Optimization: SVMs

Vibhav Gogate

THE UNIVERSITY OF TEXAS AT DALLAS
Erik Jonsson School of Engineering and Computer Science

Machine Learning: Optimization

- Most machine learning algorithms involve some form of optimization
- We have covered a generic optimization algorithm: **gradient descent/ascent**
- As presented, it requires:
  - Unconstrained Objective function
  - Differentiable Objective function

Example:

$$\text{Rep:}\ \ f(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i + w_0 \ \text{ where }\ \mathbf{x} = (x_1, \ldots, x_d)$$

$$\text{Obj:}\ \ g(w_0, \ldots, w_d) = \sum_{k=1}^{n} \left(y^{(k)} - f\left(\mathbf{x}^{(k)}\right)\right)^2 + \lambda \sum_{i=1}^{d} w_i^2$$

Unconstrained versus Constrained Optimization

Unconstrained

$$\text{Minimize: } g(w_1, \ldots, w_d)$$

Read as: find values of $w_1, \ldots, w_d$ such that the function g is minimized.

Constrained

$$\begin{aligned} \text{Minimize:} \quad & g_0(w_1, \ldots, w_d) \\ \text{Subject to:} \quad & g_i(w_1, \ldots, w_d) \leq 0 \text{ for } i = 1 \text{ to } n \end{aligned}$$

Note: This formulation is general because equality constraints $g_i(w_1, \ldots, w_d) = 0$ can be written as two constraints $g_i(w_1, \ldots, w_d) \leq 0$ and $-g_i(w_1, \ldots, w_d) \leq 0$.

Some Terminology

- $w_1, \ldots, w_d$ are the optimization variables or **parameters** and g is the **objective function**.
- $g_i(w_1, \ldots, w_d) \leq 0$, $i = 1$ to n are called the **constraints**.
- The set of points satisfying the constraints is called the **feasible set**.
- A point $w_1, \ldots, w_d$ in the feasible set is called a **feasible point**.
- The **optimal value** p^* of the problem is defined as

  $$p^* = \min \{g_0(w_1, \ldots, w_d) \mid (w_1, \ldots, w_d) \text{ satisfies all constraints}\}$$

  (technically min should be inf).
- $(w_1^*, \ldots, w_d^*)$ is the **optimal point** if it is feasible and $g_0(w_1^*, \ldots, w_d^*) = p^*$

Lagrangian Formulation

Constrained: For simplicity of notation, let $w = (w_1, \ldots, w_d)$.
Our optimization problem can be stated as:

$$\begin{aligned} \text{Minimize:} \quad & g_0(w) \\ \text{Subject to:} \quad & g_i(w) \leq 0 \text{ for } i = 1 \text{ to } n \end{aligned}$$

The Lagrangian for the optimization problem is

$$L(w, \alpha) = g_0(w) + \sum_{i=1}^{n} \alpha_i g_i(w)$$

where α_i's are called Lagrange multipliers (also called the dual variables).

Why Lagrangian?

Maximum over Lagrangian is equivalent to the original problem!

$$\begin{aligned}\max_{\alpha \geq 0} L(w, \alpha) &= \max_{\alpha \geq 0} \left(g_0(w) + \sum_{i=1}^{n} \alpha_i g_i(w) \right) \\ &= \begin{cases} g_0(w) & \text{if } g_i(w) \leq 0 \text{ for all } i \\ \infty & \text{otherwise.} \end{cases}\end{aligned}$$

- Let us say a constraint, $g_i(w)$ is violated. Then $g_i(w) > 0$. Thus, the max value of the term in brackets is reached when $\alpha_i = \infty$ which means that the max over the sum in the brackets will be ∞.
- If all the constraints are satisfied then $g_i(w) \leq 0$, which means to maximize, we should have $\alpha_i = 0$ (or $g_i(w) = 0$). Then $\sum_i \alpha_i g_i(w)$ will be zero. Thus, the max over the sum in the brackets $= g_0(w)$.

Therefore, the optimal value of the optimization problem is

$$p^* = \min_{w} \max_{\alpha \geq 0} L(w, \alpha)$$

Primal versus Dual Formulation

- Primal problem:

$$p^* = \min_{w} \max_{\alpha \geq 0} L(w, \alpha)$$

- Dual problem (Flip max and min):

$$d^* = \max_{\alpha \geq 0} \min_{w} L(w, \alpha)$$

- Verify that min of max is always greater than or equal to max of min. Therefore, $p^* \geq d^*$.

For any point (w', α'), we have:
$\min_w L(w, \alpha') \leq L(w', \alpha') \leq \max_\alpha L(w', \alpha)$

Therefore, $\max_\alpha \min_w L(w, \alpha) \leq \min_w \max_\alpha L(w, \alpha)$
`https://en.wikipedia.org/wiki/Max-min_inequality`

Primal and Dual Solution

- Primal problem:

$$p^* = \min_{w} \max_{\alpha \geq 0} L(w, \alpha)$$

- Dual problem (Flip max and min):

$$d^* = \max_{\alpha \geq 0} \min_{w} L(w, \alpha)$$

- When we have a Convex objective function and affine constraints $p^* = d^*$. Thus, we can solve the dual in lieu of the primal problem.
- Why use the dual? It might be easier.

Karush-Kuhn-Tucker (KKT) conditions

At the optimal solution (w^*, α^*):

$$\begin{aligned}
\frac{\partial L(w^*, \alpha^*)}{\partial w_i} &= 0 \qquad \text{for } i = 1 \text{ to } d \\
\alpha_i^* g_i(w^*) &= 0 \qquad \text{for } i = 1 \text{ to } n \\
g_i(w^*) &\leq 0 \qquad \text{for } i = 1 \text{ to } n \\
\alpha_i^* &\geq 0 \qquad \text{for } i = 1 \text{ to } n
\end{aligned}$$

Linear SVM Optimization Problem: Revisited

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} = \frac{1}{2}||\mathbf{w}||^2 \qquad \text{(objective function)} \\ \text{subject to} \quad & y_i(\mathbf{x}_i^T\mathbf{w} + b) \geq 1 \quad (i = 1, \cdots, n) \end{aligned}$$

OR

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||\mathbf{w}||^2 \qquad \text{(objective function)} \\ \text{subject to} \quad & 1 - y_i(\mathbf{x}_i^T\mathbf{w} + b) \leq 0, \quad (i = 1, \cdots, n) \end{aligned}$$

Lagrange Formulation for Linear SVMs

$$\begin{aligned} &\text{minimize} \quad \frac{1}{2}||\mathbf{w}||^2 \qquad \text{(objective function)} \\ &\text{subject to} \quad 1 - y_i(\mathbf{x}_i^T\mathbf{w} + b) \le 0, \quad (i = 1, \cdots, n) \end{aligned}$$

The problem can be solved by Lagrange multipliers method.

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}||\mathbf{w}||^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i(\mathbf{x}_i^T\mathbf{w} + b))$$

Primal or Dual Problem

The primal problem is given by:

$$\min_{\mathbf{w},b} \max_{\alpha} L(\mathbf{w}, b, \alpha)$$

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

with respect to **w**, b and the Lagrange coefficients $\alpha_i \geq 0$.

The dual problem is given by:

$$\max_{\alpha} \min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)$$

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

with respect to **w**, b and the Lagrange coefficients $\alpha_i \geq 0$.

Apply KKT conditions on the Dual problem

$$\max_{\alpha} \min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)$$

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

with respect to $\mathbf{w}$, b and the Lagrange coefficients $\alpha_i \geq 0$. We let

$$\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, b, \alpha) = 0, \quad \frac{\partial}{\partial b} L(\mathbf{w}, b, \alpha) = 0$$

These lead, respectively, to

$$\mathbf{w} = \sum_{j=1}^{n} \alpha_j y_j \mathbf{x}_j, \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Dual Problem

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

Substituting the two equations

$$\mathbf{w} = \sum_{j=1}^{n} \alpha_j y_j \mathbf{x}_j, \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

into the Dual problem, we get:

$$\max_{\alpha} L(\alpha) = \max_{\alpha} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right\}$$

$$\text{subject to} \quad \alpha_i \geq 0, \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Simplified Dual versus Primal Form

Primal:

$$\min_{\mathbf{w},b} \max_{\alpha} L(\mathbf{w}, b, \alpha)$$

$$= \min_{\mathbf{w},b} \max_{\alpha} \left\{ \frac{1}{2}||\mathbf{w}||^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\mathbf{x}_i^T \mathbf{w} + b)) \right\}$$

with respect to $\mathbf{w}$, b and the Lagrange coefficients $\alpha_i \geq 0$.

Dual:

$$\max_{\alpha} L(\alpha) = \max_{\alpha} \left\{ \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right\}$$

subject to $\alpha_i \geq 0, \quad \sum_{i=1}^{n} \alpha_i y_i = 0$

Example

$$Dual: \quad L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to} \quad \alpha_i \geq 0, \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Consider the following 2-D dataset (x_1 and x_2 are the attributes and y is the class variable).

x_1	x_2	y
0	0	$+1$
0	1	-1
1	0	-1
1	1	$+1$

Write the expression for the dual problem. Let $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ be the Lagrangian multipliers associated with the four data points.

$$(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) - \frac{1}{2} \{ \text{ sixteen tuples} \ldots \}$$

subject to $\alpha_1, \alpha_2, \alpha_3, \alpha_4 \geq 0$ and $\alpha_1(+1) + \alpha_2(-1) + \alpha_3(-1) + \alpha_4(+1) = 0$.

The last constraint simplifies to $\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4 = 0$.

Steps in constructing the Dual

Start with an empty objective function

- Add the term $\sum_{i=1}^{n} \alpha_i$ to the objective function
- Construct the so-called Kernel matrix $K(\mathbf{x}_i, \mathbf{x}_j)$ which stores $\mathbf{x}_i^T \mathbf{x}_j$ for all indexes i, j over the example. The cell (i, j) in the matrix is the dot product of the features associated with the i-th and j-th example respectively.
  - For example, the dot product of the examples (x_1, x_2, y): (1,0,-1) and (1,1,+1) is 1*1+0*1=1.
- For (i, j), compute $K(\mathbf{x}_i, \mathbf{x}_j) * y_i * y_j$ and add $-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) * y_i * y_j$ to the objective function
- Add the constraints $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$.

Complexity

$O(d)$ for each element of the kernel matrix. There are n^2 elements. Therefore, the complexity of constructing the Kernel matrix is $O(n^2 d)$. There are $O(n^2)$ terms in the objective function and each takes $O(1)$ for lookup (once the Kernel matrix is constructed). Therefore the overall complexity is $O(n^2 d)$ for constructing the optimization problem.