CS711008Z Algorithm Design and Analysis

Lecture 9. Lagrangian duality and SVM

Dongbo Bu

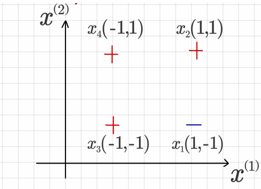
Institute of Computing Technology
Chinese Academy of Sciences, Beijing, China

Outline

- Classification problem and maximum margin strategy;
- Solving maximum margin problem using Lagrangian duality;
- SMO technique;
- Kernel tricks;

Classification problem and maximum margin strategy

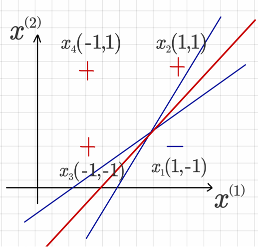
Classification problem



	$x^{(1)}$	$x^{(2)}$	y
x_1	1	-1	-1
x_2	1	1	$+1$
x_3	-1	-1	$+1$
x_4	-1	1	$+1$

- Given a set of samples with their category labels (denoted as $(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), ..., (\mathbf{x_n}, y_n)$, $y_i \in \{-1, +1\}$, the goal of classification problem is to find an appropriate function $f(\mathbf{x})$ that can describe the dependency between y_i and $\mathbf{x_i}$; thus, for a new sample $\mathbf{x}'$, we can infer its category based on $f(\mathbf{x}')$.
- A great variety of classification algorithms have been designed, including Fisher's linear discriminant, logistic regression, decision tree, neural network and SVM.

Linear classifier

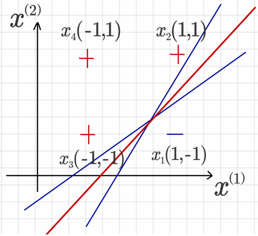


	$x^{(1)}$	$x^{(2)}$	y
x_1	1	-1	-1
x_2	1	1	$+1$
x_3	-1	-1	$+1$
x_4	-1	1	$+1$

- Unlike decision tree, SVM adopts the classifier with the following type:
  - If $f(\mathbf{x}) > 0$ then $y = +1$;
  - If $f(\mathbf{x}) < 0$ then $y = -1$;
- Let's first restrict the $f(\mathbf{x})$ to be linear, i.e.

$$f(\mathbf{x}) = \omega^T \mathbf{x} + b$$

The hyperplane $\omega^T \mathbf{x} + b = 0$ is denoted as separating hyperplane.



	$x^{(1)}$	$x^{(2)}$	y
x_1	1	-1	-1
x_2	1	1	$+1$
x_3	-1	-1	$+1$
x_4	-1	1	$+1$

- The objective of training procedure is to find an appropriate setting of ω and b such that all samples in the training set can be correctly labelled using the classifier. We will consider the torlerance of several mislabelled samples later.

Maximum margin strategy

- There are always multiple settings of ω and b that the corresponding classifier works perfectly on all samples. Which one should we use?
- We prefer the one such that the margin between positive and negative samples is maximized: The wider the margin is, the larger the generality performance on new samples. Thus, we needs to solve the following optimization problem:

$$\begin{array}{rl} \min_{w,b} & \frac{2}{||\omega||} \\ s.t. & y_i(w \cdot x_i + b) - 1 \geqslant 0 \quad i = 1, 2, \cdots, n \end{array}$$

- Note:
  - The restriction $f(\mathbf{x}) > 0$ for positive sample x is implemented as $f(\mathbf{x}) = 1$.
  - The distance for any point x to the hyperplane $\omega^T \mathbf{x} + b = 0$ is $\frac{|\omega^T \mathbf{x} + b|}{||\omega||}$. Thus, the margin is: $\frac{2}{||\omega||}$.

An equivalent form with quadratic objective function

- An equivalent form is:

$$\begin{array}{rl} \min_{w,b} & \frac{1}{2}\left\|w\right\|^2 \\ s.t. & y_i(w \cdot x_i + b) - 1 \geqslant 0 \quad i = 1, 2, \cdots, n \end{array}$$

- Question: how to solve this optimization problem subject to inequality constraints?
- Of course we solve the problem (called primal problem hereafter) directly using convex quadratic programming techniques; however, consider its dual problem will bring great benefits.
- Let's review the conditions of the optimal solution first.

Lagrangian dual explanation of maximum margin problem

- Primal problem:

$$\begin{array}{lrcl} \min_{w,b} & \frac{1}{2}\left\|w\right\|^2 & & \\ s.t. & y_i(w^T x_i + b) & \geq & 1, \quad i \in \{1, ..., n\} \end{array}$$

- Lagrangian:
$$L(w,b,\alpha) = \frac{1}{2}\left\|w\right\|^2 - \sum_{i=1}^{n} \alpha_i y_i (w \cdot x_i + b) + \sum_{i=1}^{n} \alpha_i$$
- Notice that **Lagrangian is a lower bound of the primal objective function**, i.e. $\frac{1}{2}\left\|w\right\|^2 \geq L(w,b,\alpha)$, when $\alpha \geq \mathbf{0}$ and $\mathbf{w}, \mathbf{b}$ is feasible.
- Furthermore we have

$$\frac{1}{2}\left\|w\right\|^2 \geq L(w,b,\alpha) \geq \inf_{w,b} L(w,b,\alpha)$$

when $\alpha \geq 0$ and w, b is feasible.
- Denote **Lagrangian dual** $g(\alpha) = \inf_{w,b} L(w,b,\alpha)$. The above inequality can be rewritten as:

$$\frac{1}{2}\left\|w\right\|^2 \geq L(w,b,\alpha) \geq g(\alpha)$$

Lagrangian dual function

- What is the Lagrangian dual $g(\alpha)$?

$$g(\alpha) \quad = \quad \inf_{w,b} L(w, b, \alpha)$$

- To calculate the inferior bound of $L(w, b, \alpha)$, we set its derivates to be 0, i.e.,

$$\frac{\partial L(w, b, \alpha)}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0$$

$$\frac{\partial L(w, b, \alpha)}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i = 0$$

and obtain Lagrangian dual function:

$$g(\alpha) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i^T x_j) + \sum_{i=1}^{n} \alpha_i$$

- Thus $g(\alpha)$ is a lower bound of $\frac{1}{2} \|w\|^2$ when $\sum_{i=1}^{n} \alpha_i y_i = 0$ and $\alpha \geq 0$.

Lagrangian dual problem

- Now let's try to find **the tightest lower bound** of $\frac{1}{2}\left\|w\right\|^2$, which can be calculated by solving the following Lagrangian dual problem:

$$\begin{array}{lrcl} \max & -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j (x_i^T x_j) + \sum_{i=1}^{n}\alpha_i & & \\ s.t. & \sum_{i=1}^{n}\alpha_i y_i & = & 0 \\ & \alpha & \geq & 0 \end{array}$$

- The dual problem has an identical optimal objective function value to the primal problem as the Slater's conditions hold.
- One advantage of the dual problem is that x_i and x_i appears in the form of inner product $x_i^T x_j$; thus, we can simply define a kernel function $k(x_i, x_j) = \phi(x_i)^T\phi(x_j)$ without knowing the details of map $\phi(.)$.