Optimization Theory and Duality for SVMs

Outline:

- Tools for designing learning (training) algorithms.
- How to make the optimization problem more tractable?
- A dual representation of the optimal hyperplane in terms of the training examples.
- What insight do we gain from the dual representation?
- What are the properties of the dual optimization problem?

Quadratic Program

$$\text{minimize} \quad P(\vec{w}) = -\left(\sum_{i=1}^{n} w_i\right) + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j H_{ij}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} w_i g_i^{(1)} \le 0 \quad \dots \quad \sum_{i=1}^{n} w_i g_i^{(k)} \le 0$$

$$\sum_{i=1}^{n} w_i h_i^{(1)} = 0 \quad \dots \quad \sum_{i=1}^{n} w_i h_i^{(m)} = 0$$

- k linear inequality constraints
- m linear equality constraints
- Gram Matrix $H=H_{(i,j)}$ is pos. semi-definite $\forall w_1 \dots w_n: \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j H_{ij} \ge 0$ => convex, no local optima
- $\vec{w}$ is feasible, if it fulfills constraints

Fermat Theorem

Given an unconstrained optimization problem

$$\text{minimize } P(\vec{w})$$

with $P(\vec{w})$ convex, a necessary and sufficient conditions for a point $\vec{w}^{\circ}$ to be an optimum is that

$$\frac{\delta P(\vec{w}^{\circ})}{\delta \vec{w}} = 0$$

Lagrange Function

Given an optimization problem

$$\begin{aligned} &\text{minimize } P(\vec{w}) \\ &\textbf{s.t.} \quad g_1(\vec{w}) \le 0 \quad \dots \quad g_k(\vec{w}) \le 0 \\ &\qquad h_1(\vec{w}) = 0 \quad \dots \quad h_m(\vec{w}) = 0 \end{aligned}$$

the Lagrangian function is defined as

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

- $\vec{\alpha}$ and $\vec{\beta}$ are called Lagrange Multipliers

Lagrange Theorem

Given an optimization problem

$$\text{minimize } P(\vec{w}) \quad \textbf{s.t. } h_1(\vec{w}) = 0 \quad \ldots \quad h_m(\vec{w}) = 0$$

with $P(\vec{w})$ convex and all h affine (w*x+b), necessary and sufficient conditions for a point $\vec{w}^\circ$ to be an optimum are the existence of $\vec{\beta}^\circ$ such that

$$\frac{\delta L(\vec{w}^\circ, \vec{\beta}^\circ)}{\delta \vec{w}} = 0 \qquad \frac{\delta L(\vec{w}^\circ, \vec{\beta}^\circ)}{\delta \vec{\beta}} = 0 \qquad L(\vec{w}, \vec{\beta}) = P(\vec{w}) + \sum_{i=1}^{m} \beta_i h_i(\vec{w})$$

=> $L(\vec{w}^\circ, \vec{\beta}) \le L(\vec{w}^\circ, \vec{\beta}^\circ) \le L(\vec{w}, \vec{\beta}^\circ)$

Karush-Kuhn-Tucker Theorem

Given an optimization problem

$$\begin{aligned} &\text{minimize } P(\vec{w}) \\ &\textbf{s.t. } g_1(\vec{w}) \le 0 \quad \ldots \quad g_k(\vec{w}) \le 0 \\ &\qquad h_1(\vec{w}) = 0 \quad \ldots \quad h_m(\vec{w}) = 0 \end{aligned}$$

with $P(\vec{w})$ convex and all g and h affine, necessary and sufficient conditions for a point $\vec{w}^\circ$ to be an optimum are the existence of $\vec{\alpha}^\circ$ and $\vec{\beta}^\circ$ such that

$$\frac{\delta L(\vec{w}^\circ, \vec{\alpha}^\circ, \vec{\beta}^\circ)}{\delta \vec{w}} = 0 \qquad \frac{\delta L(\vec{w}^\circ, \vec{\alpha}^\circ, \vec{\beta}^\circ)}{\delta \vec{\beta}} = 0$$

$$\alpha_i^\circ g_i(\vec{w}^\circ) = 0, i = 1, \ldots, k$$

$$g_i(\vec{w}^\circ) \le 0, i = 1, \ldots, k$$

$$\alpha_i^\circ \ge 0, i = 1, \ldots, k$$

Sufficient for convex QP: $max_{\vec{\alpha} \ge 0, \vec{\beta}} \left[min_{\vec{w}} L(\vec{w}, \vec{\alpha}, \vec{\beta}) \right]$

Dual Optimization Problem

Primal OP: minimize $P(\vec{w}, b) = \frac{1}{2}\vec{w} \cdot \vec{w}$, with $\forall i \left[y_i[\vec{w} \cdot \vec{x}_i + b] \ge 1 \right]$

Lemma: The solution w° can always be written as a linear combination

$$\vec{w}^\circ = \sum_{i=1}^{n} \alpha_i y_i \vec{x}_i \qquad \alpha_i \ge 0$$

of the training data.

==> Lagrange multipliers

Dual OP: maximize $D(\vec{\alpha}) = \left(\sum_{i=1}^{n} \alpha_i \right) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$

s.t. $\sum_{i=1}^{n} \alpha_i y_i = 0 \quad and \quad 0 \le \alpha_i$

==> positive semi-definite quadratic program

Primal <=> Dual

Theorem: The primal OP and the dual OP have the same solution. Given the solution α_i° of the dual OP,

$$\vec{w}^\circ = \sum_{i=1}^{n} \alpha_i^\circ y_i \vec{x}_i \qquad b^\circ = \frac{1}{2}(\vec{w}_0 \cdot \vec{x}^{pos} + \vec{w}_0 \cdot \vec{x}^{neg})$$

is the solution of the primal OP.

Theorem: For any feasible points $P(\vec{w}, b) \ge D(\vec{\alpha})$.

=> two alternative ways to represent the learning result

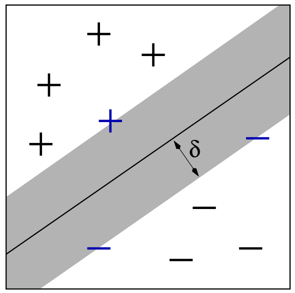
- weight vector and threshold $\langle \vec{w}, b \rangle$
- vector of " influences" $\alpha_1, \ldots, \alpha_n$

Properties of the Dual OP

Dual OP: maximize $D(\vec{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\vec{x}_i \cdot \vec{x}_j)$

s.t. $\sum_{i=1}^{n} \alpha_i y_i = 0 \quad and \quad 0 \le \alpha_i$

- single solution (*i.e.* $\langle \vec{w}, b \rangle$ is unique)
- one factor α_i for each training example
  - describes the "influence" of training examples *i* on the result
  - $\alpha_i > 0$ <=> training example is a support vector
  - $\alpha_i = 0$ else
- depends exclusively on inner product between training examples



Properties of the Soft-Margin Dual OP

Dual OP: maximize $D(\vec{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (\vec{x}_i \cdot \vec{x}_j)$

s. t. $\sum_{i=1}^{n} \alpha_i y_i = 0 \quad and \quad 0 \le \alpha_i \le C$

- (mostly) single solution (i. e. $\langle \vec{w}, b \rangle$ is almost always unique)
- one factor α_i for each training example
  - "influence" of single training example limited by *C*
  - $0 < \alpha_i < C$ <=> SV with $\xi_i = 0$
  - $\alpha_i = C$ <=> SV with $\xi_i > 0$
  - $\alpha_i = 0$ else
- based exclusively on inner product between training examples