Machine Learning 1 - Homework Week 5

Minh Ngo 10897402[1]

October 12, 2015

[1] University of Amsterdam
minh.ngole@student.uva.nl

Task 2

We define the primal program for kernel outlier detection as:

$$\begin{aligned} &\min_{\mathbf{a},R,\boldsymbol{\xi}} R^2 + C\sum_{i=1}^{N}\xi_i \\ &s.t.\ \forall i :\| \mathbf{x}_i - \mathbf{a} \|^2 \leq R^2 + \xi_i, \xi_i \geq 0 \end{aligned} \tag{1}$$

We can rewrite our constraints as:

$$\forall i :\| \mathbf{x}_i - \mathbf{a} \|^2 - R^2 + \xi_i \leq 0, -\xi_i \leq 0 \tag{2}$$

1. The primal Lagrangian will be:

$$\mathcal{L}(\mathbf{a}, R, \boldsymbol{\xi}, \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = R^2 + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\alpha_i(\| \mathbf{x}_i - \mathbf{a} \|^2 - R^2 - \xi_i) - \sum_{i=1}^{N}\mu_i\xi_i \tag{3}$$

2. KKT conditions are:

   (a)

$$\nabla_{R^2}\mathcal{L} = 1 - \sum_{i=1}^{N}\alpha_i = 0 \tag{4}$$

$$\Leftrightarrow \sum_{i=1}^{N} \alpha_i = 1 \tag{5}$$

(b)

$$\begin{aligned} \nabla_{\mathbf{a}} \mathcal{L} &= \nabla_{\mathbf{a}} \sum_{i=1}^{N} \alpha_i (\mathbf{x}_i - \mathbf{a})^T (\mathbf{x}_i - \mathbf{a}) \\ &= \sum_{i=1}^{N} \alpha_i \nabla_{\mathbf{a}} (\mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{x}_i^T \mathbf{a} + \mathbf{a}^T \mathbf{a}) \\ &= \sum_{i=1}^{N} \alpha_i (-2\mathbf{x}_i + 2\mathbf{a}) = 0 \end{aligned} \tag{6}$$

$$\begin{aligned} &\Leftrightarrow \sum_{i=1}^{N} \alpha_i (\mathbf{x}_i - \mathbf{a}) = 0 \\ &\Leftrightarrow \sum_{i=1}^{N} \alpha_i \mathbf{x}_i = \mathbf{a} \sum_{i=1}^{N} \alpha_i = \mathbf{a} \end{aligned} \tag{7}$$

(c)

$$\forall i : \nabla_{\xi_i} \mathcal{L} = C + \alpha_i - \mu_i = 0 \tag{8}$$

$$\Leftrightarrow \forall i : \mu_i = \alpha_i + C \tag{9}$$

(d) Complementary slackness conditions:

$$\begin{aligned} \forall i : &\alpha_i (\| \mathbf{x}_i - \mathbf{a} \|^2 - R^2 - \xi_i) = 0 \\ &\mu_i \xi_i = 0 \end{aligned} \tag{10}$$

(e)

$$\forall i : \alpha_i \geq 0, \mu_i \geq 0 \tag{11}$$

(f)

$$\forall i : \| \mathbf{x}_i - \mathbf{a} \|^2 - R^2 - \xi_i \leq 0, \ -\xi_i \leq 0 \tag{12}$$

3. Complementary slackness conditions have been defined [Eq 10]. From KKT conditions the following facts can be concluded:

$$\begin{aligned} \alpha_i > 0 \Rightarrow \parallel \mathbf{x}_i - \mathbf{a} \parallel^2 - R^2 - \xi_i = 0 \\ \Leftrightarrow \parallel \mathbf{x}_i - \mathbf{a} \parallel^2 = R^2 + \xi_i \end{aligned} \tag{13}$$

$$\begin{aligned} \alpha_i > 0 \Rightarrow \mu_i = \alpha_i + C > 0 \Rightarrow \xi_i = 0 \\ \Rightarrow \parallel \mathbf{x}_i - \mathbf{a} \parallel^2 = R^2 \end{aligned} \tag{14}$$

Therefore $\alpha_i > 0$ means that a point $\mathbf{x}_i$ is a support vector and defines a radius of the kernel outlier detector.

$$\begin{aligned} \mu_i > 0 \Rightarrow \xi_i = 0 \\ \Leftrightarrow \parallel \mathbf{x}_i - \mathbf{a} \parallel^2 \leq R^2 \end{aligned} \tag{15}$$

that means that a point $\mathbf{x}_i$ is an inlier and situated inside the circle.

4. If constraints 5, 7, 9 are **not** satisfied then the dual problem $\theta_D(\boldsymbol{\alpha}, \boldsymbol{\mu}) = -\infty$, otherwise:

$$\begin{aligned} \theta_D(\boldsymbol{\alpha}, \boldsymbol{\mu}) &= \min_{\boldsymbol{\alpha}, R, \boldsymbol{\xi}} \mathcal{L}(\mathbf{a}, R, \boldsymbol{\xi}, \mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\mu}) \\ &= \min_{\boldsymbol{\alpha}, R, \boldsymbol{\xi}} (R^2 + C \sum_{i=1}^{N} \xi_i + \sum_{i=1}^{N} \alpha_i (\parallel \mathbf{x}_i - \mathbf{a} \parallel^2 - R^2 - \xi_i) - \sum_{i=1}^{N} \mu_i \xi_i) \\ &= \min_{\boldsymbol{\alpha}, R, \boldsymbol{\xi}} (R^2 + \sum_{i=1}^{N} \alpha_i \parallel \mathbf{x}_i - \mathbf{a} \parallel^2 - \sum_{i=1}^{N} \alpha_i R^2) \\ &= \min_{\boldsymbol{\alpha}, R, \boldsymbol{\xi}} \sum_{i=1}^{N} \alpha_i (\mathbf{x}_i - \mathbf{a})^T (\mathbf{x}_i - \mathbf{a}) \\ &= \min_{\boldsymbol{\alpha}, R, \boldsymbol{\xi}} \sum_{i=1}^{N} \alpha_i (\mathbf{x}_i^T \mathbf{x}_i - \mathbf{x}_i^T \mathbf{a} - \mathbf{a}^T \mathbf{x}_i + \mathbf{a}^T \mathbf{a}) \end{aligned} \tag{16}$$

$$\begin{aligned}\theta_D(\boldsymbol{\alpha}, \boldsymbol{\mu}) &= \min_{\boldsymbol{\alpha}, R, \boldsymbol{\xi}} \sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x}_i - \sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \mathbf{a} - \mathbf{a}^T \sum_{i=1}^{N} \alpha_i (\mathbf{x}_i - \mathbf{a}) \\ &= \min_{\boldsymbol{\alpha}, R, \boldsymbol{\xi}} \sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x}_i - \sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \sum_{j=1}^{N} \alpha_j \mathbf{x}_j \\ &= \sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \mathbf{x}_i - \sum_{i,j=1}^{N} (\alpha_i \alpha_j)(\mathbf{x}_i^T \mathbf{x}_j)\end{aligned} \tag{17}$$

Let $k(\mathbf{x}, \mathbf{z}) = \mathbf{x}^T \mathbf{z}$ then

$$\theta_D(\boldsymbol{\alpha}, \boldsymbol{\mu}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i,j=1}^{N} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{18}$$

5.

$$\begin{aligned}\max_{\boldsymbol{\alpha}} \theta_D(\boldsymbol{\alpha}, \boldsymbol{\mu}) &= \max_{\boldsymbol{\alpha}} \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i,j=1}^{N} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \\ s.t. \forall i : \alpha_i \geq 0,\ &\sum_{i=1}^{N} \alpha_i = 1,\ \mu_i = C + \alpha_i\end{aligned} \tag{19}$$

will return optimal values for $\{\alpha_i\}$, after which you can solve the primal variables in terms of the dual variables $\alpha_i,\ \mu_i$:

$$\mathbf{a} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i \tag{20}$$

As it was mentioned in the task 2.3 we can determine if a point $\mathbf{x}_k$ forms a support vector by checking $\alpha_k > 0$. For these points $\xi_k = 0$. Let K be an amount of points that satisfy this constraint and form support vectors, then radius can be computed as following:

$$R = \frac{1}{K} \sum_{k; \alpha_k > 0} \parallel \mathbf{x}_k - \mathbf{a} \parallel \tag{21}$$

If $\| \mathbf{x}_i - \mathbf{a} \|^2 < R^2 \Rightarrow \xi_i = 0$.

If $\| \mathbf{x}_i - \mathbf{a} \|^2 > R^2 \Rightarrow \xi_i = \| x_i - a \|^2 - R^2$

6. Suppose a new point is $\mathbf{x}_t$ then we can detect if it's outlier by the following inequality:

$$\| \mathbf{x}_t - \mathbf{a} \|^2 > R^2 \tag{22}$$

$$\begin{aligned} \mathbf{x}_t^T \mathbf{x}_t - 2\mathbf{x}_t^T \mathbf{a} + \mathbf{a}^T \mathbf{a} &> \frac{1}{K} \sum_{k;\alpha_k>0} (\mathbf{x}_k - \mathbf{a})^T (\mathbf{x}_k - \mathbf{a}) \\ \mathbf{x}_t^T \mathbf{x}_t - 2\mathbf{x}_t^T \mathbf{a} + \mathbf{a}^T \mathbf{a} &> \frac{1}{K} \sum_{k;\alpha_k>0} (\mathbf{x}_k^T \mathbf{x}_k - 2\mathbf{x}_k^T \mathbf{a} + \mathbf{a}^T \mathbf{a}) \end{aligned} \tag{23}$$

In the term of kernels and Lagrange multipliers:

$$\begin{aligned} & k(\mathbf{x}_t, \mathbf{x}_t) - 2\sum_{i=1}^{N} \alpha_i k(\mathbf{x}_t, \mathbf{x}_i) + \sum_{i,j=1}^{N} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \\ & > \frac{1}{K} \sum_{k;\alpha_k>0} \left(k(\mathbf{x}_k, \mathbf{x}_k) - 2\sum_{i=1}^{N} \alpha_i k(\mathbf{x}_k, \mathbf{x}_i) \right) + \sum_{i,j=1}^{N} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \end{aligned} \tag{24}$$

$$\begin{aligned} & k(\mathbf{x}_t, \mathbf{x}_t) - 2\sum_{i=1}^{N} \alpha_i k(\mathbf{x}_t, \mathbf{x}_i) \\ & > \frac{1}{K} \sum_{k;\alpha_k>0} \left(k(\mathbf{x}_k, \mathbf{x}_k) - 2\sum_{i=1}^{N} \alpha_i k(\mathbf{x}_k, \mathbf{x}_i) \right) \end{aligned} \tag{25}$$

7. C $= 0 \Rightarrow$ no penalty for outliers. The primal program becomes $\min_{\boldsymbol{\alpha}, R, \boldsymbol{\xi}} R^2$ that has a solution $R = 0$ that can be satisfied by constraints $\| \mathbf{x}_i - \mathbf{a} \|^2 \leq \xi_i,\ \xi_i \geq 0$.

   $C = \infty \Rightarrow \mu_i = \infty \Rightarrow \xi_i = 0\ \forall i \Rightarrow \| \mathbf{x}_i - \mathbf{a} \|^2 - R^2 \leq 0$ that means that all points are in the circle.

8. Gaussian kernel has a form:

$$k(\mathbf{x}, \mathbf{z}) = exp\Big(- \frac{\| \mathbf{x} - \mathbf{z} \|^2}{2\sigma^2} \Big) \tag{26}$$

with a small bandwidth $\sigma = 0$, $k(\mathbf{x}, \mathbf{z})$ becomes close to 1 if only $\mathbf{x}$ and $\mathbf{z}$ are very similar, and close to 0 otherwise. If we put it to the equation 25 we obtain the following criteria for outlier detection:

$$\begin{aligned} 1 - 2\sum_{i=1}^{N} \alpha_i k(\mathbf{x}_t, \mathbf{x}_i) > 1 - 2\sum_{i=1}^{N} \frac{1}{K} \sum_{k;\alpha_k>0} \alpha_i k(\mathbf{x}_k, \mathbf{x}_i) \\ \sum_{i=1}^{N} \alpha_i (\frac{1}{K} \sum_{k;\alpha_k>0} k(\mathbf{x}_k, \mathbf{x}_i) - k(\mathbf{x}_t, \mathbf{x}_i)) > 0 \\ \sum_{i=1,\alpha_i>0}^{N} \alpha_i (\frac{1}{K} \sum_{k;\alpha_k>0} k(\mathbf{x}_k, \mathbf{x}_i) - k(\mathbf{x}_t, \mathbf{x}_i)) > 0 \end{aligned} \tag{27}$$

that means that $\mathbf{x}_t$ will be detected as an inlier if $\mathbf{x}_t$ is close to some support vector $\mathbf{x}_i^*$ that makes this sum above a negative value:

$$\sum_{i=1,\alpha_i>0}^{N} \alpha_i (\frac{1}{K} \sum_{k;\alpha_k>0} k(\mathbf{x}_k, \mathbf{x}_i) - k(\mathbf{x}_t, \mathbf{x}_i)) = -\alpha_i k(\mathbf{x}_t, \mathbf{x}_i^*) < 0 \tag{28}$$

$\mathbf{x}_t$ will be detected as an outlier otherwise.

Gaussian and linear kernel behave a bit differently. Let's say for vectors that have the same direction but significantly different size (for example $\mathbf{x}$ and $1000\mathbf{x}$). Linear kernel will return high value for them, but Gaussian kernel will return a high value only if vectors are really similar.

9. Given labels for outliers (y = 1) and inliners (y = -1) we can change the the primal problem to include these labels:

$$\begin{aligned} \min_{\mathbf{a},R,\boldsymbol{\xi}} R^2 + C\sum_{i=1}^{N} \xi_i \\ y_i = 1 \Rightarrow \| \mathbf{x}_i - \mathbf{a} \|^2 - R^2 - \xi_i \leq 0,\ \xi_i \geq 0 \\ y_i = -1 \Rightarrow \| \mathbf{x}_i - \mathbf{a} \|^2 - R^2 \leq 0 \end{aligned} \tag{29}$$