Assignment 2
AI2000
Foundations of Machine Learning

Darpan Gaur
CO21BTECH11004

Problem 1

Margin boundaries are defined as:

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_+ + b &= 1 \quad \text{(Positive Margin Boundary)} \\ \mathbf{w}^T\mathbf{x}_- + b &= -1 \quad \text{(Negative Margin Boundary)} \end{aligned} \tag{1}$$

Now, margin becomes:

$$\rho = (+1) * \frac{\mathbf{w}^T\mathbf{x}_+ + b}{\|\mathbf{w}\|} + (-1) * \frac{\mathbf{w}^T\mathbf{x}_- + b}{\|\mathbf{w}\|} = \frac{2}{\|\mathbf{w}\|} \tag{2}$$

To find maximum margin hyperplane, we need to maximize ρ, and solve:

$$\begin{aligned} &\max_{\mathbf{w},b} \quad \frac{2}{\|\mathbf{w}\|} \text{ or } \min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1, \quad \forall i, where \quad y_i \in \{-1, +1\} \end{aligned} \tag{3}$$

If we replace $y_i \in \{-1, +1\}$ with $y_i \in \{\gamma, -\gamma\}$, then the margin boundaries will be:

$$\begin{aligned} \mathbf{w}^T\mathbf{x}_+ + b &= \gamma \quad \text{(Positive Margin Boundary)} \\ \mathbf{w}^T\mathbf{x}_- + b &= -\gamma \quad \text{(Negative Margin Boundary)} \end{aligned} \tag{4}$$

Now, margin becomes:

$$\rho = \gamma * \frac{\mathbf{w}^T\mathbf{x}_+ + b}{\|\mathbf{w}\|} + (-\gamma) * \frac{\mathbf{w}^T\mathbf{x}_- + b}{\|\mathbf{w}\|} = \frac{2\gamma}{\|\mathbf{w}\|} \tag{5}$$

To find maximum margin hyperplane, we need to maximize ρ, and solve:

$$\begin{aligned} &\max_{\mathbf{w},b} \quad \frac{2\gamma}{\|\mathbf{w}\|} \text{ or } \min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 \\ &\text{subject to} \quad y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq \gamma, \quad \forall i, where \quad y_i \in \{-\gamma, +\gamma\} \end{aligned} \tag{6}$$

Here, margin is scaled by γ. But our optimization problem remains the same, i,e, we need to maximize margin i.e, minimize $\|\mathbf{w}\|^2$. Hence solution for the maximum margin hyperplane remains the same.

Problem 2

The half-margin of maximum-margin SVM defined by ρ, i.e., $\rho = \frac{1}{\|\mathbf{w}\|}$.
The optimization problem for maximum-margin SVM is:

$$\begin{aligned} \max_{\mathbf{w},b} \quad & \frac{1}{\|\mathbf{w}\|} \text{ or } \min_{\mathbf{w},b} \quad \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1, \quad \forall i, where \quad y_i \in \{-1, +1\} \end{aligned} \tag{7}$$

$$L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\mathbf{w}^T\mathbf{x_i} + b) - 1] \tag{8}$$

Can solve for $\mathbf{w}$, b as function of α.

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} = 0 \quad & \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x_i} \\ \frac{\partial L}{\partial b} = 0 \quad & \Rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{9}$$

Substitute $\mathbf{w}$ and b back into L to get the dual optimization problem:

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x_i}^T \mathbf{x_j} \\ \text{subject to} \quad & \alpha_i \geq 0, \quad \forall i \\ & \sum_{i=1}^{n} \alpha_i y_i = 0 \end{aligned} \tag{10}$$

Say, (x_j, y_j) is the support vector, then $w^T x^j + b = y_j$

$$b = y_j - w^T x^j = y_j - \sum_{i=1}^{n} \alpha_i y_i x_i^T x_j$$

Taking sum by multiplying with $\alpha_j.y_j$ on both sides, we get:

$$\sum_{j=1}^{n} \alpha_j y_j b = \sum_{j=1}^{n} \alpha_j y_j^2 - \sum_{j=1}^{n}\sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{11}$$

$$\implies \sum_{j=1}^{n} \alpha_j - \|\mathbf{w}\|^2 = 0 \implies \sum \alpha_j = \|\mathbf{w}\|^2$$

$$\implies \sum_{j=1}^{n} \alpha_j = \frac{1}{\rho^2}$$

Problem 3

(a) $k(x,z) = k_1(x,z) + k_2(x,z)$

Let k_1 has corresponding feature map ϕ_1 and k_2 has corresponding feature map ϕ_2,
Then, $k_1(x,z) = \langle \phi_1(x), \phi_1(z) \rangle$ and $k_2(x,z) = \langle \phi_2(x), \phi_2(z) \rangle$
For all x and z,

$$k(x,z) = k_1(x,z) + k_2(x,z) = \langle \phi_1(x), \phi_1(z) \rangle + \langle \phi_2(x), \phi_2(z) \rangle$$

$$k(x,z) = \langle \phi_1(x) + \phi_2(x), \phi_1(z) + \phi_2(z) \rangle$$

As $k(x,z)$ is represented using inner product of concatenation of feature maps ϕ_1 and ϕ_2, hence, $k(x,z)$ is a valid kernel.

(b) $k(x,z) = k_1(x,z) \cdot k_2(x,z)$

Let k_1 has corresponding feature map ϕ^1 and k_2 has corresponding feature map ϕ^2,
Then, $k_1(x,z) = \langle \phi^1(x), \phi^1(z) \rangle$ and $k_2(x,z) = \langle \phi^2(x), \phi^2(z) \rangle$
For all x and z, where $x, z \in \mathbb{R}^d$,

$$k(x,z) = k_1(x,z) \cdot k_2(x,z) = \langle \phi^1(x), \phi^1(z) \rangle \cdot \langle \phi^2(x), \phi^2(z) \rangle$$

$$k(x,z) = \left(\sum_{i=1}^{d} \phi_i^1(x)\phi_i^1(z) \right) \cdot \left(\sum_{j=1}^{d} \phi_j^2(x)\phi_j^2(z) \right)$$

$$k(x,z) = \sum_{i=1}^{d} \sum_{j=1}^{d} \phi_i^1(x)\phi_i^1(z)\phi_j^2(x)\phi_j^2(z)$$

$$k(x,z) = \sum_{i=1}^{d} \sum_{j=1}^{d} \langle \phi_i^1(x)\phi_j^2(x), \phi_i^1(z)\phi_j^2(z) \rangle$$

$$k(x,z) = \langle \phi(x), \phi(z) \rangle, \quad where \quad \phi(x) = \left[\phi_1^1(x)\phi_1^2(x), \phi_1^1(x)\phi_2^2(x), \ldots, \phi_d^1(x)\phi_d^2(x) \right]$$

As $k(x,z)$ is represented using inner product of feature maps ϕ, hence, $k(x,z)$ is a valid kernel.

(c) $k(x,z) = h(k_1(x,z))$

h is a polynomial function with positive coefficients.
Say, h be a d degree polynomial function, then $h(k_1(x,z)) = \sum_{i=0}^{d} a_i k_1(x,z)^i$
$h(k_1(x,z))$ has terms products of form:

- product of kernels, i.e., $k_1(x,z)^i$, which is valid kernel by part (b).
- summation of kernels, i.e., $\sum_{i=0}^{d} a_i k_1(x,z)^i$, which is valid kernel by part (a).
- scaler multiplication of kernel, i.e., $c \cdot k_1(x,z)$.

- addition of constant term.

$$k(x,z) = c \cdot k_1(x,z) = c \cdot \langle \phi_1(x), \phi_1(z) \rangle = \langle \sqrt{c}\phi_1(x), \sqrt{c}\phi_1(z) \rangle$$

Therefore, $k(x,z)$ is a valid kernel for scalar multiplication with $c > 0$ also given positive coefficients. Simlarly for addition of constant it is valid.

Combining results of all four properties, $k(x,z) = h(k_1(x,z))$ is a valid kernel.

(d) $k(x,z) = \exp(k_1(x,z))$

$$\exp(k_1(x,z)) = \sum_{i=0}^{\infty} \frac{k_1(x,z)^i}{i!}$$

$\exp(k_1(x,z))$ has terms products of form:

- product of kernels, i.e., $k_1(x,z)^i$, which is valid kernel by part (b).
- summation of kernels, i.e., $\sum_{i=0}^{\infty} \frac{k_1(x,z)^i}{i!}$, which is valid kernel by part (a).
- scaler multiplication of kernel, i.e., $\frac{1}{i!} \cdot k_1(x,z)$.

In part (c), we shown that all above properties hold.
Hence $k(x,z) = \exp(k_1(x,z))$ is a valid kernel.

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

$$exp(-\frac{\|x-z\|^2}{\sigma^2}) = exp(-\frac{(x-z)(x-z)^T}{\sigma^2}) = exp(-\frac{x^Tx - 2x^Tz + z^Tz}{\sigma^2})$$

$$exp(-\frac{\|x-z\|^2}{\sigma^2}) = exp(-\frac{x^Tx}{\sigma^2}) \cdot exp(\frac{2x^Tz}{\sigma^2}) \cdot exp(-\frac{z^Tz}{\sigma^2})$$

$$exp(-\frac{\|x-z\|^2}{\sigma^2}) = exp(-\frac{\|x\|^2}{\sigma^2}) \cdot exp(\frac{2x^Tz}{\sigma^2}) \cdot exp(-\frac{\|z\|^2}{\sigma^2})$$

- $exp(-\frac{\|x\|^2}{\sigma^2})$ is a valid kernel by part (d).
- $exp(-\frac{\|z\|^2}{\sigma^2})$ is a valid kernel by part (d).
- $exp(\frac{2x^Tz}{\sigma^2})$ is a valid kernel by part (d).
- product of kernels, i.e., $exp(-\frac{\|x-z\|^2}{\sigma^2})$, which is valid kernel by part (b).

Combining results of all four properties, $k(x,z) = exp(-\frac{\|x-z\|^2}{\sigma^2})$ is a valid kernel.

Problem 4

Part (a)

Used linear kernel for SVM model, wiht deafult parameters.

- Accuracy of the model over entire test set is **0.97877**.
- Number of support vectors are **28**.

Part (b)

Trained using first 50, 100, 200, 800 samples using linear kernel with default parameters.

Number of Samples	Accuracy	Number of Support Vectors
50	0.98113	2
100	0.98113	4
200	0.98113	8
800	0.98113	14

Part (c)

Used polynomial kernel with degee = q, C (regPar) = C, gamma = 1,and coef0 = 1.
Training Error = 1 - Accuracy(train)

C	Q = 2	Q = 5
0.0001	0.008969	0.004484
0.001	0.004484	0.004484
0.01	0.004484	0.003844
1	0.003203	0.003203

Test Error = 1 - Accuracy(test)

C	Q = 2	Q = 5
0.0001	0.016509	0.018868
0.001	0.016509	0.021226
0.01	0.018868	0.021226
1	0.018868	0.021226

Number of Support Vectors

C	Q = 2	Q = 5
0.0001	236	26
0.001	76	25
0.01	34	23
1	24	21

- (i) **False** At C = 0.0001, training error at Q=2 is 0.008969, Q=5 is 0.004484.
- (ii) **True** At C = 0.001, number of support vectors at Q=2 is 76, Q=5 is 25.
- (iii) **False** At C=0.01 training error at Q=2 is 0.004484, Q=5 is 0.003844.
- (iv) **False** At C=1, test error at Q=2 is 0.018868, Q=5 is 0.021226.

Part (d)

Used RBF kernel with gamma = 1, C = C.
Table for training error (1-Accuracy(train)), test error (1 - Accuracy(test)), and number of support vectors.

C	Training Error	Test Error	Number of Support Vectors
0.01	0.003844	0.023585	403
1	0.004484	0.021226	31
100	0.003203	0.018868	22
10^4	0.002562	0.023585	20
10^6	0.000641	0.023585	17

- Training error is decreasing with increase in C, and lowest (0.000641) for C = 10^6.
- Test error first decreases and then increases with increase in C, and lowest (0.018868) for C = 100.

Problem 5

Here training error = 1-accuracy(train) and test error = 1-accuracy(test).

(a): Standard run

Trained using linear kernel with default parameters.

- Train error **0.0**.
- Test error **0.024**.
- Number of support vectors are **1084**.

(b): Kernel variations

RBF kernel with gamma = 0.001.

- Train error **0.0**.

- Test error **0.5**.
- Number of support vectors are **6000**.

Polynomial kernel with degree = 2, gamma = 1, coef0 = 1.

- Train error **0.0**.
- Test error **0.021**.
- Number of support vectors are **1755**.

We get same train error for both kernels, but test error is higher for RBF kernel.