Sheet 7 | Based on a paper: „A Revisit to SVDD", National Taiwan University, Taipei

EX 1 a) Derive the dual program for the one-class SVM

SVDD primal is given by: ~~$(0 \leq v \leq 1)$~~ $(0 < v \leq 1)$

$$\text{(1)} \quad \min_{R, c, (\xi_i)_{i=1}^n} R^2 + \frac{1}{nv} \sum_{i=1}^n \xi_i$$

$$\text{s.t.} \quad \forall_{i=1}^n \ \|\phi(x_i) - c\|^2 \leq R^2 + \xi_i \ \text{and} \ \xi_i \geq 0$$

For minimizing an optimization problem, before deriving the Lagrangian dual problem and checking whether the dual optimal value is identical to the primal optimal value, one should make sure that the original problem is convex. Otherwise, the duality theory of convex programming may not be applicable.

The primal problem (1) is non-convex, because $\|\phi(x_i) - c\|^2 - R^2 - \xi_i$ is concave wrt. R. Therefore problem (1) has a non-convex feasible region, so it is not a convex optimization problem.

Nevertheless, by defining $\bar{R} = R^2$, (1) is equivalent to the following convex problem:

$$\text{(2)} \quad \min_{\bar{R}, c, (\xi_i)_{i=1}^n} \bar{R} + \frac{1}{nv} \sum_{i=1}^n \xi_i$$

$$\text{s.t.} \quad \forall_{i=1}^n \ \|\phi(x_i) - c\|^2 \leq \bar{R} + \xi_i \ \text{and} \ \xi_i \geq 0$$

$$\bar{R} \geq 0$$

(A new constraint specifying the non-negativity of $\bar{R}$ is added).

A related difficulty is whether the non-negativity constraint $\bar{R} \geq 0$ is active. We handle this, by splitting the derivation to two cases:

1° for any $0 < v < 1$, the constraint $\bar{R} \geq 0$ in (2) is not necessary i.e. without this constraint any optimal solution still satisfies $\bar{R} \geq 0$

2° for any $v = 1$ at least one optimal solution has $\bar{R} = 0$.

We derive the dual problem by considering these two cases separately.

Case 1 : $0 < v < 1$

The Lagrangian of ② is :

$$\mathcal{L}(c, \bar{R}, \xi, \alpha, \gamma) = \bar{R} + \frac{1}{nv}\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i(\bar{R} + \xi_i - \|\phi(x_i) - c\|^2) - \sum_{i=1}^{n} \gamma_i \xi_i =$$

$$= \bar{R}\left(1 - \sum_{i=1}^{n}\alpha_i\right) + \sum_{i=1}^{n}\xi_i\left(\frac{1}{nv} - \alpha_i - \gamma_i\right) + \sum_{i=1}^{n}\alpha_i\left(\|\phi(x_i) - c\|^2\right)$$

where α and γ are Lagrange multipliers.

The Lagrange dual problem is : $\max\limits_{\alpha \geq 0, \gamma \geq 0}\left(\inf\limits_{c, \bar{R}, (\xi_i)_{i=1}^{n}} \mathcal{L}(c, \bar{R}, \xi, \alpha, \gamma)\right)$

From Karush-Kuhn-Tucker (KKT) Conditions :

1: Stationarity : condition for primal variables $\nabla\mathcal{L} = 0$ gives : (zero gradient of Lagrangian wrt. primal variables)

$$\frac{\partial\mathcal{L}}{\partial\bar{R}} = 0 \; : \quad \sum_{i=1}^{n}\alpha_i = 1$$

$$\frac{\partial\mathcal{L}}{\partial c} = 0 : \quad 2c\underbrace{\sum_{i=1}^{n}\alpha_i}_{=1} - 2\sum_{i=1}^{n}\alpha_i\phi(x_i) = 0 \implies c = \sum_{i=1}^{n}\alpha_i\phi(x_i)$$

$$\frac{\partial\mathcal{L}}{\partial\xi_i} = 0 : \quad \frac{1}{nv} - \alpha_i - \gamma_i = 0 \implies \alpha_i, \gamma_i \geq 0 \implies 0 \leq \alpha_i \leq \frac{1}{nv}$$

~~Substituting the results back into $\mathcal{L}(\cdot)$ can yield the dual problem formulated as:~~

2. primal admisson :

$$\|\phi(x_i) - c\|^2 \leq \bar{R} + \xi_i \quad \forall_{i=1}^{n}$$

$$\xi_i \geq 0 \quad \forall_{i=1}^{n}$$

3: dual admission : $\alpha_i \geq 0$, $\gamma_i \geq 0$

4. complementarity : (slackness conditions)

$$\alpha_i(\|\phi(x_i) - c\|^2 - \bar{R} - \xi_i) = 0 \quad \text{and} \quad \gamma_i\xi_i = 0 \quad \forall_{i=1}^{n}$$

↳ Complementarity tells us : two groups of points :

1) the support vectors $\|\phi(x_i) - c\|^2 = \bar{R}$

2) and the insiders $\alpha_i = 0$

Resubstituting the optimality conditions in $\mathcal{L}(\cdot)$ gives the Lagrange function only in dual variables α_i

$$\mathcal{L}(\alpha_i) = \bar{R}\underbrace{\left(1-\sum_{i=1}^{n}\alpha_i\right)}_{=0} + \sum_{i=1}^{n}\xi_i\underbrace{\left(\frac{1}{n\nu}-\alpha_i-\gamma_i\right)}_{=0} + \sum_{i=1}^{n}\alpha_i\left(\|\phi(x_i)-\underbrace{c}_{c=\sum_{i=1}^{n}\alpha_i\phi(x_i)}\|^2\right) =$$

$$= \sum_{i=1}^{n}\alpha_i\left(\|\phi(x_i)-\sum_{j=1}^{n}\alpha_j\phi(x_j)\|^2\right) =$$

$$= \sum_{i=1}^{n}\alpha_i\left(\phi(x_i)\phi(x_i) - 2\phi(x_i)\sum_{j=1}^{n}\alpha_j\phi(x_j) + \left(\sum_{j=1}^{n}\alpha_j\phi(x_j)\right)^T\left(\sum_{j=1}^{n}\alpha_j\phi(x_j)\right)\right)$$

$$= \sum_{i=1}^{n}\alpha_i\phi(x_i)^T\phi(x_i) - 2\sum_{i=1}^{n}\alpha_i\phi(x_i)^T\sum_{j=1}^{n}\alpha_j\phi(x_j) + \underbrace{\sum_{i=1}^{n}\alpha_i}_{=1}\cdot\left(\sum_{i=1}^{n}\alpha_i\phi(x_i)\right)^T\left(\sum_{j=1}^{n}\alpha_i\phi(x_i)\right)$$

$$= \sum_{i=1}^{n}\alpha_i\phi(x_i)^T\phi(x_i) - \left(\sum_{i=1}^{n}\alpha_i\phi(x_i)\right)^T\left(\sum_{j=1}^{n}\alpha_j\phi(x_j)\right) =$$

$$= \sum_{i=1}^{n}\alpha_i\phi(x_i)^T\phi(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\phi(x_i)^T\phi(x_j)$$

That yields the dual problem formulated as:

$$\max_{\alpha}\ \sum_{i=1}^{n}\alpha_i\phi(x_i)^T\phi(x_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j\phi(x_i)^T\phi(x_j)$$

$$\text{s.t.}\ \sum_{i=1}^{n}\alpha_i = 1 \quad \text{and} \quad 0 \le \alpha_i \le \frac{1}{n\nu} \quad \forall_{i=1}^{n}$$

Case 2 : $\nu = 1$

then at least one optimal solution has $\bar{R}=0$, so we can remove the variable $\bar{R}$ from the problem

The minimum must occur when $\xi_i = \|\phi(x_i)-c\|^2 \geq 0$

Thus problem (2) can be reduced to:

$$\min_{c}\ \sum_{i=1}^{n}\|\phi(x_i)-c\|^2$$

This problem is strictly convex to c, so setting the gradient to be zero leads to:

$$c = \frac{\sum_{i=1}^{n}\phi(x_i)}{n}$$

1 b) Show that the kernelized dual has the form

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i k(x_i, x_i) - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)$$

$$\text{s.t. } \sum_{i=1}^{n} \alpha_i = 1 \quad \text{and } \forall_{i=1}^{n}: 0 \leq \alpha_i \leq \frac{1}{n\nu}$$

and with center $c = \sum_{i=1}^{n} \alpha_i \phi(x_i)$

where k is the kernel associated to the feature map.

SVDD can be simply kernalized $\phi(x_i)^T \phi(x_j)$ in the computations by $k(x_i, x_j)$, where $k(\cdot,\cdot)$ is some suitable kernel function

Sidenote:

Let $\mathcal{H}$ be a RKHS on some domain $\mathbb{R}^p$ endowed with kernel k. If there exist some constant c s.t. $\forall x \in \mathbb{R}^p$ $k(x, x) = c$, then the SVDD and One-Class SVM are equivalent (for translation invariant kernels).

Let consider the kernel $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$

Then the dual problem after kernalization is:

from ~~...~~ excercise 1a):

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i \phi(x_i)^T \phi(x_i) - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \phi(x_i)^T \phi(x_j)$$

$$\text{s.t. } \sum_{i=1}^{n} \alpha_i = 1 \quad \text{and } \forall_{i=1}^{n}: 0 \leq \alpha_i \leq \frac{1}{n\nu}$$

replacing $\phi(x_i)^T \phi(x_j) = k(x_i, x_j)$ we get:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i k(x_i, x_i) - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)$$

$$\text{s.t. } \sum_{i=1}^{n} \alpha_i = 1 \quad \text{and } \forall_{i=1}^{n} \; 0 \leq \alpha_i \leq \frac{1}{n\nu}$$

Note that the KKT stationarity condition $\left(\frac{\partial \mathcal{L}}{\partial c} = 0\right)$ is equivalent to the Representer Theorem i.e. $c = \sum_{i=1}^{n} \alpha_i \phi(x_i) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$

EX2. Show that the dual program from EX1 is a linearly constrained QP:

$$\min_{\alpha} \alpha^T P \alpha + q^T \alpha$$
$$\text{s.t } G\alpha \leq h \text{ and } A\alpha = b$$

From EX1 we have the dual program:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i k(x_i, x_i) - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)$$
$$\text{s.t. } \sum_{i=1}^{n} \alpha_i = 1 \text{ and } \forall_{i=1}^{n} \; 0 \leq \alpha_i \leq \frac{1}{n\nu}$$

which is equivalent to minimization problem:

$$\min_{\alpha} \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) - \sum_{i=1}^{n} \alpha_i k(x_i, x_i)$$
$$\text{s.t. } \sum_{i=1}^{n} \alpha_i = 1 \text{ and } \forall_{i=1}^{n} \; 0 \leq \alpha_i \leq \frac{1}{n\nu}$$

Rewriting it in a matrix form:

→ matrix $\underline{P}$ is a kernel matrix with the elements $P_{ij} = k(x_i, x_j)$ of dimensions $n \times n$

→ vector $\underline{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$ of dimensions $n \times 1$

→ vector $\underline{q} = \begin{bmatrix} k(x_1, x_1) \\ \vdots \\ k(x_n, x_n) \end{bmatrix}$ with the diagonal elements of the kernel matrix of dimensions $n \times 1$

→ matrix $\underline{G}$ is an identity matrix of dimensions $n \times n$ $G = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & & & \vdots \\ \vdots & & \ddots & & 0 \\ 0 & & & 0 & 1 \end{bmatrix}$

→ vector $\underline{h}$ is a vector with all the same elements $h = \begin{bmatrix} 1/n\nu \\ \vdots \\ 1/n\nu \end{bmatrix}$ of dimensions $n \times 1$

$\Rightarrow$ matrix $\underline{A}$ contains only number 1 $\quad A = [1, \dots 1]$
and it's actually a transposed unity vector
A has dimensions $1 \times n$.

$\Rightarrow$ vector $\underline{b}$ is a scalar: $b = \underline{1}$ (vector of dim. 1×1)