10-715: Adv. Intro to Machine Learning **Fall 2015**

Recitation 2: October 20

Eric Wong

2.1 Duality

In these notes, we define the basics of solving dual problems and go through SVM soft regression as an example. For more detailed coverage, we refer to Barnabas slides from last year `https://www.cs.cmu.edu/~epxing/Class/10715/lectures/Duality.pdf`.

2.1.1 Lagrangian

Consider a general minimization problem:

$$\min f(x) \tag{2.1}$$
$$\text{s.t. } h_i(x) \leq 0, \quad \forall i \in [m], \tag{2.2}$$
$$l_j(x) = 0 \quad \forall j \in [k], \tag{2.3}$$

The Lagrangian is defined as follows:

$$L(x,u,v) = f(x) + \sum_i u_i h_i(x) + \sum_j v_j l_j(x)$$

subject to $u \geq 0$. What is this? This is simply a lower bound on $f(x)$ when x is feasible (satisfies the constraints). More importantly:

$$f(x^*) \geq L(x^*,u,v) \geq \min_x L(x,u,v) = g(u,v)$$

where g is called the Lagrange dual function. This is a lower bound of $f(x^*)$.

2.1.2 Motivation

Why do we care about dual problems? The dual problem is always a concave maximization problem, regardless of the primal problem! (pointwise maximum of convex functions is convex, and $u \geq 0$ is convex)

$$g(u,v) = \min_x f(x) + \sum_i u_i h_i(x) + \sum_j v_j l_j(x)$$

Thus, nasty primal problems can be transformed and thrown into a convex optimization solver.

2.1.3 Strong and Weak Duality

Let $u^*, v^* = \arg\max_{u,v} g(u,v)$. Then, $f(x^*) \geq g(u^*,v^*)$. This property is called weak duality and is always true regardless of the original problem (including non-convex primal problems).

If $f(x^*) = g(u^*,v^*)$ then we have strong duality. One useful conditino for strong duality is Slater's condition:

Theorem 2.1 *Slater's Condition: If the primal is a convex problem (convex f, h_i and affine l_j), and if there exists at least one strictly feasible x, then we have strong duality.*

2.1.4 KKT conditions

The KKT conditions are a set of conditions that are sufficient for optimality, and necessary under strong duality. A more detailed proof can be found in last year's slides.

1. Stationarity: $0 \in \partial f(x) + \sum_i u_i \partial h_i(x) + \sum_j v_j \partial l_j(x)$
2. Complementary slackness: $u_i \cdot h_i(x) = 0$
3. Primal feasibility: $h_i(x) \leq 0, l_j(x) = 0$
4. Dual feasibility: $u_i \geq 0$

Theorem 2.2 *If x^*, u^*, v^* satisfy the KKT conditions, then x^*, u^*, v^* are primal and dual solutions.*

2.2 SVM Regression Example

For SVM Regression we have the following objective:

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{n} L_\epsilon(x_i, y_i, w^T x_i), \tag{2.4}$$

where L is ϵ-sensitive loss:

$$L_\epsilon(x, y, w^T x) = \max(0, |y - w^T x - b| - \epsilon), \tag{2.5}$$

Introduce slack variables ξ_i, ξ_i^*. These represent the actual amount of loss incurred:

$$\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*), \tag{2.6}$$

subject to $y_i - \epsilon - \xi_i^* \leq w^T x_i + b \leq y_i + \epsilon + \xi_i$ and $\xi_i, \xi^* \geq 0$. so the Lagrange is

$$L(w, u, u^*, v, v^*) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) + \sum_i u_i^*(y_i - \epsilon - \xi_i^* - w^T x_i - b) + \sum_i u_i(w^T x_i + b - y_i - \epsilon - \xi_i) - \sum_i v_i \xi_i - \sum_i v_i^* \xi_i^*$$

Stationarity conditions:

$$\partial_w L = 0 = w + \sum_i (u_i - u_i^*) x_i$$

$$\partial_b L = 0 = \sum_i (u_i - u_i^*)$$

$$\partial_{\xi_i} L = 0 = C - u_i - v_i$$

$$\partial_{\xi_i^*} L = 0 = C - u_i^* - v_i^*$$

Plug these in to get

$$g(u, u^*) = -\frac{1}{2}(u - u^*)^T K(u - u^*) + \sum_i (u_i - u_i^*) y_i - \sum_i (u_i + u_i^*)\epsilon$$

subject to the constraints from the stationarity conditions, that $0 = \sum_i (u_i - u_i^*)$ and $u_i, u_i^* \in [0, C]$. After solving the dual, we can retrieve the primal with $w = \sum_i (u_i - u_i^*) x_i$ and so $f(x) = \sum_i (u_i - u_i^*) x_i^T x + b$. We can retrieve b by exploiting complementary slackness.