The Report of the Exercice SVM and duality

Zuoyu Zhang

February 06 2023

1 Introduction to the problem

In machine learning, a support vector machine, often abbreviated as SVM, aka support vector network) is a supervised learning model for analyzing data in classification and regression analysis with associated learning algorithms. Given a set of training instances, each labeled as belonging to one or the other of two classes, the SVM training algorithm creates a model that assigns new instances to one of the two classes, making it a non-probabilistic binary linear classifier.

SVM techniques rely on two main ideas: classify data by (i)searching for a separation boundary between two classes such that labelled data exhibit largest minimum distance to the boundary and (ii) transform data into a space of larger, possibly infinite, dimension where linear separation of classes works better (but then in the original classes separation becomes nonlinear).

In SVM, the two classes are noted as $c \in \{-1, 1\}$. The linear separation can simly achieved by looking for an hyperplane with parameters $(\mathbf{w}, b)$ such that in the training set $\{(\mathbf{x}_k, c_k)\}_{k=1:K}$ elements satisfy $c_k(\mathbf{w}_T x_k - b) \geq 0$. We can also search for two hyper planes such that $\mathbf{w}^T\mathbf{x}_k - b \leq -1$ for class -1 and $\mathbf{w}^T\mathbf{x}_k - b \geq 1$ for class 1. We would like that they represent limiting hyper planes for both classes with maximum separation.

2 Answer the questions

1) Assuming the hyper planes $\mathbf{w}^T\mathbf{x}_k - b \leq -1$ and $\mathbf{w}^T\mathbf{x}_k - b \geq 1$ are limiting hyperplanes for classes $c = -1$ and $c = 1$ respectively, prove that the maximum distance between these class limiting hyper planes is $2||\mathbf{w}||^{-1}$, that is,

$$\underset{c_k=1, c_l=-1}{max} ||\mathbf{x}_k - \mathbf{x}_l||^2 = \frac{2}{||\mathbf{w}||}$$

Then, check that we can get $(\mathbf{w}, b)$ by solving

$$\begin{cases} min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2, \\ c_k(\mathbf{w}^T\mathbf{x}_k - b) \geq 1, \quad k = 1 : K \end{cases} \tag{1}$$

This way of determining w enables maximum margin between class limiting hyper planes $\mathbf{w}^T\mathbf{x}_k - b \leq -1$ and $\mathbf{w}_T\mathbf{x}_k - b \geq 1$.

First, we can express the hyperplane by the following formulation:

$$\mathbf{w}^T\mathbf{x}_k - b = 0$$

$\mathbf{w}$ is the normal vector to determine the direction of the hyperplane and $\mathbf{w} = (w_1, w_2, ..., w_m)$ and d is the intercept.

The geometric distance of all sample points in the training set to the hyperplane can be defined as:

$$\gamma_k = \frac{|(\mathbf{w}^T\mathbf{x}_k - b)|}{||\mathbf{w}||} = \frac{y_k(\mathbf{w}^T\mathbf{x}_k - b)}{||\mathbf{w}||}$$

and define the minimum value in the interval of the function from the sample points in the training set to the hyperplane as:

$$\gamma = \min_{k=1,2,...,K} \gamma_k = \frac{1}{||\mathbf{w}||}$$

As we have the two classes here, the gap is the sum of the minimum distance from the two class sample points to the hyperplane, so the maximum distance between these class limiting hyperplane is $\frac{2}{||\mathbf{w}||}$.

2) When both classes are not exactly linearly separable, a possible linear approximation is obtained by weighting the objective $||\mathbf{w}||^2$ by the sum of distances to the class boundary for wrongly classified data. Explain how this can be expressed by the following problem by discussing the meaning of variables ζ_k and parameter λ:

$$\begin{cases} min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_{k=1}^{K} \zeta_k, \\ c_k(\mathbf{w}^T\mathbf{x}_k - b) \geq 1 - \zeta_k, \qquad k = 1 : K, \zeta_k \geq 0. \end{cases} \tag{2}$$

In the SVM formula expressed earlier, all the samples must be divided correctly, and this interval is called hard margin. In reality, there may be some sample points that cannot be linearly divisible, so we will allow some sample points to not satisfy the constraint $c_k(\mathbf{w}^T\mathbf{x}_k - b) \geq 1$ and maximize the interval while making the number of samples that do not satisfy the constraint as minimum as possible, so the optimization objective of the SVM soft interval can be expressed as

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_{k=1}^{K} l(c_k(\mathbf{w}^T\mathbf{x}_k - b) - 1)$$

Moreover, we can introduce slack variables to rewrite this equation

$$\min_{\mathbf{w},b,\boldsymbol{\zeta}} \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_{k=1}^{K} \zeta_k$$

Where λ is called the penalty parameter. When λ is set larger, it means that the penalty for outliers is larger and eventually fewer points will cross the interval boundary and the model will become complex. The smaller λ is set, the more points will cross the interval boundary, and the model will be smoother. So, we can say when λ is infinite, the optimization objective requires that all samples satisfy the constraint, i.e., the initial hard margin, and allows some samples not to satisfy the constraint when λ takes a finite value.

where l is the loss function and it can be described by slack variable, where each sample has a corresponding slack variable to characterize the extent to which the constraint is not satisfied by that sample, we can surely use the sum of distances to the class boundary for wrongly classified data as the loss function and try to minimize it.

3) Write the Lagrangian of the problem and express KKT conditions.

The Lagrangian of the problem can be expressed as:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\zeta}, \boldsymbol{\mu}) = \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_{k=1}^{K} \zeta_k + \sum_{k=1}^{K} \alpha_k (1 - \zeta_k - c_k(\mathbf{w}^T \mathbf{x}_k - b)) - \sum_{k=1}^{K} \mu_k \zeta_k$$

where α_k, μ_k are the Lagrangian multipliers.

Meanwhile, we can obtain the KKT conditions:

$$\begin{cases} \frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{k=1}^{K} \alpha_k c_k \mathbf{x}_k = 0, \\ \frac{\partial L}{\partial b} = \sum_{k=1}^{K} \alpha_k c_k = 0, \\ \frac{\partial L}{\partial \zeta_k} = \lambda - \alpha_k - \mu_k = 0, \\ \alpha_k \geq 0, \mu_k \geq 0, \\ c_k(\mathbf{w}^T \mathbf{x}_k - b) - 1 + \zeta_k \geq 0, \\ \alpha_k(c_k(\mathbf{w}^T \mathbf{x}_k - b) - 1 + \zeta_k) = 0, \\ \zeta_k \geq 0, \mu_k \zeta_k = 0. \end{cases} \tag{3}$$

4) From KKT conditions, show that the dual of the problem writes

$$\begin{cases} max_\alpha f(\alpha) = \sum_{1:K} \alpha_k - \frac{1}{2} \sum_{i,j=1:K} \alpha_i \alpha_j c_i c_j (\mathbf{x}_j^T \mathbf{x}_i), \\ \sum_{k=1:K} \alpha_k c_k = 0, \\ 0 \leq \alpha_k \leq \lambda, \qquad\qquad k = 1 : K. \end{cases} \tag{4}$$

What do variables $(\alpha_k)_k = 1 : K$ represent?

We substitute the results of the above KKT conditions for the three variables to derive zero into the Lagrangian function to obtain the pairwise problem expression for the soft interval SVM problem.

We already have :

$$\mathbf{w} = \sum_{k=1}^{K} \alpha_k c_k \mathbf{x}_k$$

$$\sum_{k=1}^{K} \alpha_k c_k = 0$$

$$\lambda = \alpha_k + \mu_k$$

Then, substitute these equations into the above expressions for the Lagrangian functions

$$\begin{aligned}
L(\mathbf{w}, b, \boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\zeta}) &= \frac{1}{2}||\mathbf{w}||^2 + \lambda \sum_{k=1}^{K} \zeta_k + \sum_{k=1}^{K} \alpha_k (1 - \zeta_k - c_k(\mathbf{w}^T \mathbf{x}_k - b)) - \sum_{k=1}^{K} \mu_k \zeta_k \\
&= \frac{1}{2}||\mathbf{w}||^2 + \sum_{k=1}^{K} \alpha_k (1 - \zeta_k - c_k(\mathbf{w}^T \mathbf{x}_k - b)) + \sum_{k=1}^{K} \alpha_k \zeta_k \\
&= \frac{1}{2}\mathbf{w}^T \mathbf{w} - \sum_{k=1}^{K} \alpha_k c_k (\mathbf{w}^T \mathbf{x}_k - b) + \sum_{k=1}^{K} \alpha_k \\
&= -\frac{1}{2}\mathbf{w}^T \mathbf{w} + \sum_{k=1}^{K} \alpha_k + \sum_{k=1}^{K} \alpha_k c_k b \\
&= -\frac{1}{2} \sum_{k=1}^{K} \alpha_k c_k x_k^T \sum_{k=1}^{K} \alpha_k c_k \mathbf{x}_k + \sum_{k=1}^{K} \alpha_k \\
&= \sum_{k=1}^{K} \alpha_k - \frac{1}{2} \sum_{i,j=1:K} \alpha_i \alpha_j c_i c_j (\mathbf{x}_j^T \mathbf{x}_i)
\end{aligned}$$

Then we can get the dual problem of the soft margin SVM problem.

Each sample point corresponds to an α, where the α's represent only the relationship between the corresponding sample point and the maximum interval boundary, as specified by the following interpretation.

- if $\alpha_k = 0$, then $c_k(\mathbf{w}^T \mathbf{x}_k - b) - 1 \geq 0$, these sample points are far away from the spacing Boundary and have no influence on the model.
- if $0 < \alpha_k < \lambda$, then $\mu_k > 0$, and thus we have $\zeta_k = 0$, the sample is exactly on the maximum interval boundary.
- if $\alpha_k = \lambda$, then $\mu_k = 0$, thus we have if $0 \leq \zeta_k \leq 1$, the sample has been correctly classified, but is between the hyperplane and the interval boundary of its own category, if $\zeta_k > 1$, the sample is wrongly classified.

5) Check that the solution of the primal for $\mathbf{w}$ is then given by $\mathbf{w} = \sum_k \alpha_k c_k \mathbf{x}_k$. To find b, check first that if $0 < \alpha_k < \lambda$ then $\mathbf{x}_k$ lies on the margin. Then , $c_k(\mathbf{w}^T \mathbf{x}_k - b) = 1$, that is, $b = \mathbf{w}^T \mathbf{x}_k - c_k$.

In SVM, the objective function is convex and the constraints are linear, so it is a convex optimization problem and therefore a strong dyadic problem. That means the maximum value of the dual problem is the minimum value

of the original problem, so the the solution of the primal for $\mathbf{w}$ is then given by $\mathbf{w} = \sum_k \alpha_k c_k \mathbf{x}_k$.

As we have mentioned before, if $0 < \alpha_k < \lambda$, then $\mu_k > 0$, and thus we have $\zeta_k = 0$, the sample is exactly on the maximum interval boundary. And we say these samples are support vectors, then we can simply calculate the intercept as $b = \mathbf{w}^T \mathbf{x}_k - c_k$