SVM Notebook

Ruoqi Wei

- Three important points in Statistical Learning :

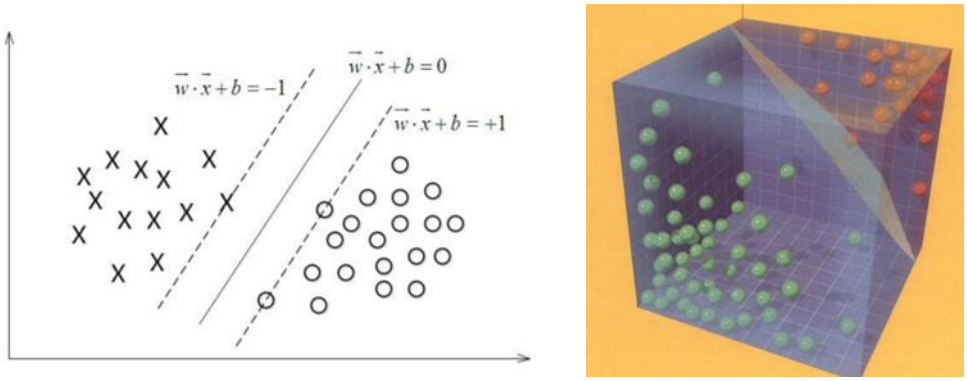
Model: Probability distribution or function. There are usually many models in the model's hypothesis space.
Strategy: How to choose the optimal model from the hypothesis space (such as the loss function)
Algorithm: The solution to the optimal model.

- Support Vector Machine (SVM)

Model: Two-category model (conditional probability distribution).
Strategy: Find a hyper-plane in the data space, this hyperplane has the largest margin from the nearest data belonging to the two categories.
Algorithm: An optimization algorithm for solving convex quadratic programming.

- What are the differences between linear SVM and nonlinear SVM?

When the training data is linearly separable, the linear SVM is learned by margin maximization.
When the training data is nonlinear data, the data is mapped to the high-dimensional feature space by kernel function first ,and then SVM is learned by margin maximization.



Linear SVM and Nonlinear SVM

Definition 1 **linear separable support vector machine**

Given a linearly separable training data set, the hyperplane obtained by solving the corresponding convex quadratic programming problem by margin maximization can be expressed as:

$$w^* \cdot x + b^* = 0 \tag{1}$$

And the Classification Decision function can be expressed as:

$$f(x) = \text{sign}(w^* \cdot x + b^*) \tag{2}$$

Is called Linear separable support vector machine.

Definition 2 Functional margin

For a given training data set T and hyperplane (w, b), define the hyperplane (w, b) function margin for the sample point (x_i, y_i) as

$$\hat{\gamma}_i = y_i(w \cdot x_i + b) \tag{3}$$

Definition 3 Geometrical Margin:

For a given training data set T and hyperplane (w, b), define the hyperplane (w, b) Geometrical margin for the sample point (x_i, y_i) as

$$\gamma_i = y_i\left(\frac{w}{\|w\|} \cdot x_i + \frac{b}{\|w\|}\right) \tag{4}$$

That is,the geometric Margin is the function Margin divided by L2 norm.

(Note: What is the L2 norm?

The L2 norm is our most common and most commonly used norm. The most metric distance we use is the L2 norm, which is defined as follows:

$$||x||_2 = \sqrt{\sum_i x_i^2}$$

)

From expression (3) and (4),we can see that The function margin and geometric margin have the following relationship

$$\gamma = \frac{\hat{\gamma}}{\|w\|} \tag{5}$$

Now let's consider how to find a separate hyperplane with the largest Geometrical margin. Specifically, this problem can be expressed as the following constraint optimization problem

$$\begin{aligned} &\max_{w,b} \quad \gamma \\ &\text{s.t.} \quad y_i\left(\frac{w}{\|w\|} \cdot x_i + \frac{b}{\|w\|}\right) \geqslant \gamma, \quad i = 1,2,\cdots,N \end{aligned} \tag{6}$$

Considering (5), this problem can be rewritten as

$$\max_{w,b} \quad \frac{\hat{\gamma}}{\|w\|}$$
$$\text{s.t.} \quad y_i(w \cdot x_i + b) \geqslant \hat{\gamma}, \quad i = 1,2,\cdots,N \tag{7}$$

Since maximization $\frac{1}{||w||}$ is equivalent to minimization$\frac{1}{2}||w||^2$, the optimization problem of SVM can be obtained as follows

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2$$
$$\text{s.t.} \quad y_i(w \cdot x_i + b) - 1 \geqslant 0, \quad i = 1,2,\cdots,N \tag{8}$$

In summary, the Maximum Margin Classifier algorithm can be described as follows:

Algorithm 1 Maximum Margin Classifier
Input: Linear separable training data set T={x1,y1), (x2,y2),...(xn,yn)}
Output: Maximum interval separation hyperplane and classification decision function
step
(1) Construct and solve the constraint optimization problem:

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2$$
$$\text{s.t.} \quad y_i(w \cdot x_i + b) - 1 \geqslant 0, \quad i = 1,2,\cdots,N$$

Find the optimal solution w, b.
(2) The resulting hyperplane is thus obtained:

$$w^* \cdot x + b^* = 0$$

And the Classification decision function:

$$f(x) = \text{sign}(w^* \cdot x + b^*)$$

Next, to make the Maximum Margin Classifier,we need to solve the constraint optimization problem(8)
How can we do that?
The answer is by Lagrange Duality.

Definition 4 Lagrange Duality

By solving the dual problem equivalent to the original problem, the optimal solution of the original problem is obtained. Format: Add a Lagrange multiplier to each inequality constraint.

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{N}\alpha_i y_i(w\cdot x_i+b)+\sum_{i=1}^{N}\alpha_i$$

(9)

According to Lagrange duality, the dual problem of the original problem is to find the Maximum and Minimum.

Therefore, in order to get the solution to the dual problem, we need to find L(w,b,a) to the minimum of w and b, and then find the maximum a .

(1) Find L(w,b,a) to the minimum of w,b

Method: Let L find the partial derivative value for w and b respectively, and let it = 0.

$$\nabla_w L(w,b,\alpha)=w-\sum_{i=1}^{N}\alpha_i y_i x_i=0$$

$$\nabla_b L(w,b,\alpha)=\sum_{i=1}^{N}\alpha_i y_i=0$$

$$w=\sum_{i=1}^{N}\alpha_i y_i x_i$$

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

Next,substituting the above results into the previous L:

$$\mathcal{L}(\mathrm{w},\mathrm{b},\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{m}\alpha_i\left[y^{(i)}\left(w^T x^{(i)}+b\right)-1\right]$$

$$=\frac{1}{2}\mathrm{w}^T\mathrm{w}-\sum_{i=1}^{m}\alpha_i y^{(i)}w^T x^{(i)}-\sum_{i=1}^{m}\alpha_i y^{(i)}b+\sum_{i=1}^{m}\alpha_i$$

$$=\frac{1}{2}\mathrm{w}^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}-\sum_{i=1}^{m}\alpha_i y^{(i)}w^T x^{(i)}-\sum_{i=1}^{m}\alpha_i y^{(i)}b+\sum_{i=1}^{m}\alpha_i$$

$$=\frac{1}{2}\mathrm{w}^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}-w^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}-\sum_{i=1}^{m}\alpha_i y^{(i)}b+\sum_{i=1}^{m}\alpha_i$$

$$=-\frac{1}{2}\mathrm{w}^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}-\sum_{i=1}^{m}\alpha_i y^{(i)}b+\sum_{i=1}^{m}\alpha_i$$

$$=-\frac{1}{2}\mathrm{w}^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}-b\sum_{i=1}^{m}\alpha_i y^{(i)}+\sum_{i=1}^{m}\alpha_i$$

$$=-\frac{1}{2}\left(\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}\right)^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}-b\sum_{i=1}^{m}\alpha_i y^{(i)}+\sum_{i=1}^{m}\alpha_i$$

$$=-\frac{1}{2}\sum_{i=1}^{m}\alpha_i y^{(i)}\left(x^{(i)}\right)^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}-b\sum_{i=1}^{m}\alpha_i y^{(i)}+\sum_{i=1}^{m}\alpha_i$$

$$=-\frac{1}{2}\sum_{i=1,j=1}^{m}\alpha_i y^{(i)}\left(x^{(i)}\right)^T\alpha_j y^{(j)}x^{(j)}-b\sum_{i=1}^{m}\alpha_i y^{(i)}+\sum_{i=1}^{m}\alpha_i$$

Finally we got

$$\min_{w,b}\ L(w,b,\alpha)=-\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)+\sum_{i=1}^{N}\alpha_i$$

(2) Find the maximum of L for a, that is, the optimization problem for the dual problem.

$$\max_{\alpha}\ -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)+\sum_{i=1}^{N}\alpha_i$$

$$\text{s.t.}\quad \sum_{i=1}^{N}\alpha_i y_i=0$$

$$\alpha_i\geqslant 0,\quad i=1,2,\cdots,N$$

If the KTT* condition(See Appendix) is true,

$$\nabla_w L(w^*, b^*, \alpha^*) = w^* - \sum_{i=1}^{N} \alpha_i^* y_i x_i = 0$$

$$\nabla_b L(w^*, b^*, \alpha^*) = -\sum_{i=1}^{N} \alpha_i^* y_i = 0$$

$$\alpha_i^* (y_i (w^* \cdot x_i + b^*) - 1) = 0, \quad i = 1, 2, \cdots, N$$

$$y_i (w^* \cdot x_i + b^*) - 1 \geqslant 0, \quad i = 1, 2, \cdots, N$$

$$\alpha_i^* \geqslant 0, \quad i = 1, 2, \cdots, N$$

Thus,we can get

$$w^* = \sum_i \alpha_i^* y_i x_i$$

$$b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i (x_i \cdot x_j)$$

Thus,the separation hyperplane can be express as

$$\sum_{i=1}^{N} \alpha_i^* y_i (x \cdot x_i) + b^* = 0$$

And the Classification decision function:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i^* y_i (x \cdot x_i) + b^* \right)$$

(3) After obtaining L(w, b, a) for the minimization of w and b, and the maximum of the pair, the last step can use the SMO algorithm to solve the Lagrangian multiplier in the dual problem.

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

$$s.t. \quad \alpha_i \geq 0, i = 1, ..., n$$

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

Is equivalent to solving:

$$\min_{\alpha} \Psi(\vec{\alpha}) = \min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j K(x_i, x_j) \alpha_i \alpha_j - \sum_{i=1}^{n} \alpha_i$$

$$s.t. \quad 0 \leq \alpha_i \leq C, i = 1, ..., n$$

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

In 1998, John C. Platt of Microsoft Research proposed a solution to the above problem in the paper "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines": SMO algorithm, which quickly became the fastest quadratic programming optimization algorithm, especially when it comes to linear SVM and data sparse performance.

So far, The SVM is still weak and can only deal with linear datasets. Next i will study kernel functions and then generalize them to nonlinear classification problems.

To be continue

Appendix

Karush–Kuhn–Tucker conditions

Consider the following nonlinear minimization or maximization problem:

Optimize $f(x)$

subject to

$$g_i(x) \leq 0,$$
$$h_j(x) = 0,$$

where x is the optimization variable, f is the objective or utility function, $g_i \ (i = 1, \ldots, m)$ are the inequality constraint functions, and $h_j \ (j = 1, \ldots, \ell)$ are the equality constraint functions. The numbers of inequality and equality constraints are denoted m and ℓ, respectively.

Suppose that the objective function $f : \mathbb{R}^n \to \mathbb{R}$ and the constraint functions $g_i : \mathbb{R}^n \to \mathbb{R}$ and $h_j : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable at a point x^*. If x^* is a local optimum and the optimization problem satisfies some regularity conditions (see below), then there exist constants $\mu_i \ (i = 1, \ldots, m)$ and $\lambda_j \ (j = 1, \ldots, \ell)$, called KKT multipliers, such that

Stationarity

For maximizing $f(x)$: $\nabla f(x^*) = \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{j=1}^{\ell} \lambda_j \nabla h_j(x^*),$

For minimizing $f(x)$: $-\nabla f(x^*) = \sum_{i=1}^{m} \mu_i \nabla g_i(x^*) + \sum_{j=1}^{\ell} \lambda_j \nabla h_j(x^*),$

Primal feasibility

$$g_i(x^*) \leq 0, \text{ for } i = 1, \ldots, m$$
$$h_j(x^*) = 0, \text{ for } j = 1, \ldots, \ell$$

Dual feasibility

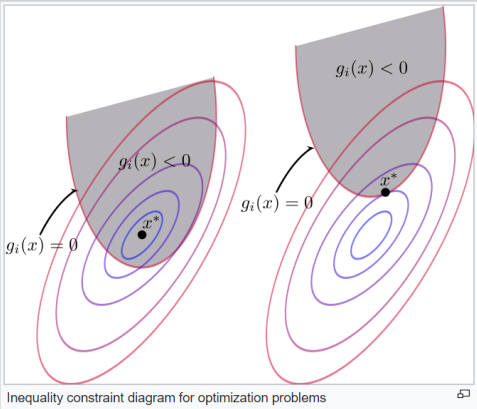
$$\mu_i \geq 0, \text{ for } i = 1, \ldots, m$$

Complementary slackness

$$\mu_i g_i(x^*) = 0, \text{ for } i = 1, \ldots, m.$$

In the particular case $m = 0$, i.e., when there are no inequality constraints, the KKT conditions turn into the Lagrange conditions, and the KKT multipliers are called Lagrange multipliers.

If some of the functions are non-differentiable, subdifferential versions of Karush–Kuhn–Tucker (KKT) conditions are available.[5]



Inequality constraint diagram for optimization problems

References:

- Cristianini, N., & Shawe-Taylor, J. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.
- Hang, L. (2012). Statistical learning method. Beijing: Tsinghua University Press, 2012, 80-87.
- http://web.mit.edu/6.034/wwwbob/svm.pdf
- Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.