$$

  - Substitute in the Lagrangian

Properties of Lagrangian Duality

- Weak duality always holds:
  - $d^* <= p^*$
- 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^*, α^*, and β^* satisfy KKT, it is the solution to the primal and the dual problems

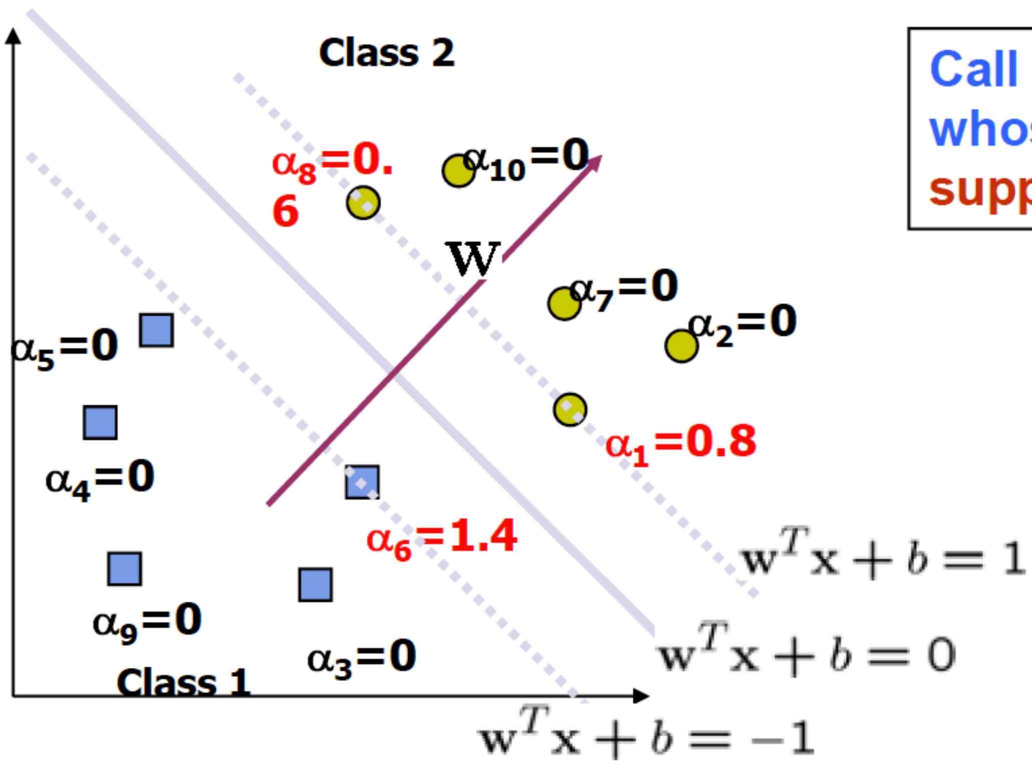
Geometric Interpretation



SUPPORT VECTORS

- After training, we only need the support vectors

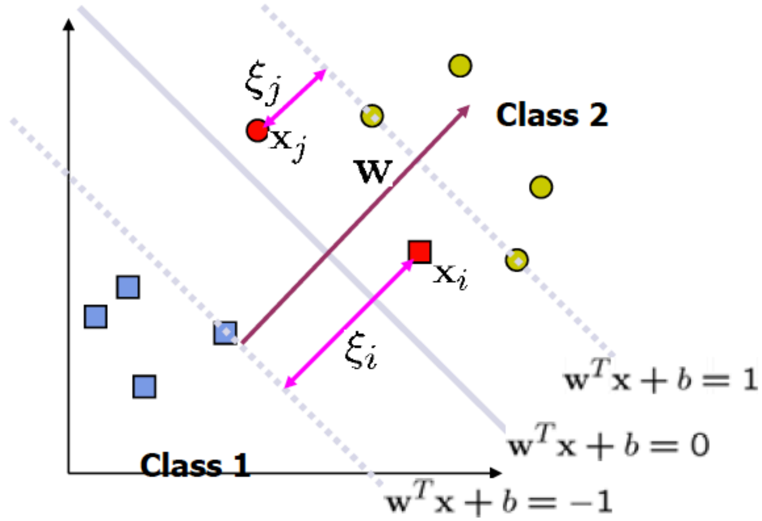
$$\alpha_i g_i(w) = 0, \quad i = 1, \ldots, m$$



Call the training data points whose α_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



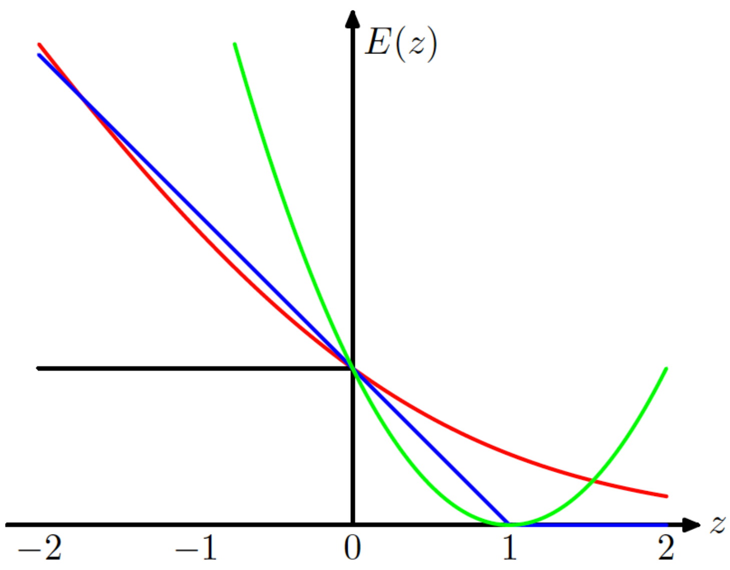
$$\min_{w,b} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i$$

$$\text{s.t} \quad \begin{array}{l} y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i \\ \xi_i \geq 0, \quad \forall i \end{array}$$

- For misclassified point x_i, $\xi_i > 1$
- For correctly classified point that lies inside the margin x_i, $0 < \xi_i <= 1$
- 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

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$$

- Handles non-linearly separable data

$$\max_\alpha \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, k$$

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

- Symmetry
- Positive-semidefinite