Foundations of Machine Learning

Module 5:

Part C: Support Vector Machine: Dual

Sudeshna Sarkar
IIT Kharagpur

Solving the Optimization Problem

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$$

- Optimization problem with convex quadratic objectives and linear constraints
- Can be solved using QP.
- Lagrange duality to get the optimization problem's dual form,
  - Allow us to use kernels to get optimal margin classifiers to work efficiently in very high dimensional spaces.
  - Allow us to derive an efficient algorithm for solving the above optimization problem that will typically do much better than generic QP software.

Lagrangian Duality in brief

The Primal Problem

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \le 0, \quad i = 1, \dots, k \\ & h_i(w) = 0, \quad i = 1, \dots, l \end{aligned}$$

The generalized Lagrangian:

$$L(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

the α's ($\alpha_i \geq 0$) and β's are called the Lagrange multipliers

Lemma:

$$\max_{\alpha, \beta, \alpha_i \ge 0} L(w, \alpha, \beta) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise} \end{cases}$$

A re-written Primal:

$$\min_w \max_{\alpha, \beta, \alpha_i \ge 0} L(w, \alpha, \beta)$$

Lagrangian Duality, cont.

The Primal Problem: $p^* = \min_w \max_{\alpha,\beta,\alpha_i \geq 0} \mathrm{L}(w,\alpha,\beta)$

The Dual Problem: $d^* = \max_{\alpha,\beta,\alpha_i \geq 0} \min_w \mathrm{L}(w,\alpha,\beta)$

Theorem (weak duality):

$$d^* = \max_{\alpha,\beta,\alpha_i \geq 0} \min_w \mathrm{L}(w,\alpha,\beta) \leq \min_w \max_{\alpha,\beta,\alpha_i \geq 0} \mathrm{L}(w,\alpha,\beta) = p^*$$

Theorem (strong duality):

Iff there exist a saddle point of $L(w,\alpha,\beta)$, we have

$$d^* = p^*$$

The KKT conditions

If there exists some saddle point of *L*, then it satisfies the following "Karush-Kuhn-Tucker" (KKT) conditions:

$$\frac{\partial}{\partial w_i} \mathrm{L}(w, \alpha, \beta) = 0, \quad i = 1, \ldots, k$$

$$\frac{\partial}{\partial \beta_i} \mathrm{L}(w, \alpha, \beta) = 0, \quad i = 1, \ldots, l$$

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

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

$$\alpha_i \geq 0, \quad i = 1, \ldots, m$$

Theorem: If *w**, *a** and *b** satisfy the KKT condition, then it is also a solution to the primal and the dual problems.

Support Vectors

- Only a few α_i's can be nonzero
- Call the training data points whose α_i's are nonzero the support vectors

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

If $\alpha_i > 0$ then $g(w) = 0$

Solving the Optimization Problem

Quadratic programming with linear constraints

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$$

Lagrangian Function

$$\text{minimize} \quad L_p(\mathbf{w}, b, \alpha_i) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 \right)$$

$$\text{s.t.} \quad \alpha_i \geq 0$$

Solving the Optimization Problem

$$\text{minimize} \quad L_p(\mathbf{w}, b, \alpha_i) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left(y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1 \right)$$

$$\text{s.t.} \qquad \alpha_i \geq 0$$

Minimize wrt w and b for fixed α

$$\frac{\partial L_p}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial L_p}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$\mathrm{L}_p(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j) - b \sum_{i=1}^{m} \alpha_i y_i$$

$$\mathrm{L}_p(w, b, \alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i^T \mathbf{x}_j)$$

The Dual problem

Now we have the following dual opt problem:

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

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

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

This is a **quadratic programming** problem.

– A global maximum of α_i can always be found.

Support vector machines

- Once we have the Lagrange multipliers $\{\alpha_j\}$ we can reconstruct the parameter vector w as a weighted combination of the training examples:

$$w = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \qquad w = \sum_{i \in SV} \alpha_i y_i \mathbf{x}_i$$

- For testing with a new data ***z***
  - Compute

$$w^T z + b = \sum_{i \in SV} \alpha_i y_i \left(\mathbf{x}_i^T z \right) + b$$

  and classify ***z*** as class 1 if the sum is positive, and class 2 otherwise

Note: *w* need not be formed explicitly

Solving the Optimization Problem

- The discriminant function is:

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = \sum_{i \in \mathrm{SV}} \alpha_i \mathbf{x}_i^T \mathbf{x} + b$$

- It relies on a *dot product* between the test point $\boldsymbol{x}$ and the support vectors $\boldsymbol{x}_i$
- Solving the optimization problem involved computing the dot products $\boldsymbol{x}_i^T \boldsymbol{x}_j$ between all pairs of training points
- The optimal $\boldsymbol{w}$ is a linear combination of a small number of data points.