Introduction to Machine Learning
Prof. Sudeshna Sarkar
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Module - 5
Lecture - 22
SVM: The Dual Formulation

Good morning. We will now talk about Part C of this lecture, where we will look at The Dual Formulation of support vector machine.

(Refer Slide Time: 00:28)

In the last class, we looked at the formulation of the optimization problem corresponding to support vector machine where we have to minimize half w square, this is the convex quadratic optimization function subject to this linear constrains y i w T x i plus b greater than equal to 1 for all examples.

So, before we look at how to get the dual of this particular formulation, let us very briefly talk about Lagrangian Duality. Suppose, we take a general primal general problem and its primal formulation is given by, this is an optimization problem where you want to minimize f w, w are the parameters. You want to find values of w, so that is to minimize f w and you have a set of linear constraints.

There are two type of linear constraints, equality constraints - we have l equality constraints h i w equal to 0 and k inequality constraints g i w less than equal to 0, all these constraints are linear. Corresponding to this problem the generalized lagrangian is given by a function of w alpha beta, f w plus summation over 1 to k for all the number of non-linear constraints i equal to 1 to k alpha i g i w plus summation i equal to 1 to l beta i h i w, where h i w are the equality constraints, g i w are the inequality constraints. The alphas and betas are called Lagrange multipliers and the value of alpha is greater than equal to 0.

So, this is the lagrangian of this optimization function. Now, what we want to do is that, we want to; so this particular lagrangian if you take the values of a w alpha and beta such that if the primal constraints are not satisfied then the value of this lagrangian will be infinity. If the constraints are not satisfied the value of the lagrangian is infinity and it

will be equal to f w, if the constraints are satisfied right. So, we want to find out the values of alpha beta for which l; so if you look at the maximum of the value of l w alpha beta then if w satisfies the primal constraints, it will be equal to f w and it will infinity otherwise.

And we can rewrite the primal, as finding the value of this expression max of l w alpha beta, we want to find out those. So, we first do maximize keeping w fixed for a particular w, we can maximize over alpha beta and find the expression max l w alpha beta and max l w alpha beta is either f w or infinity and if we find the minimum of this, it will be giving us the solution of this. Because, otherwise it has a value of infinity, so the primal can be rewritten as, you take the lagrangian and you find out alpha beta for a fixed w you can take maximization over alpha beta and then you can do minimum over w. So, minimum over w maximum over alpha beta, L w alpha beta is a rewriting of the primal formulation.

(Refer Slide Time: 04:42)

Lagrangian Duality, cont.

The Primal Problem $p^* = \min_w \max_{\alpha,\beta,\alpha_i \geq 0} L(w,\alpha,\beta)$

The Dual Problem: $d^* = \max_{\alpha,\beta,\alpha_i \geq 0} \min_w L(w,\alpha,\beta)$

Theorem (weak duality):
$d^* = \max_{\alpha,\beta,\alpha_i \geq 0} \min_w L(w,\alpha,\beta) \leq \min_w \max_{\alpha,\beta,\alpha_i \geq 0} L(w,\alpha,\beta) = p^*$

Theorem (strong duality):
Iff there exist a saddle point of $L(w,\alpha,\beta)$, we have $d^* = p^*$

So, this is the primal problem and it has a solution let us say the solution is p star. P star is the solution of the primal problem - minimum over w maximum over alpha beta of the lagrangian and the dual problem is we are just putting the minimum here and maximum here. So, we take first, we take max of our alpha beta minimum over w, L w alpha beta

this is the dual formulation and it has the solution d star. So, this is the primal problem solution, this is the dual problem solution and we have two theorems. This theorem says that first of all if you change the order of max min and min max, it is general expression that - max over min of this expression, any expression is less than equal to min of max of this expression, right.

And d star is the max of min of this expression and therefore, d star is less than equal to p star. So, d star is always less than equal to p star. Now, if there exist a saddle point of this lagrangian where they are equal that is called the saddle point and that is the optimum value of both. So, the optimum value of the primal formulation and the optimum value of the dual formulation will be identical when there is a saddle point.

(Refer Slide Time: 06:32)

The KKT conditions

If there exists some saddle point of L, then it satisfies the following "Karush-Kuhn-Tucker" (KKT) conditions:

$$\frac{\partial}{\partial w_i} L(w, \alpha, \beta) = 0, \quad i = 1, \ldots, k$$

$$\frac{\partial}{\partial \beta_i} L(w, \alpha, \beta) = 0, \quad i = 1, \ldots, l$$

$$\alpha_i g_i(w) = 0, \quad i = 1, \ldots, m$$

$$g_i(w) \le 0, \quad i = 1, \ldots, m$$

$$\alpha_i \ge 0, \quad i = 1, \ldots, m$$

Theorem: If w^*, a^* and b^* satisfy the KKT condition, then it is also a solution to the primal and the dual problems.

And if the saddle point exists, then the saddle point satisfies the following condition called KKT condition or Karush-Kuhn-Tucker condition. Now, the condition says that the partial derivative of this lagrangian, with respect to w i and with respect to beta i will be equal to 0, according to the KKT condition. And from these two, you will find out that what you get is that alpha i g i w will be equal to 0, for i equal to 1 to m, g i w is less than equal to 0 and alpha i greater than equal to 0. So, these are the conditions that you get when the saddle point exists.

And the theorem says if w star, alpha star, and beta star, satisfy the KKT condition then it is also a solution to be primal and dual problems. With this brief description, brief outline of lagrangian duality let us go back to SVM and see how it can be applied there. The details of this theory are beyond the scope of this class and you can read some material on convex optimization if you want to learn more about this.

(Refer Slide Time: 07:59)

Now, if we look at our SVM formulation what we have is we have f w as half w square and we have the g w as y i w x plus b greater than equal to 1. We do not have the h, we do not have the equality constraints, we have only the objective function and the g i constraints. So, we are only dealing with alpha i, not the beta i. So, we are dealing with f w plus alpha i g i w.

So, this KKT conditions also says that this alpha i g i w equal to 0 and g i w is less than 0, it is because alpha i g i w is 0 only when alpha is 0 then g i w can be non-zero, and otherwise, g i w is . And which says that only the few of the alpha i's can be non-zero and the training data points whose alpha i's are non-zero are called the support vectors. So, some of the alpha is are non-zero and the training data corresponding to data support alpha is greater than equal to 0, if alpha i greater than 0, g i w will be equal to 0.

(Refer Slide Time: 09:45)

Solving the Optimization Problem

Quadratic programming with linear constraints

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2$$

$$\text{s.t.} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$$

Lagrangian Function

$$\text{minimize} \quad L_p(\mathbf{w}, b, \alpha_i) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n}\alpha_i\left(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1\right)$$

$$\text{s.t.} \quad \alpha_i \geq 0$$

Now, let us see the implication. So, this is the original optimization problem and when we take, this is the SVM optimization problem we take the lagrangian which gives us minimization of l w b alpha, we have written L p - p denotes primal. So, minimize L p w b alpha, half w square minus sigma alpha i y i w i x i plus b minus 1, subject to the constraints alpha i greater than equal to 0. This is by getting the lagrangian of the optimization problem.

Now, if we take the partial derivative of this L p with respect to w and b what we get here is, w equal to sigma alpha i y i x i, and from the second one by taking partial derivative of L p with respect to b and setting it to 0 we get sigma alpha i y i equal to 0.

So, what it means is that if I substitute this value of w in this expression here, if I substitute this value of w sigma alpha i y i x i in this expression here what I get is L p w b alpha equal to sigma alpha i minus sigma alpha i alpha j y i x T x j. I am sorry. So, we put w is here, so this w becomes half of alpha i alpha j y i y j x T x j minus b sigma alpha i y i. So, this is my L p when I substitute this value of w. But we know that sigma alpha i y i equal to 0 from this constraint. So, this expression on the rights side can be ignored and finally, we get L w b alpha as this expression.

So, L w b alpha is sigma alpha i minus half of i j equal to 1 to m, alpha i alpha j y i y j x i transpose x j. Now, this very important formulation and we will look at the properties of this formulation to get certain properties of the support vector machine algorithm.

The Dual problem

Now we have the following dual opt problem:

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

$$\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, k$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0.$$

This is a **quadratic programming** problem.
  – A global maximum of α_i can always be found.

So, the Dual problem, before we go to that let us look at the dual problem the dual problem is maximizing of j alpha where j alpha is the expression we saw earlier and these are the constraints alpha i greater than equal to 0 and sigma alpha i y i equal to 0. This is the dual problem, which is a quadratic programming problem and from this quadratic programming problem we can solve and find the global maximum value of alpha i. We can find out the values of alpha i, by solving this quadratic programming problem.

Support vector machines

- Once we have the Lagrange multipliers $\{\alpha_j\}$ we can reconstruct the parameter vector w as a weighted combination of the training examples:

$$w = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \qquad w = \sum_{i \in SV} \alpha_i y_i \mathbf{x}_i$$

- For testing with a new data z
  - Compute

$$w^T z + b = \sum_{i \in SV} \alpha_i y_i \left(\mathbf{x}_i^T z\right) + b$$

  and classify z as class 1 if the sum is positive, and class 2 otherwise

  Note: w need not be formed explicitly

And this quadratic programming problem is much easier to solve than the primal formulation. This is much simpler, because the constraints are simpler and we will see it has certain nice properties.

So, once we solve and get the lagrange multipliers alpha, we can reconstruct the parameter vectors. We can find w as sigma alpha i y i x i. In fact, we noted that alpha i is non-zero only for few of the examples. Those examples are the one, once which are the support vectors. So, w is obtained from sigma alpha i y i x i where i ranges among the support vectors and usually the support vectors are few in number and w can computed from the coordinates of those support vectors.

Also when we get a new data point z in order to find out the output corresponding to this we can compute w T z plus b, which is alpha sigma alpha i y i x i T z plus b and we classify z as class 1 if the sum is positive and class 2 if otherwise. Now, you note that w need not be found explicitly we can just use this expression and this expression has a very nice property when you put z what you are doing is - this alpha i, this is y i, this is x i T z. So, you are taking the dot product of the support vector with your test point.

(Refer Slide Time: 14:59)

Solving the Optimization Problem

- The discriminant function is:

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = \sum_{i \in SV} \alpha_i \mathbf{x}_i^T \mathbf{x} + b$$

- It relies on a *dot product* between the test point *x* and the support vectors $\mathbf{x}_i$
- Solving the optimization problem involved computing the dot products $\mathbf{x}_i^T \mathbf{x}_j$ between all pairs of training points
- The optimal $\mathbf{w}$ is a linear combination of a small number of data points.

So, the discriminant function is given by this dot product of x i T and x right. So, the computation reduces to mainly finding these dot products. So, you have the dot product between the test point x and the support vector x i. Why is this, such an exciting thought? Now, x i is a vector and this can be a high dimensional vector, if you take the dot product of these two linear vectors what you get is a scalar. So, the dot product is a scalar value.

And we will look at what are the implications later, when we solve the optimization problem also, if we look at this formulation where we solve the optimization problem. In here you see what we have is the dot product of the training points, so alpha i alpha j y i y j is either plus 1 minus 1. So, these are very simple to compute multiply and x i T x j the dot product of x i x j. So, when we solve the optimization problems it involved computing the dot products between all the pairs of training points and the optimal w is linear combination of a small number of data points. So, these are some of the features about this SVM formulation.

We stop here today, in the next class we will look at certain properties of SVM and how these properties can be used for those formulations of SVM. With this I end today's lecture.

Thank you.