Optimization Methods	Date: *10 April 2018*
Instructor: *Sujit Prakash Gujar*	Scribes: Shyam Nandan Rai(20172088), Gokul B. Nair(201502034), Hemanth Vemuri(201564126)

Lecture 23: Lagrangian Duality and Applications

1 Recap

- **Equality Constraint Optimization:** As we learnt in previous lecture, Lagrangian and Lagrange Multipliers are proposed to handle the optimization of non-linear functions subject to some equality constraints.

$$minf(x)$$
$$s.t. : h_i(x) = 0, i = 1..n$$

The Lagrange Multiplier for the optimization is:

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{i=1}^{n} \lambda_i h_i(x)$$

Here λ is Lagrange Multiplier for each constraint. If x*,λ^* is the local minima for $\mathcal{L}$, then:

  - $\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$
  - $\nabla_\lambda \mathcal{L}(x^*, \lambda^*) = 0$

- **Inequality Constraint Optimization:** As we learnt in the previous lecture, we use the KKT conditions to generalize an unconstrained optimization problem.

$$minf(x) \quad s.t. : h_j(x) = 0, g_i(x) \leq 0 \quad i = 1, ..m; j = 1, ..m$$

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

The four KKT conditions are:

  - First order condition: $\nabla f(x^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(x^*) + \sum_{i=1}^{m} \lambda_j^* \nabla h_j(x^*) = 0$
  - Feasibility conditions: $h_j(x^*) = 0 \quad j = 1, ..m$
    $g_i(x^*) \leq 0 \quad , i = 1, ..m$
  - Complimentary slackness: $\mu_i^* g_i(x^*), \forall i = 1, ..m$
  - Dual Feasibility Condition: $\mu_i^* \geq 0, \forall i = 1, ..m$

- **Fritz John Conditions:** KKT conditions are not necessary conditions for a solution in non-linear programming to be optimal. To overcome this we have Fritz John condiotions which are same as KKT conditions except the first order condition which is:

$$\mu_0 \nabla f(x^*) + \sum_{i=1}^{m} \mu_i^* \nabla g_i(x^*) + \sum_{i=1}^{m} \lambda_j^* \nabla h_j(x^*) = 0, \mu_0 \geq 0$$

2 Water Filling Algorithm

We will now discuss the problem of calculation of the amount of power to be transmitted for maximum bit rate given a limited amount of power and a given set of channels. This can be solved using the Water Filling Algorithm.
Let there be n communication channels and x_i be the power in the i^{th} channel. The optimization problem in the standard form is as follows:

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

$$s.t. : \sum_{i=1}^{n} x_i - 1 = 0$$

$$-x_i \leq 0, \forall i$$

Lagrangian:

$$\mathcal{L}(x, \mu, \lambda) = -\sum_{i=1}^{n} \log(\alpha_i + x_i) - \sum_{i=1}^{n} \mu_i x_i + \lambda(\sum_{i=1}^{n} x_i - 1)$$

The KKT conditions are:

- First Order Condition: $-\frac{1}{\alpha_i + x_i^*} - \mu_i^* + \lambda^* = 0, \forall i$
- Feasibility Condition:

$$x_i^* \geq 0, \forall i$$

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

$$\mu_i^* \geq 0, \forall i$$

- Complimentary Slackness Condition: $\mu_i^* x_i^* = 0, \forall i$

From First Order Condition:

$$\lambda^* \geq \frac{1}{\alpha_i + x_i} \forall i (\because \mu_i^* \geq 0 \forall i)$$

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

$$x_i^* > 0 \Rightarrow \mu_i^* = 0 (from \quad Complementary \quad Slackness \quad Condition)$$

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

- **Case 2:** $\lambda^* \geq \frac{1}{\alpha_i}$

$$If \quad x_i^* > 0, \quad \lambda^* > \frac{1}{\alpha_i + x_i^*}$$

$$\Rightarrow \mu_i^* > 0 (Contradiction \quad to \quad Complementary \quad Slackness \quad Condition)$$

$$\Rightarrow x_i^* = 0$$

Hence,

$$x_i^* = \max(0, \frac{1}{\lambda^*} - \alpha_i), \forall i = 1, ..n$$

$$s.t : \sum_{i=1}^{n} x^* = 1$$

3 Lagrange Duality

Consider the following non-linear optimization problem:

$$\min f(x) \tag{1}$$

$$\text{s.t: } g_i(x) \leq 0, i = 1, ..., m \tag{2}$$

$$h_j(x) = 0, j = 1, ..., l \tag{3}$$

The Lagrangian dual problem is defined as the following non-linear form:

$$\max \theta(\mu, \lambda) \tag{4}$$

$$\text{s.t: } \mu \geq 0 \tag{5}$$

Where,

$$\theta(\mu, \lambda) = \min_x \mathcal{L}(x, \mu, \lambda) \tag{6}$$

or,

$$\theta(\mu, \lambda) = \min_x (f(x) + \sum_{i=1}^{m} \mu_i g_i(x) + \sum_{j=1}^{l} \lambda_j h_j(x)) \tag{7}$$

The above dual formulation of Lagrangian results into a convex optimization problem, whose optimal value can be given as:

$$\theta^* = \theta(\mu^*, \lambda^*) = \max_{\mu,\lambda} \theta(\mu, \lambda) \tag{8}$$

Proof: Primal and Dual optimal values of Lagrangian always satisfy weak duality.

From the above equation 7, we get:

$$\theta(\mu, \lambda) = \min_x (f(x) + \sum_{i=1}^{m} \mu_i g_i(x) + \sum_{j=1}^{l} \lambda_j h_j(x)), \forall \text{ feasible x} \tag{9}$$

Since, $\sum_{i=1}^{m} \mu_i g_i(x)$ results in a negative value, and $\sum_{j=1}^{l} \lambda_j h_j(x))$ is zero, which results into:

$$\min_x (f(x) + \sum_{i=1}^{m} \mu_i g_i(x) + \sum_{j=1}^{l} \lambda_j h_j(x)) \leq \min_x f(x) = f^*, \forall \mu \geq 0, \forall \text{ feasible x} \tag{10}$$

$$\text{Hence, } \boxed{\theta^* \leq f^*} \tag{11}$$

Slater's Conditions: Strong duality holds if there exists a strictly feasible point, i.e. some x such that the inequality constraints are strictly satisfied, with:

$$g_i(x) < 0 \text{ and } h_j(x) = 0 \tag{12}$$

4 Support Vector Machines

Support Vector Machines(SVM) are optimization methods that are widely used in the field of Computer Vision and Machine Learning. The general idea is to find a hyperplane that separates the negative and positive examples with maximum margin.

Formulating the problem more mathematically, assume we have two class problem dataset, containing training samples x_i where $x_i \in \mathcal{R}^D$ and x_i either belongs to class C_1(Positive Samples) or C_2(Negative Samples). We assume that the data is linearly separable, i.e there is a separating hyperplane which separates the positive examples from the negative ones.

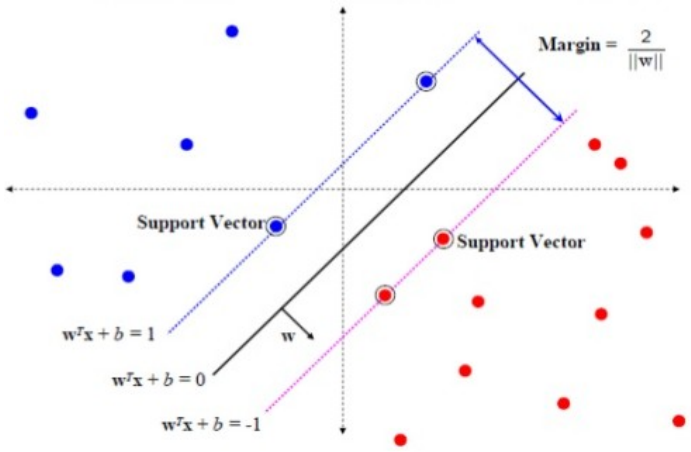

Figure 1: Hard Margin SVM [Source : Slideshare]

Thus, the goal is to find the optimal linear classifier (a hyperplane), such that it classifies every training example correctly, and maximizes the classification margin. The training data must satisfy the following constraints:

$$w^T x_i + b \geq 1, x_i \in C_1 \tag{13}$$

$$w^T x_i + b \leq -1, x_i \in C_2 \tag{14}$$

where $x_i \in \mathrm{R}^D$, $w \in \mathrm{R}^D$ and b $\in$R.

Let y_i be the variable containing the information about class i.e

$$y_i = 1, \forall x_i \in C_1 \tag{15}$$

$$y_i = -1, \forall x_i \in C_2 \tag{16}$$

which implies

$$y_i(w^T x_i + b) \geq 1, \forall i \tag{17}$$

Now consider the points for which quality holds in equation 13. These points lie on the hyperplane H1 : $w^T x_i + b = 1$ with normal w and perpendicular distance from the origin $\frac{|1-b|}{||w||}$. Similarly, the points for which the equality in equation 14 holds lie on the hyperplane H2 : $w^T x_i + b$ = -1, with normal again w, and perpendicular distance from the origin $\frac{|-1-b|}{||w||}$.

Hence, the perpendicular distance between two hyperplane or total margin is $\frac{2}{||w||}$. Since H1 and H2 are parallel and they have the same normal w, thus we can find the pair of hyperplanes which gives the maximum margin by minimizing $||w||^2$, subject to the above constraints. Hence, the optimization problem formulates to be:

$$\max \frac{2}{||w||} \tag{18}$$

$$\text{Subject to: } 1 - y_i(w^T x_i + b) \leq 0, \forall i \tag{19}$$

$$-x_i \leq 0, \forall i \tag{20}$$

or we can write,

$$\min \frac{||w||^2}{2} \tag{21}$$

$$\text{Subject to: } 1 - y_i(w^T x_i + b) \leq 0, \forall i \tag{22}$$

$$-x_i \leq 0, \forall i \tag{23}$$

The above minimization problem is convex, therefore there exists a unique global minimum value (when feasible), and there is a unique minimizer, i.e. w and b value that provides the minimum (given that the data is indeed linearly separable).It can be solved using standard Quadratic Programming (QP) optimization techniques. But, in order to extend this idea to nonlinear decision surfaces we will use the technique of Lagrange Multipliers.

4.1 Lagrangian Formulation

There are two main reasons behind the reformulation of the SVM optimization problem.

- Constraints will be replaced by constraints on the Lagrange multipliers themselves, which will be much easier to handle.
- Training data will appear in the form of dot products between vectors which allows us to generalize to the nonlinear case.

For the reformulation of the problem into Lagrangian, we introduce Lagrange Multipliers μ_i, i = 1,...,m, for every inequality constraint in *Equation* 17 which gives Lagrangian as:

$$\mathcal{L}(w, b, \mu) = \frac{||w||^2}{2} + \sum_{i=1}^{m} \mu_i [1 - y_i(w^T x_i + b)] \tag{24}$$

The above equation is Primal formulation of hard-margin SVM. Now, by applying KKT condition on *Equation* 24 we get following conditions:

- First Order Conditions:

$$\bigtriangledown_w \mathcal{L} = 0 \tag{25}$$

$$\Rightarrow w^* = \sum_{i=1}^{m} \mu_i^* y_i x_i \tag{26}$$

$$\bigtriangledown_b \mathcal{L} = 0 \tag{27}$$

$$\Rightarrow \sum_{i=1}^{m} \mu_i^* y_i = 0 \tag{28}$$

- Feasibility Conditions:

$$1 - y_i(w^{*T} x_i + b^*) \leq 0, \forall i \tag{29}$$

- Complementary Slackness Condition:

$$\mu_i^* [1 - y_i(w^{*T} x_i + b^*)] = 0, \forall i \tag{30}$$

- Dual Feasibility Condition:

$$\mu_i^* \geq 0, \forall i \tag{31}$$

Now, from the above KKT conditions we can infer following observations:

- If $\mu_i^* > 0$ then from Complementary Slackness Condition:

$$y_i(w^{*T} x_i + b^*) = 1 \tag{32}$$

  which shows that these point lies on the boundary of decision hyperplane.Thus, these points are called Support Vectors.

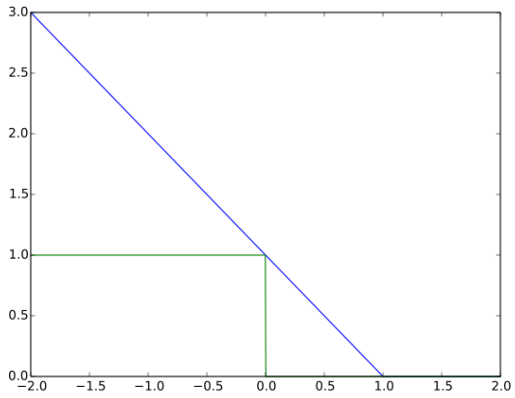

Figure 2: Hinge Loss(In Blue)[Source:Wikipedia]

- If $y_i(w^{*T}x_i + b^*) > 1$ then from Complementary Slackness Condition:

$$\mu_i^* = 0 \tag{33}$$

  which shows that these point doesn't lies on the boundary of decision hyperplane.Thus, are not Support Vectors as these point do not play any role in defining the decision hyperplane.

- From First Order Condition we can infer that:

$$\sum_{i \in C_1} \mu_i^* = \sum_{j \in C_2} \mu_j^* \tag{34}$$

4.2 Dual Form of Lagrangian Formulation

Now, by substituting First Oder Conditions in Primal Formulation of Lagrangian, we get the dual form:

$$\mathcal{L}(\mu) = \max_{\mu_1, \ldots \mu_m} \sum_{i=1}^{m} \mu_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \mu_i \mu_j y_i y_j (x_i^T x_j) \tag{35}$$

$$\text{s.t.} \sum_{i=1}^{m} \mu_i y_i = 0 \tag{36}$$

$$\mu_i \geq 0, \forall i \tag{37}$$

The above formulation is the dual form of Lagrangian for the SVM problem. Now, it is interesting to note that the dual form only depends on the class labels and the dot product of the data samples which allows SVM to be used in non-linear problems. More specifically, they are called kernel tricks.

4.3 Hinge Loss

In Machine Learning or Optimization , hinge loss is a loss function primarily used for training support vector machines (SVMs). Mathematically,the loss can be formulated for a particular data sample as :

$$\xi_i = \max[0, 1 - y_i(w^T x_i + b)] \tag{38}$$

where x_i is the data sample , y_i is the class label, b is the offset and ξ_i is the loss.

4.4 Soft Margin SVM

There are cases where two classes can not be separated linearly , we introduce a slight relaxation in the condition for finding optimal hyperplane by including an extra term ξ_i, due to which the SVM objective for the primal form becomes:

$$\min \frac{||w||^2}{2} + C\sum_{i=1}^{m} \xi_i \tag{39}$$

$$\text{Subject to: } 1 - \xi_i - y_i(w^T x_i + b) \leq 0, \forall i \tag{40}$$

$$-\xi_i \leq 0, \forall i \tag{41}$$

Here, C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. Small C tends to emphasize the margin while ignoring the outliers in the training data, while large C may tend to overfit the training data. When C approaches 0, all data points are free to be as large as possible to maximize the margin and when C approaches infinity, we end up with a hard-margin SVM classifier.

5 Refrences

1. A tutorial on support vector machines for pattern recognition, , Burges, C.J.C. and Burges, Chris J.C.,

Nonlinear Programming ,Dimitri Bertsekas