-W2-SVM

asd135441

May 2019

Abstract

In this week's group discussion, Yao Xingcheng delivered the lecture naming *Supporting Vector Machine* systematically, through which we got to learn the fundamental concept and the methods to train the model which based on the Lagrange Duality and optimization problem. Also, we learned the Kernel Function and the training algorithm of SVM.

1 Supporting Vector Machine

Supporting vector machine, first proposed before 1980, was originally used as a linear classifier. In 1992, Bernhard E.Boser, Isabelle M. Guyon and Vladimir N. Vapnik applied the kernel trick to find the maximum-margin hyperplanes which can be used as nonlinear classifiers. The current standard incarnation of SVM was proposed in 1993.

1.1 Optimization Goal

What we are dealing with here is a classification problem that we need to learn a linear function

$$g(x) = w^T + b$$

based on the given training set with labels.

Here, we simply discussed the dual classification and use a sign function to indicate the classification.

1.1.1 Definitions

• functional margin:

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x + b)$$

• geometric margin:

$$\gamma^{(i)} = y^{(i)}((\frac{w}{||w||})^T x + \frac{b}{||w||})$$

• margin for a training set:

$$\gamma = min_{i=1,\ldots,m}\gamma^{(i)}$$

Here, we should pay attention to difference and meaning of the functional margin and geometric margin. **Functional margin** can define whether the classification is correct or not based on the sign, while the numerical value is meaningless. **Geometric margin** gets rid of the influence caused by the value and carrys a geometric meaning which indicates the distance between the datapoint and the hyperplane. Specially, if we set $\hat{\gamma}$ with a constraint, we can always find corresponding w and b with same proportion to define the same hyperplane.

1.1.2 Lagrange Duality

When solving optimization problems with constraints, lagrange duality is always used. If the primal optimization problem is as follow:

$$min_{x \in R^n} f(x)$$
$$s.t. g_i(x) \geq 0, i = 1, ..., k$$
$$h_j(x) = 0, j = 1, ..., l$$

We define the generalized Lagrangian:

$$L(x, \alpha, \beta) = f(x) - \sum_{i=1}^{k} \alpha_i g_i(x) - \sum_{j=1}^{l} \beta_j h_j(x), s.t. \alpha_i \geq 0$$

With some analyzation, we can see that

$$q^* = max_{\alpha, \beta : \alpha_i \geq 0} min_x L(x, \alpha, \beta) \leq min_x max_{\alpha, \beta : \alpha_i \geq 0} L(x, \alpha, \beta) = p^*$$

in which, p^* is the primal problem.
•KKT conditions: under certain conditions, we will have

$$p^* = q^*$$

Under the ideal assumptions, there must exist x*, α* and β* so that x* is the solution to the primal problem, α* and β* are the solution to the dual problem and p^*=q^*=L(x*,α*,β*).

- $$\frac{\partial L(x*, \alpha*, \beta*)}{\partial x_i} = 0, i \in [1, N]$$
- $$\frac{\partial L(x*, \alpha*, \beta*)}{\partial \alpha_i} = 0, i \in [1, k]$$
- $$\frac{\partial L(x*, \alpha*, \beta*)}{\partial \beta_i} = 0, i \in [1, l]$$
- $$\alpha_i^* g_i(x^*) = 0, i \in [1, k]$$
- $$g_i(x^*) \geq 0, i \in [1, k]$$
- $$\alpha_i^* \geq 0, i \in [1, k]$$

•KKT dual complementarity condition: if α*¿0, then g_i(x)=0.

1.1.3 Maximization problem

For a given dataset, there are infinite boundaries to achieve the correct classification. We would like to find the hyperplane with the biggest margin, so that we can be more sure to classify the data correctly.
We want to adjust w and b to maximize γ, for every datapoint, there is,

$$y^{(i)}(w^T x + b) \geq \gamma, i = 1, ...m, ||w|| = 1$$

With the explanation above, we can make some transformation and add a constraint

$$\hat{\gamma} = 1,$$

and changing the maximization problem into a minimization problem:

$$min_{w,b} \frac{1}{2} ||w||^2 s.t. y^{(i)}(w^T x + b) \geq 1, i = 1, ..., m$$

In Lagrangian form, it will be,

$$L(w, b, \alpha) = \frac{||w||^2}{2} - \sum_{i=0}^{N} \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1], s.t. \alpha_i \geq 0, 1 \leq i \leq N$$

•Since we have transformed the minimization problem to the lagragian form, our problem becomes

$$min_{w,b} max_{\alpha} L(w, b, \alpha)$$

According to the Lagrange Duality discussed above, we can have the dual problem

$$max_{\alpha} min_{w,b} L(w, b, \alpha)$$

By applying the KKT conditions, we have

$$\frac{\partial L(w, b, \alpha)}{\partial w} = 0$$

$$\frac{\partial L(w, b, \alpha)}{\partial b} = 0$$

which are

$$w = \sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)}$$

$$\sum_{i=1}^{N} \alpha_i y^{(i)} = 0$$

After subsituting these outcomes back to the lagrangian equation, we have a function only related to vectorα,

$$L(\alpha) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)} x^{(j)} + \sum_{i=1}^{N} \alpha_i$$

This is a quadratic programming problem which can be solved by a program and a global maximum of α_i can always be found.
Many of the α_i may be 0, therefore, x_i with non-zero α_i are called **support vectors** because they determine the decision boundary according to KKT dual complementarity condition.
Based on the model trained above, we can predict a new sample x by calculating the value of $w^Tx + b$.

$$w^Tx + b = (\sum_{i=1}^{N} \alpha_i y^{(i)} x^{(i)})^T x + b$$

classify x as class +1 if the sum is positive, and class -1 otherwise.

1.1.4 Non LInearly Separable Case

In some cases, we allow error in classification, and record the errors with ξ_i. The optimization problem becomes:

$$min_{w,b,\xi} \frac{||w||^2}{2} + C\sum_{i=1}^{N} \xi_i,$$

$$s.t. y^{(i)}(w^Tx^{(i)} + b) \geq 1 - \xi_i, \xi_i \geq 0, 1 \leq i \leq N$$

Similarly, we have the lagragian form and dealing with it by applying the Lagrange Duality.
Specificlly, we can pay attention to the meanings of the values of α.
• α_i=0 means $y^{(i)}(w^Tx^{(i)} + b) \geq 1$
• α_i=C means $y^{(i)}(w^Tx^{(i)} + b) \leq 1$
• 0¡α_i¡C means $y^{(i)}(w^Tx^{(i)} + b) = 1$

1.2 Kernel Function

When we meet a case where the datapoints cannot be linearly classified, we apply kernel function to transform x_i to a higher dimensional space to achieve the linear classification. **Input Space** is the space the point x_i is located and **Feature Space** is the space of f(x_i) after transformation.

$$x \to \Phi(x)$$

$$K(x, z) = \Phi(x)^T\Phi(z)$$

From the analyzation of the optimization problem above, we can easily notice that what matters is the inner product ¡x,z¿ instead of the individual feature vectors. Therefore, as long as we can define and calculate the inner product in the feature space, we can train the model and make predictions by changing all inner products to kernel functions.

$$f(z) = \sum_{x_i \in S} \alpha_i y_i K(z, x_i) + b$$

S: the set of support vectors
Notice: Not all similarity measure can be used as kernel function.