Optimization Methods		Date:	*10/4/2018*
Instructor: *Sujit Prakash Gujar*		Scribes:	M Sreenivasan M Vamsidhar

Lecture 23: Non Linear Programming - Lagrangian Duality and its Application

1 Recap

In the last lecture, we covered the following:

1.1 Lagrangian formulation

Generalized Optimization problem:

$$\begin{aligned} \min \;\; & f(x) \;\; x \in \mathcal{R}^n \\ \text{subject to} \;\; & g_i(x) \leq 0; \;\; \forall i = 1, 2, ..m. \\ & h_j(x) = 0; \;\; \forall j = 1, 2, ..l. \end{aligned} \tag{1}$$

Writing in Lagrangian form:

$$L(x, \mu, \lambda) \;\; = \;\; f(x) \;\; + \;\; \sum_{i=1}^{m} \mu_i g_i(x) + \sum_{j=1}^{l} \lambda_j h_j(x) \tag{2}$$

1.2 KKT conditions

(I) First order conditions:

$$\nabla_x L(x^*, \mu^*, \lambda^*) \;\; = 0 \tag{3a}$$

Expanding on it, we get

$$\nabla f(x^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(x^*) + \sum_{i=1}^{l} \lambda_i^* \nabla h_j(x^*) \;\; = \;\; 0 \tag{3b}$$

(II) Feasibility conditions:

$$h_j(x^*) \;\; = \;\; 0 \;\; \forall \;\; j = 1, 2, ..l \tag{3c}$$

$$g_i(x^*) \leq 0 \;\; \forall \;\; i = 1, 2, ..m. \tag{3d}$$

(III) Complimentary slackness conditions:

$$\mu_i^* g_i(x^*) = 0 \;\; \forall \;\; i = 1, 2, ..m \tag{3e}$$

(IV) Dual feasibility conditions:

$$\mu_i^* \;\; \geq \;\; 0 \;\; \forall \;\; i = 1, ..., m \tag{3f}$$

These equations 3a,3c,3d, 3e, 3f are called KKT conditions and are necessary conditions and not always. Modified version of this is called Fritz-John Multiplier which is broader than KKT and is obtained by multiplying μ_0 to f(x).

1.3 Water filling formulation

Consider, for a communication system there are n communication channels and x_i is the power on the i^{th} channel. Then the transmission rate of the communication system is given as follows

$$\sum_{i=1}^{n} \log(\alpha_i + x_i) \tag{4}$$

where α_i are constants. Our objective here is to maximize the transmission rate of the system i.e *maximizing* equation 4. Now, our optimization problem can be written as

$$\begin{aligned} &\max \sum_{i=1}^{n} \log(\alpha_i + x_i) \ \ \forall \ \ \alpha_i > 0 \\ \text{subject to } &\sum_{i=1}^{n} x_i = 1 \\ &x_i \geq 0, \ \ \forall i \end{aligned} \tag{5}$$

Converting into standard form,

$$\begin{aligned} f(x) &= -\sum_{i=1}^{n} \log(\alpha_i + x_i) \ \ \forall \ \ \alpha_i > 0 \\ g_i(x) &= -x_{\mathrm{i}} \leq 0 \\ h(x) &= -\sum_{i=1}^{n} x_i - 1 = 0 \end{aligned} \tag{6}$$

2 Solving Water filling

From equation 6 and equations 3a, 3c, 3d, 3e, 3f, we can write the following

First order condition:

$$-\frac{1}{\alpha_i + x_i} - \mu^* + \lambda^* = 0 \ \ \forall \ \ i \tag{7a}$$

Feasibility condition:

$$x_i^* \geq 0 \ \ \forall \ \ i \tag{7b}$$

$$\sum_{i=1}^{n} x_i^* = 1 \tag{7c}$$

$$\mu_i^* \geq 0 \ \ \forall \ \ i \tag{7d}$$

Complementary Slackness Conditions:

$$\mu_i^* x_i^* = 0 \ \ \forall \ \ i \tag{7e}$$

(7f)

From equations 7a and 7d, we can say that

$$\begin{aligned} \lambda^* &= \frac{1}{\alpha_i + x_i} + \mu^* \ \ \forall \ \ i \\ &\geq \frac{1}{\alpha_i + x_i} \end{aligned} \tag{8}$$

Now, we get two cases to solve.

1. Case 1: $\lambda^* < \frac{1}{\alpha_i}$

- $x_i^* > 0$
- $\mu_i^* = 0$ (from equation 7e)

Therefore, $\lambda^* = \frac{1}{\alpha_i + x_i^*} or x_i^* = \frac{1}{\lambda^*} - \alpha_i$

2. Case 2: $\lambda^* \geq \frac{1}{\alpha_i}$
   - If $x_i^* > 0$, then $\lambda^* \geq \frac{1}{\alpha_i} > \frac{1}{\alpha_i + x_i}$ and also $\mu_i^* > 0$
   - Both $\mu_i^* > 0$ and $x_i^* > 0$, directly contradict equation 7e
   - Therefore, $x_i^* = 0$

Basically, we can write x_i^* as follows.

$$x^* = \max(0, \frac{1}{\lambda^*} - \alpha_i) \tag{9}$$

Equating $R.H.S$ part of equation 9 to 1 gives us the Water Filling Algorithm.

3 Lagrangian Duality

For a given optimization problem $f(x)$, it's Lagrangian form is defined as detailed in equation 2. We define Lagrange dual function as follows.

$$\theta(\mu, \lambda) = \min_x L(x, \mu, \lambda) \tag{10}$$

Now, the dual problem to the original minimization problem $f(x)$ can be written as

$$\begin{aligned} \theta^* = & \max_{\mu,\lambda} \theta(\mu, \lambda) \\ \text{such that } & \mu_i \geq 0 \ \ \forall \ \ i \end{aligned} \tag{11}$$

The primal and dual optimal values, f^* and θ^*, always satisfy weak duality.

$$f^* \geq \theta* \tag{12}$$

3.1 Weak duality proof for Lagrangian

Given a LP:

$$\begin{aligned} \max_x \ & c^T x \\ \text{subject to } & Ax \leq b \\ & x \geq 0 \end{aligned}$$

It's dual is written as:

$$\begin{aligned} \min_y \ & f = b^T y \\ \text{subject to } & c - A^T y \leq 0 \\ & -y \leq 0 \end{aligned}$$

The Lagrangian of above dual is written as:

$$\begin{aligned} L(y, x, z) &= b^T y + \sum_{i=1}^{n} x_i^T (c - A^T y)_i - \sum_{j=1}^{m} z_j y_j \\ &= b^T y + x_i^T (c - A^T y) - z^T y \end{aligned}$$

The dual of above written Lagrangian can be written as

$$\begin{aligned}\theta(x,z) &= \min_y(b^T y + x_i^T(c - A^T y) - z^T y) \\ &= c^T x + \min_y((b^T - (Ax)^T - z^T)y)\end{aligned}$$

$$\begin{aligned}\text{If} \quad & b^T - (Ax)^T - z^T = 0 \\ \text{then} \quad & \theta^* = c^T x^*\end{aligned}$$

Therefore

$$b^T y^* \geq c^T x^* \qquad \because f^* \geq \theta^*$$

which is weak duality of LP

For the case of Strong duality to exist in our given optimization problem, it needs to satisfy Slater conditions.

4 Support Vector Machines

The primary objective or the optimization problem of a support vector machine(SVM) can be written as

$$\begin{aligned}&\max \frac{2}{||w||} \\ &\text{subject to} \\ &w^T x_i + b \geq 1 \quad \forall \quad i \; \epsilon \text{ Class 1} \\ &w^T x_i + b \leq -1 \quad \forall \quad i \; \epsilon \text{ Class 2}\end{aligned} \tag{15}$$

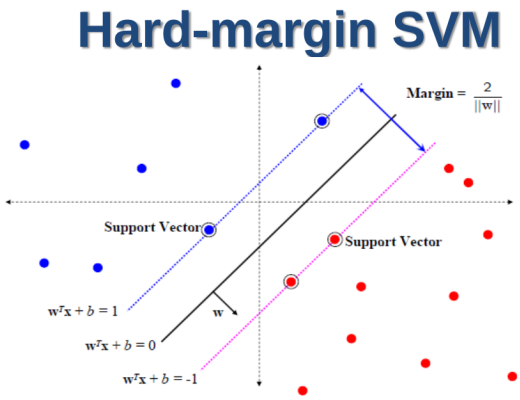

Figure 1: A support vector machine: Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.
Source: Wikipedia

The object here is to maximize the margin between support vectors. It is hard to maximize the above equation 15. So we change it another form to make it feasible for optimization. Also, for m classes, we need m constraints. We also need a way to incorporate all these constraints in a simple and elegant way while reformulating our problem. So, the form can be written as

$$\begin{aligned}&\min \frac{||w||^2}{2} \\ &\text{subject to} \quad 1 - y_i(w^T x_i + b) \leq 0 \quad \forall \quad i = 1, ..., m\end{aligned} \tag{16}$$

4.1 SVM Lagrangian problem

The idea behind above transformation is that max $\frac{1}{||w||}$ is same as min $||w||$ which again is same as the objective in 16. To solve the above problem, we need to convert it into it's Lagrangian form $\alpha(w, b, \mu)$. Therefore, we introduce it as follows

$$\alpha(w, b, \mu) = \frac{||w||^2}{2} + \sum_{i=1}^{m} \mu_i [1 - y_i(w^T x_i + b)] \tag{17}$$

To get the solution of the above problem 16, we need to solve the Lagrangian problem 17. But, since we have inequalities present in equation 16, we know that the solution to Lagrangian in 17 will satisfy Karush-Kuhn-Tucker(KKT) conditions. These first-order necessary conditions for our problem can be written formulated as follows.

(I) First order conditions:

$$\nabla_w \alpha = 0 \quad \Rightarrow \quad w^* = \sum_{i=1}^{m} \mu_i^* y_i x_i \tag{18a}$$

$$\nabla_b \alpha = 0 \quad \Rightarrow \quad \sum_{i=1}^{m} \mu_i^* y_i = 0 \tag{18b}$$

(II) Feasibility conditions:

$$1 - y_i(w^T x_i + b) \leq 0 \ \ \forall \ i = 1, ..., m \tag{18c}$$

(III) Complimentary slackness conditions:

$$\mu_i [1 - y_i(w^T x_i + b)] = 0 \ \ \forall \ i = 1, ..., m \tag{18d}$$

(IV) Dual feasibility conditions:

$$\mu^* \geq 0 \ \ \forall \ i = 1, ..., m \tag{18e}$$

If we consider the complimentary slackness condition in equation 18d, we see that we have two possible scenarios.

1. $\mu_i^* > 0 \quad \Rightarrow \quad y_i(w * x_i + b) = 1$
2. $\mu_i^* = 0 \quad \Rightarrow \quad y_i(w * x_i + b) > 1$

In SVM, we know that support vectors have $w * x_i + b = 1$. We can see that above first scenario satisfies this. Thus, we can say that Support vectors are examples which have positive Lagrange Multiplier.

4.2 Wolfe dual problem

The duality principle tells us that an optimization problem can be viewed from two perspectives. The first one is a primal problem, a minimization problem in our case and the second one is the dual problem, which will be a maximization problem. In this subsection, we try to find the dual form of above Lagrangian problem 17.

$$\text{Lagrangian Function:} \quad \alpha(w, b, \mu) = \frac{1}{2} w^T w + \sum_{i=1}^{m} \mu_i [1 - y_i(w^T x_i + b)] \tag{19}$$

Substituting the value of w from equation 18a into α in equation 19, we get

$$\begin{aligned} D(w,b) &= \frac{1}{2}(\sum_{i=1}^{m}\mu_i y_i x_i)(\sum_{j=1}^{m}\mu_j y_j x_j) + \sum_{i=1}^{m}\mu_i[1 - y_i((\sum_{j=1}^{m}\mu_j y_j x_j).x_i + b)] \\ &= \frac{1}{2}(\sum_{i=1}^{m}\sum_{j=1}^{m}\mu_i\mu_j y_i y_j x_i x_j) - \sum_{i=1}^{m}\mu_i y_i((\sum_{j=1}^{m}\mu_j y_j x_j).x_i + b) + \sum_{i=1}^{m}\mu_i \\ &= \frac{1}{2}(\sum_{i=1}^{m}\sum_{j=1}^{m}\mu_i\mu_j y_i y_j x_i x_j) - \sum_{i=1}^{m}\sum_{j=1}^{m}\mu_i\mu_j y_i y_j x_i x_j - b\sum_{i=1}^{m}\mu_i y_i + \sum_{i=1}^{m}\mu_i \\ &= \sum_{i=1}^{m}\mu_i - \frac{1}{2}(\sum_{i=1}^{m}\sum_{j=1}^{m}\mu_i\mu_j y_i y_j x_i x_j) - b\sum_{i=1}^{m}\mu_i y_i \end{aligned} \tag{20}$$

So in the above equation 20, we removed w. But b is still there. But from equation 18b, we know that $\sum_{i=1}^{m}\mu_i y_i = 0$. So, the third term in equation 20 becomes zero. We can write the final equation as

$$D(\mu) = \sum_{i=1}^{m}\mu_i - \frac{1}{2}(\sum_{i=1}^{m}\sum_{j=1}^{m}\mu_i\mu_j y_i y_j x_i x_j) \tag{21}$$

This is called Wolfe dual Lagrangian function. We can rewrite our original optimization problem 17 as Wolfe dual problem.

$$\begin{aligned} \max_{\mu} \quad & \sum_{i=1}^{m}\mu_i - \frac{1}{2}(\sum_{i=1}^{m}\sum_{j=1}^{m}\mu_i\mu_j y_i y_j x_i x_j) \\ \text{subject to} \quad & \mu_i \geq 0. \ \ \forall \ \ i = 1, .., m \\ & \sum_{i=1}^{m}\mu_i y_i = 0 \end{aligned} \tag{22}$$

4.3 Soft SVM and Hinge Loss

Till now what we discussed is hard SVM which doesn't allow any mis-classifications. But, for some data which might not get classified without any mis-classifications, we need a different objective function than equation 16 which also integrates the misclassification loss ζ_i for a misclassified sample i. In fact, the soft SVM does just that. It's formulation of optimization problem is given as follows.

$$\begin{aligned} \min \quad & \frac{||w||^2}{2} + C\sum_{i=1}^{m}\zeta_i \\ \text{subject to} \quad & \zeta_i \geq 1 - y_i(w^T x_i + b) \ \ \forall \ \ i \\ & \zeta_i \geq 0 \ \ \forall \ \ i \end{aligned} \tag{23}$$

From the constraints of the above in equation 23, we can write ζ as follows.

$$\zeta_i = \max(0, 1 - y_i(w^T x_i + b)) \tag{24}$$

This above equation 24 is called Hinge Loss.