One-class Support Vector Machines

Chunpai Wang

March 13, 2018

1 Problem

Given some dataset drawn from an underlying probability distribution P, we want to estimate a subset S of input space such that the probability that a test point drawn from P lies outside of S equals some a priori specified v between 0 and 1. In other words, we would like to know if the test point is similar(or close) to S or not, and quantify the similarity with probability. A more simple example is, we only have training data of one class, and the goal is to test new data and find out whether it is alike or not like the training data.

Question: supervised learning or unsupervised learning ? Think of solving this problem in your own way first.

2 Ideas

- We would like a function that takes the value $+1$ in a "small" region capturing most of the data points, and -1 elsewhere.
- The strategy is to map the data into the feature space corresponding to the kernel, and to separate them from the origin with maximum margin.
- For a new point x, the value $f(x)$ is determined by evaluating which side of the hyperplane it falls on.
- The method proposed by Scholkopf (vSVM)is to separates all the data points from the origin (in feature space F) and maximizes the distance between this separating hyperplane and the origin.
- Another method proposed by Tax and Duin (SVDD) uses separating hypersphere.
- These two methods are intrinsically same, since we may use kernel trick to map the hyperplane to a hypersphere.

3 Support Vector Data Description (SVDD)

We define a model which gives a closed boundary around the data: a hypersphere. The sphere is characterized by center $\mathbf{a}$ and radius $R > 0$. We minimize the volume of the sphere by minimizing R^2, and demand that the sphere contains all training objects $\mathbf{x}_i$. Thus, the problem can be formulated as

$$\begin{aligned} \min_{R} \quad & R^2 \\ \text{subject to} \quad & \|\mathbf{x}_i - \mathbf{a}\| \leq R^2, \quad \forall i \end{aligned} \tag{1}$$

To allow the possibility of outliers in the training set, the distance from $\mathbf{x}_i$ to the center $\mathbf{a}$ should not be strictly smaller than R^2, but larger distance should be penalized. Therefore we introduce slack variables $\xi_i \geq 0$ and the minimization problem changes into:

$$\begin{aligned} \min_{R} \quad & R^2 + C\sum_i \xi_i \\ \text{subject to} \quad & \|\mathbf{x}_i - \mathbf{a}\| \leq R^2 + \xi_i, \quad \forall i \\ & \xi_i \geq 0 \qquad \forall i \end{aligned} \tag{2}$$

3.1 Dual Problem

The Lagrangian of primal problem is

$$L(R, \mathbf{a}, \alpha_i, \gamma_i, \xi_i) = R^2 + C\sum_i \xi_i - \sum_i \alpha_i\{R^2 + \xi_i - \|\mathbf{x}_i - \mathbf{a}\|^2\} - \sum_i \gamma_i \xi_i \tag{3}$$

with the Lagrangian multiplier $\alpha_i, \gamma_i \geq 0$.

Then, we have the Lagrangian dual

$$g(\alpha, \gamma) = \min_{R,\mathbf{a},\xi} L(R, \mathbf{a}, \alpha_i, \gamma_i, \xi_i), \tag{4}$$

and the dual problem of our primal problem is

$$\max_{\alpha\geq 0, \gamma \geq 0} g(\alpha, \gamma) = \max_{\alpha,\gamma} \min_{R,\mathbf{a},\xi} L(R, \mathbf{a}, \alpha_i, \gamma_i, \xi_i) \tag{5}$$

We can solve the Lagrange dual by taking derivative w.r.t. $R, \mathbf{a}$,and ξ and equaling to 0, and we have

$$\frac{\partial L}{\partial R} = 0: \quad 2R - 2R\sum_i \alpha_i = 0 \quad \Rightarrow \quad \sum_i \alpha_i = 1 \tag{6}$$

$$\frac{\partial L}{\partial \mathbf{a}} = 0: \quad \sum_i 2\alpha_i(\mathbf{x}_i - \mathbf{a}) = 0 \quad \Rightarrow \quad \mathbf{a} = \frac{\sum_i \alpha_i \mathbf{x}_i}{\sum_i \alpha_i} = \sum_i \alpha_i \mathbf{x}_i \quad \text{by (8).} \tag{7}$$

$$\frac{\partial L}{\partial \xi_i} = 0: \quad C - \alpha_i - \gamma_i = 0 \tag{8}$$

Since $\gamma_i \geq 0$, by (10) we have $C - \alpha_i \geq 0$,which implies $0 \leq \alpha_i \leq C$.

Since the primal problem is convex and slater condition holds, we can conclude strong duality holds (actually, we still need to check the relative interior is convex, but omit). In addition, if strong duality holds and $(R^*, \mathbf{a}^*, \xi^*)$ and (α^*, γ^*) are optimal solutions of primal and dual problem, then they must satisfy the KKT conditions, which are

1. primal constraints:

$$\|\mathbf{x}_i - \mathbf{a}^*\|^2 \leq R^{*2} + \xi_i^* \quad \forall i \tag{9}$$

$$\xi_i^* \geq 0 \quad \forall i \tag{10}$$

2. dual constraints:

$$\alpha_i^* \geq 0 \quad \textit{and} \quad \gamma_i^* \geq 0 \tag{11}$$

3. complementary slackness:

$$\alpha_i^*\{\|\mathbf{x}_i - \mathbf{a}^*\|^2 - R^{*2} + \xi_i^*\} = 0 \tag{12}$$

$$\gamma_i^*(-\xi_i^*) = 0 \tag{13}$$

If $\alpha_i^* = 0$, then $\|\mathbf{x}_i - \mathbf{a}^*\|^2 < R^{*2} + \xi_i^*$; if $\|\mathbf{x}_i - \mathbf{a}^*\|^2 = R^{*2} + \xi_i^*$, then $\alpha_i > 0$.
If $\gamma_i^* = 0$, then $\xi_i^* > 0$; if $\xi_i^* = 0$, then $\gamma_i^* > 0$.

4. gradient of Lagrangian with respect to $R^*, \mathbf{a}^*$,and ξ^* vanishes

$$\mathbf{a}^* = \sum_i \alpha_i \mathbf{x}_i \tag{14}$$

We can rewrite the dual problem as a constrained problem as follows:

$$\max_{\alpha,\gamma} g(\alpha, \gamma) = \max_{\alpha,\gamma} \min_{R,\mathbf{a},\xi} L(R, \mathbf{a}, \alpha_i, \gamma_i, \xi_i) \tag{15}$$

$$= \max_{\alpha,\gamma} \min_{R,\mathbf{a},\xi} R^2 + C\sum_i \xi_i - \sum_i \alpha_i\{R^2 + \xi_i - (\|\mathbf{x}_i\|^2 - 2\mathbf{a}\cdot\mathbf{x}_i + \|\mathbf{a}\|^2)\} - \sum_i \gamma_i\xi_i \tag{16}$$

$$= \max_{\alpha,\gamma} \min_{R,\mathbf{a},\xi} \underbrace{R^2 - R^2\sum_i \alpha_i}_{=0} + \sum_i \alpha_i(\|\mathbf{x}_i\|^2 - 2\mathbf{a}\cdot\mathbf{x}_i + \|\mathbf{a}\|^2) + \underbrace{C\sum_i \xi_i - \sum_i \alpha_i\xi_i - \sum_i \gamma_i\xi_i}_{=0} \tag{17}$$

$$= \max_{\alpha,\gamma} \sum_i (\alpha_i \mathbf{x}_i^\top \mathbf{x}_i - 2\alpha_i \sum_j \alpha_j x_j^\top x_i + \sum_i \sum_j \alpha_i\alpha_j \mathbf{x}_i^\top \mathbf{x}_j) \tag{18}$$

$$= \max_{\alpha,\gamma} \sum_i \alpha_i \mathbf{x}_i^\top \mathbf{x}_i - \sum_i \sum_j \alpha_i\alpha_j \mathbf{x}_i^\top \mathbf{x}_j \tag{19}$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq C, \quad \text{and} \quad \sum_i \alpha_i = 1 \tag{20}$$

The dual problem is a quadratic programming problem, We can apply QP-solver or SGD to solve it. In addition, we need to analyze if the solution is reasonable according to some observations on complementary slackness condition:

1. If $\alpha_i^* = 0$, and since $C > 0$, then we can conclude $\|\mathbf{x}_i - \mathbf{a}^*\|^2 < R^{*2} + \xi_i^*$ and $\gamma_i^* > 0$, which implies $\xi_i^* = 0$. Thus, we have $\|\mathbf{x}_i - \mathbf{a}^*\| < R^2$.
2. If $\alpha_i^* = C > 0$, then we have $\|\mathbf{x}_i - \mathbf{a}^*\|^2 = R^{*2} + \xi_i^*$ and $\gamma_i = 0$ and , which implies $\xi_i^* > 0$. Thus we have $\|\mathbf{x}_i - \mathbf{a}^*\|^2 - R^{*2} = \xi_i^* > 0$, that is $\|\mathbf{x}_i - \mathbf{a}^*\|^2 > R^{*2}$.
3. If $0 < \alpha_i^* < C$, when we have $\|\mathbf{x}_i - \mathbf{a}^*\|^2 = R^{*2} + \xi_i^*$ and $\gamma_i > 0$, which implies $\xi_i^* = 0$. Thus, we have $\|\mathbf{x}_i - \mathbf{a}^*\| = R^2$.

In addition, we know that $\mathbf{a}^* = \sum_i \alpha_i \mathbf{x}_i$, that means the center of the sphere is a linear combination of the objects. Only object $\mathbf{x}_i$ with $\alpha_i > 0$ are needed in the description and these objects will therefore be called the support vectors of the description.

When we test if an object $\mathbf{z}$ is an outlier, we only need to check if $\mathbf{z}$ falls in the sphere, which is

$$\|\mathbf{z} - \mathbf{a}\|^2 = (\mathbf{z}^\top \mathbf{z}) - 2\sum_i \alpha_i (\mathbf{z}^\top \mathbf{x}_i) + \sum_i \sum_j (\mathbf{x}_i^\top \mathbf{x}_j) \leq R^2 \tag{21}$$

where R^2 is defined as the distance from the center $\mathbf{a}$ to any support vector on the boundary, which is any $\mathbf{x}_k$ corresponding to $0 < \alpha_k < C$. Therefore,

$$R^2 = (\mathbf{x}_k^\top \mathbf{x}_k) - 2\sum_i \alpha_i (\mathbf{x}_k^\top \mathbf{x}_i) + \sum_i \sum_j (\mathbf{x}_i^\top \mathbf{x}_j) \tag{22}$$

for any $\mathbf{x}_k \in SV_{0<\alpha_k<C}$.

4 One-class ν-SVM Formulation

The method proposed by Scholkopf (ν-SVM) is to separate all the data points from the origin (in feature space F) and maximize the distance between this separating hyperplane and the origin. Hence, it is to maximize the distance between origin (in mapped feature space) and the hyperplane $y = \mathbf{w}\Phi(x) + \rho$, which is

$$\frac{\rho}{\|\mathbf{w}\|} \tag{23}$$

Φ is a kernel mapping function, $(\mathbf{w}, \rho)$ are a weight vector and an offset parameterizing a hyperplane in the feature space associated with the kernel.

Question: why use origin ? Other approaches rather than using the origin ?

If we fix the ρ, we would like the denominator small; but at the meantime, we want $\mathbf{w} \cdot \Phi(\mathbf{x}_i) \geq \rho$. Hence, we can formulate our problem as a quadratic program:

$$\begin{aligned} \min_{\mathbf{w}\in F, \xi \in \mathbb{R}^n, \rho \in \mathbb{R}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\nu n}\sum_i \xi_i \\ \text{subject to} \quad & (\mathbf{w} \cdot \Phi(x_i)) \geq \rho - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{24}$$

where $\nu \in (0, 1)$ is a trade-off parameter, similar to C in standard C-SVM. ν is introduced to avoid using grid search on C. But if we replace $\frac{1}{\nu n}$ with C, it does not effect us to solve problem here.

4.1 Dual Problem

The Lagrangian of primal problem is

$$L(\mathbf{w}, \rho, \xi_i, \alpha_i, \gamma_i) = \frac{1}{2}\|\mathbf{w}\|^2 - \rho + \frac{1}{\nu n}\sum_i \xi_i - \sum_i \alpha_i((\mathbf{w} \cdot \Phi(x_i)) - \rho + \xi_i) - \sum_i \gamma_i \xi_i \tag{25}$$

with the Lagrangian multiplier $\alpha_i, \gamma_i \geq 0$.

Then, we have the Lagrangian dual

$$g(\alpha, \gamma) = \min_{\mathbf{w}, \rho, \xi} L(\mathbf{w}, \rho, \xi_i, \alpha_i, \gamma_i), \tag{26}$$

and the dual problem of our primal problem is

$$\max_{\alpha \geq 0, \gamma \geq 0} g(\alpha, \gamma) = \max_{\alpha, \gamma} \min_{\mathbf{w}, \rho, \xi} L(\mathbf{w}, \rho, \xi_i, \alpha_i, \gamma_i), \tag{27}$$

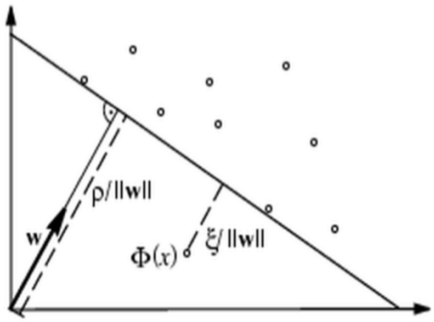

Figure 1: Geometric Interpretation of Primal Problem on ν-SVM.

We can solve the Lagrange dual by taking derivative w.r.t. $\mathbf{w}, \rho$,and ξ and equaling to 0, and we have

$$\frac{\partial L}{\partial \mathbf{w}} = 0: \quad \mathbf{w} - \sum_i \alpha_i \Phi(\mathbf{x}_i) = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_i \alpha_i \Phi(\mathbf{x}_i) \tag{28}$$

$$\frac{\partial L}{\partial \rho} = 0: \quad -1 + \sum_i \alpha_i = 0 \quad \Rightarrow \quad \sum_i \alpha_i = 1 \tag{29}$$

$$\frac{\partial L}{\partial \xi_i} = 0: \quad \frac{1}{\nu n} - \alpha_i - \gamma_i = 0 \tag{30}$$

Since $\gamma_i \geq 0$, by (10) we have $\frac{1}{\nu n} - \alpha_i \geq 0$,which implies $0 \leq \alpha_i \leq \frac{1}{\nu n}$.

The KKT conditions are

1. primal constraints:

$$(\mathbf{w}^* \cdot \Phi(x_i)) \geq \rho^* - \xi_i \quad \forall i \tag{31}$$

$$\xi_i^* \geq 0 \quad \forall i \tag{32}$$

2. dual constraints:

$$\alpha_i^* \geq 0 \quad and \quad \gamma_i^* \geq 0 \tag{33}$$

3. complementary slackness:

$$\alpha_i^*((\mathbf{w}^* \cdot \Phi(x_i)) - \rho + \xi_i^*) = 0 \tag{34}$$

$$\gamma_i^*(-\xi_i^*) = 0 \tag{35}$$

If $\alpha_i^* = 0$, then $(\mathbf{w}^* \cdot \Phi(x_i)) \geq \rho^* - \xi_i$; if $(\mathbf{w}^* \cdot \Phi(x_i)) = \rho^* - \xi_i$, then $\alpha_i > 0$.
If $\gamma_i^* = 0$, then $\xi_i^* > 0$; if $\xi_i^* = 0$, then $\gamma_i^* > 0$.

4. gradient of Lagrangian with respect to $\mathbf{w}^*, \rho^*$,and ξ^* vanishes

$$\mathbf{w}^* = \sum_i \alpha_i \Phi(\mathbf{x}_i) \tag{36}$$

We arrive at the following quadratic program which is the dual of the primal program

$$\begin{aligned} \max_{\alpha} \quad & \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \\ \text{subject to} \quad & 0 \leq \alpha_i \leq \frac{1}{\nu n}, \quad \sum_i^n \alpha_i = 1 \end{aligned} \tag{37}$$

We can solve it with SGD or QP solver. We also can validate the solution based on some observations according to complementary slackness:

1. If $\alpha_i^* = 0$, and since $\nu > 0$, then we can conclude $(\mathbf{w}^* \cdot \Phi(x_i)) > \rho^* - \xi_i$ and $\gamma_i^* > 0$, which implies $\xi_i^* = 0$. Thus, we have $(\mathbf{w}^* \cdot \Phi(x_i)) > \rho^*$.

2. If $\alpha_i^* = \frac{1}{\nu n} > 0$, then we have $(\mathbf{w}^* \cdot \Phi(x_i)) = \rho^* - \xi_i$ and $\gamma_i = 0$ and , which implies $\xi_i^* > 0$. Thus we have $\rho^* - (\mathbf{w}^* \cdot \Phi(x_i)) = \xi_i^* > 0$, that is $(\mathbf{w}^* \cdot \Phi(x_i)) < \rho^*$.

3. If $0 < \alpha_i^* < \frac{1}{\nu n}$, then we have $(\mathbf{w}^* \cdot \Phi(x_i)) = \rho^* - \xi_i$ and $\gamma_i > 0$, which implies $\xi_i^* = 0$. Thus, we have $(\mathbf{w}^* \cdot \Phi(x_i)) = \rho^*$.

Since $\mathbf{w} = \sum_i \alpha_i \Phi(\mathbf{x}_i)$, all object x_i has corresponding $\alpha_i > 0$ will contribute to determining the value of $\mathbf{w}$ (the hyperplane's direction). Those objects are called support vectors. If we test an object $\mathbf{z}$, we can check

$$f(\mathbf{z}) = sgn(\mathbf{w}\Phi(x) - \rho) = \begin{cases} + & \Rightarrow \quad \text{normal} \\ - & \Rightarrow \quad \text{outlier} \end{cases} \tag{38}$$

where ρ can be computed by choosing any x_k which has corresponding $0 < \alpha_i < \frac{1}{\nu n}$, and get

$$\rho = (\mathbf{w} \cdot \Phi(\mathbf{x}_k)) = \sum_i \alpha_i K(x_i, x_k) \tag{39}$$

5 Connection Between One-class ν-SVM and SVDD

An equivalent formulation of One-class ν-SVM is

$$\begin{aligned} \max \quad & \rho - \frac{1}{\nu n} \sum_{\xi_i} \\ \text{subject to} \quad & \mathbf{w} \cdot \Phi(\mathbf{x}_i) \geq \rho - \xi_i \\ & \xi_i \geq 0 \\ & \|\mathbf{w}\| = 1 \end{aligned} \tag{40}$$

Now, if we normalize all data in original SVDD formulation, we obtains the following optimization problem

$$\min \quad R'^2 + C' \textstyle\sum_i \xi_i' \tag{41}$$

$$\text{subject to} \quad \|\mathbf{x}_i' - \mathbf{a}'\| \leq R'^2 + \xi_i' \quad \forall i \quad \xi_i \geq 0 \tag{42}$$

where $\mathbf{x}'$ and $\mathbf{a}'$ are normalized vectors. The constraints can be rewritten as

$$\begin{aligned} \|x_i'\|^2 - 2\mathbf{a}' \cdot \mathbf{x}_i' + \|\mathbf{a}'\|^2 &\leq R'^2 + \xi_i' \\ 1 - 2\mathbf{a}' \cdot \mathbf{x}_i' + 1 &\leq R'^2 + \xi_i' \\ 2\mathbf{a}' \cdot \mathbf{x}_i' &\geq 2 - R'^2 - \xi_i' \\ \mathbf{a}' \cdot \mathbf{x}_i' &\geq \frac{1}{2}(2 - R'^2) - \frac{1}{2}\xi_i' \end{aligned} \tag{43}$$

We can further rewrite the problem as

$$\max \quad 2 - R'^2 - C' \textstyle\sum_i \xi_i' \tag{44}$$

$$\text{subject to} \quad \mathbf{a}' \cdot \mathbf{x}_i' \geq \tfrac{1}{2}(2 - R'^2) - \tfrac{1}{2}\xi_i' \quad \forall i \tag{45}$$

$$\xi_i \geq 0 \tag{46}$$

where adding a constant 2 does not effect the solution. Now We define

$$\mathbf{w} = 2\mathbf{a}', \quad \rho = (2 - R'^2), \quad \frac{1}{\nu n} = C', \quad \xi_i = \xi_i' \tag{47}$$

$$\max \quad -2 + \rho - \tfrac{1}{\nu n} \textstyle\sum_i \xi_i \tag{48}$$

$$\text{subject to} \quad \mathbf{w} \cdot \mathbf{x}_i \geq \rho - \xi_i' \quad \forall i \tag{49}$$

$$\xi_i \geq 0 \tag{50}$$

$$\|\mathbf{w}\| = 2 \tag{51}$$

For normalized data, it differs in the constraint on the norm of $\mathbf{w}$ and in an offset of 2 in the error function. Gaussian kernel implicitly normalize the data.

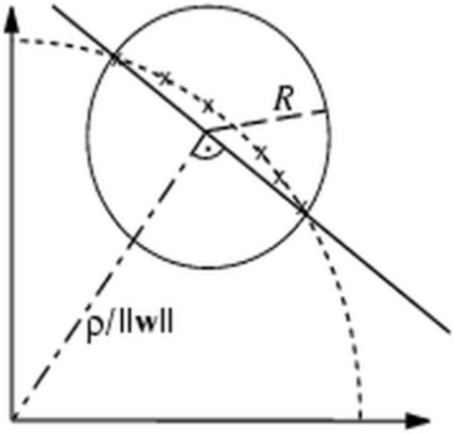

Figure 2: Geometric Interpretation of SVDD and one-class ν-SVM where the data is normalized to unit norm.

References

[1] Schölkopf, Bernhard, et al. "Support vector method for novelty detection." Advances in neural information processing systems. 2000.

[2] Schölkopf, Bernhard, et al. "Estimating the support of a high-dimensional distribution." Neural computation 13.7 (2001): 1443-1471.

[3] Chen, Pai-Hsuen, Chih-Jen Lin, Bernhard Schölkopf. "A tutorial on v-support vector machines." Applied Stochastic Models in Business and Industry 21.2 (2005): 111-136.http://vis.lbl.gov/~romano/mlgroup/papers/nusvmtutorial.pdf

[4] Wu, Xiaoyun, and Rohini K. Srihari. "Newí-Support Vector Machines and their Sequential Minimal Optimization." Proceedings of the 20th International Conference on Machine Learning (ICML-03). 2003.

[5] Crisp, David J., and Christopher JC Burges. "A geometric interpretation of v-SVM classifiers." Advances in neural information processing systems. 2000.

[6] http://www.dainf.ct.utfpr.edu.br/~kaestner/Mineracao/ArquivosExtras2016/Dan.Nick-OneClassSVM.pdf

[7] http://homepage.tudelft.nl/n9d04/thesis.pdf