CSE – 426 Homework 4

Griffin Kent

1) Given two training examples $\boldsymbol{x}^{(1)} = [2,1]^T$, $y^{(1)} = 1$ and $\boldsymbol{x}^{(2)} = [1,-1]^T$, $y^{(2)} = -1$, evaluate the linear and Gaussian kernel (with bandwidth $\sigma = 1$) functions on these two examples. [*Hints: refer to the lecture note of SVM.*]

(Solution): The linear kernel function is defined as the following

$$k_L(\boldsymbol{x}, \boldsymbol{z}) = \langle \psi(\boldsymbol{x}), \psi(\boldsymbol{z}) \rangle = \langle \boldsymbol{x}, \boldsymbol{z} \rangle = \boldsymbol{x}^T \boldsymbol{z}$$

Where $\psi(\boldsymbol{x}) = \boldsymbol{x}$ is the identity function. Plugging in the two training examples, we obtain:

$$k_L\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\right) = [2,1]\begin{bmatrix} 1 \\ -1 \end{bmatrix} = 2 - 1 = 1.$$

The Gaussian kernel is defined as the following

$$k_G(\boldsymbol{x}, \boldsymbol{z}) = e^{-\frac{\|\boldsymbol{x}-\boldsymbol{z}\|^2}{2\sigma^2}}$$

Plugging in the two training examples, we obtain:

$$k_G\left(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\right) = e^{-\frac{(2-1)^2+(1-(-1))^2}{2}} \approx 0.082.$$

■

2) Use the training example from Question 1, write down the primal problem for soft-SVM with parameters $\boldsymbol{w}$, b, and ξ. [*Hints: You need to plug the data into the general formulation and simplify the primal objective function and the constraints.*]

(Solution): The primal form of the soft-SVM problem is stated as follows:

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \xi_i$$
$$y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi_i, \quad \forall i$$
$$\xi_i \geq 0, \quad \forall i$$

Inputting the training data, we obtain the following:

$$\min \frac{1}{2}(\boldsymbol{w}_1^2 + \boldsymbol{w}_2^2) + C(\xi_1 + \xi_2)$$
$$1 - 2\boldsymbol{w}_1 - \boldsymbol{w}_2 - b \leq \xi_1$$
$$1 + \boldsymbol{w}_1 - \boldsymbol{w}_2 + b \leq \xi_2$$
$$\xi_1, \xi_2 \geq 0$$

■

3) Use the training example from Question 1, write down the dual problem for soft-SVM. [*Hints: You need to plug the data into the general formulation and simplify the primal objective function and the constraints.*]

(Solution): The dual of the soft-SVM has the form:

$$\max L(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} K_{ij}$$

$$\sum_{j=1}^{m} \alpha_i y^{(i)} = 0$$

$$C \geq \alpha_i \geq 0, \quad i = 1, \dots, m$$

Inputting the training data, we obtain the following:

$$\max L(\alpha) = \alpha_1 + \alpha_2 - \frac{1}{2}[\alpha_1\alpha_1 y^{(1)} y^{(1)} K_{11} + \alpha_1\alpha_2 y^{(1)} y^{(2)} K_{12} + \alpha_2\alpha_1 y^{(2)} y^{(1)} K_{21} + \alpha_2\alpha_2 y^{(2)} y^{(2)} K_{22}]$$

$$s.t \quad \alpha_1 y^{(1)} + \alpha_2 y^{(2)} = 0$$

$$C \geq \alpha_1 \geq 0$$

$$C \geq \alpha_2 \geq 0$$

Simplifying this, we obtain the form:

$$\max L(\alpha) = \alpha_1 + \alpha_2 - \frac{1}{2}[\alpha_1\alpha_1 K_{11} - \alpha_1\alpha_2 K_{12} - \alpha_2\alpha_1 K_{21} + \alpha_2\alpha_2 K_{22}]$$

$$s.t \quad \alpha_1 - \alpha_2 = 0$$

$$C \geq \alpha_1 \geq 0$$

$$C \geq \alpha_2 \geq 0$$

If we let K_{ij} be the linear kernel function of the two samples i and j, we obtain the kernel values as: $K_{11} = \left\|\boldsymbol{x}^{(1)}\right\|^2 = 5$, $K_{12} = K_{21} = \langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle = 1$, and $K_{22} = \left\|\boldsymbol{x}^{(2)}\right\|^2 = 2$. Plugging these into the above primal problem, we obtain the simplified form:

$$\max L(\alpha) = \alpha_1 + \alpha_2 - \frac{1}{2}[5\alpha_1\alpha_1 - 2\alpha_1\alpha_2 + 2\alpha_2\alpha_2]$$

$$s.t \quad \alpha_1 - \alpha_2 = 0$$

$$C \geq \alpha_1 \geq 0$$

$$C \geq \alpha_2 \geq 0$$

■

4) Prove the following statements using the KKT conditions for soft-SVM.

$$\alpha_i = 0 \Leftrightarrow y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1. \qquad (1)$$

$$0 < \alpha_i < C \Leftrightarrow y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) = 1. \qquad (2)$$

$$\alpha_i = C \Leftrightarrow y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \leq 1. \qquad (3)$$

[*Hints: Refer to the KKT conditions of soft-SVM and then discuss the cases when $\alpha_i = 0$, when $0 < \alpha_i < C$, and when $\alpha_i = C$. The conclusions will be used in the SMO implementation. Refer to the SMO section of the lecture notes for more clues.*]

(Proofs): The soft-SVM primal problem is stated as follows:

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi_i, \ \ \forall i$$

$$\xi_i \geq 0, \ \ \forall i$$

The corresponding Lagrangian function is

$$L(\boldsymbol{w}, b, \xi, \alpha, \mu) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\left(y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) - 1 + \xi_i\right) - \sum_{i=1}^{m}\mu_i\xi_i$$

Where α_i are Lagrange multipliers that correspond to the constraints $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi_i$ and μ_i are Lagrange multipliers that correspond to the constraints $\xi_i \geq 0$.

Proving (1):

Given $\alpha_i = 0$, by complementary slackness, we know that the constraint $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right)$ is not active, indicating that $\xi_i = 0$, and the following condition must hold,

$$\alpha_i\left(y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) - 1 + \xi_i\right) = 0 \qquad (complementary\ slackness).$$

Therefore,

$$y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) - 1 + \xi_i \geq 0$$

$$\longrightarrow y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) \geq 1.$$

■

Proving (2):

Given $0 < \alpha_i < C$, by complementary slackness, the constraint $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right)$ must be active and due to the relationship $C - \alpha_i - \mu_i = 0$, we know that $\mu_i > 0$, indicating that $\xi_i = 0$ (by the KKT condition $\mu_i\xi_i = 0$). Thus, we have

$$y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) - 1 + \xi_i = 0$$

$$\longrightarrow y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) = 1.$$

■

Proving (3):

Given $\alpha_i = C$, by the relationship $C - \alpha_i - \mu_i = 0$, we know that $\mu_i = 0$, indicating that $\xi_i > 0$ (by the KKT condition $\mu_i\xi_i = 0$). By complimentary slackness, we also know that the constraint $y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) - 1 + \xi_i \geq 0$ must be active. Thus, we have

$$y^{(i)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b\right) = 1 - \xi_i$$

Since $\xi_i > 0$, we can see that

$$y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \leq 1.$$

■

5) (Graduate Only) Prove that there is a function $\psi([x_1, x_2])$, that maps from the input space $\mathbb{R}^2$ to a higher dimensional space $\mathbb{R}^n$, $n > 2$, so that the polynomial kernel $k(\boldsymbol{x}, \boldsymbol{z}) = (\langle \boldsymbol{x}, \boldsymbol{z} \rangle + 1)^2$ can be written as $\langle \psi(\boldsymbol{x}), \psi(\boldsymbol{z}) \rangle$. [*Hints: Refer to PRML Eq. (6.12) for help and write down the formula of the function* ψ.]

(Proof): Recall that a kernel function is defined as the following:

$$k(\boldsymbol{x}, \boldsymbol{z}) = \langle \psi(\boldsymbol{x}), \psi(\boldsymbol{z}) \rangle = \psi(\boldsymbol{x})^T\psi(\boldsymbol{z}).$$

If we let $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^2$, then given the polynomial kernel $k(\boldsymbol{x}, \boldsymbol{z}) = (\langle \boldsymbol{x}, \boldsymbol{z} \rangle + 1)^2$, we have

$$k(\boldsymbol{x}, \boldsymbol{z}) = (\langle \boldsymbol{x}, \boldsymbol{z} \rangle + 1)^2 = (\boldsymbol{x}^T\boldsymbol{z} + 1)^2$$
$$= (x_1z_1 + x_2z_2 + 1)^2$$
$$= x_1^2z_1^2 + x_1x_2z_1z_2 + x_1z_1 + x_1x_2z_1z_2 + x_2^2z_2^2 + x_2z_2 + x_1z_1 + x_2z_2 + 1$$
$$= x_1^2z_1^2 + x_2^2z_2^2 + 2x_1x_2z_1z_2 + 2x_1z_1 + 2x_2z_2 + 1$$
$$= \left[x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1\right]\left[z_1^2, z_2^2, \sqrt{2}z_1z_2, \sqrt{2}z_1, \sqrt{2}z_2, 1\right]^T$$
$$= \psi(\boldsymbol{x})^T\psi(\boldsymbol{z})$$

Where the feature mapping takes the form

$$\psi(\boldsymbol{x}) = \psi([x_1, x_2]) = \left[x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1\right]^T$$

Therefore, we can see that there exists a function $\psi([x_1, x_2])$ that maps from the input space $\mathbb{R}^2$ to the higher dimensional space $\mathbb{R}^6$.■