Assignment 2

Foundations of Machine Learning
IIT-Hyderabad
Aug-Dec 2021

Submitted by:
Ankita Jain
BM21MTECH14001

Questions: Theory

1. **Support Vector Machines:**

   Statement: In the derivation for the Support Vector Machine, we assumed that the margin boundaries are given by $\mathbf{w} \cdot \mathbf{x} + b = +1$ and $\mathbf{w} \cdot \mathbf{x} + b = -1$.

   To prove: if the $+1$ and -1 on the right-hand side were replaced by some arbitrary constants $+\gamma$ and $-\gamma$ where $\gamma > 0$, the solution for the maximum margin hyperplane is unchanged.

   Proof:
   Method 1:
   The margin from our derivation is defined as:

   $$\text{Margin} = \min_i \frac{y_i[\mathbf{w^T}\phi(\mathbf{x_i}) + b]}{||w||_2} \tag{1}$$

   Let us assume three different formulations of the decision boundary as follows:

   - w,b
   - 7w,7b
   - 9w,9b

We kow that the decision boundary is given as below:

$$y_i[\mathbf{w^T}\phi(\mathbf{x_i}) + b] = 0 \tag{2}$$

Now the decision boundaries for the above three formulations can be written as below:

- $y_i[\mathbf{w^T}\phi(\mathbf{x_i}) + b] = 0$
- $y_i[7\mathbf{w^T}\phi(\mathbf{x_i}) + 7b] = 0$
- $y_i[9\mathbf{w^T}\phi(\mathbf{x_i}) + 9b] = 0$

All the above three formulations can be reduced to one common form:

$$y_i[\gamma\mathbf{w^T}\phi(\mathbf{x_i}) + \gamma b] = 0 \tag{3}$$

$$\gamma(y_i[\mathbf{w^T}\phi(\mathbf{x_i}) + b]) = 0 \tag{4}$$

$$y_i[\mathbf{w^T}\phi(\mathbf{x_i}) + b] = 0 \tag{5}$$

where γ is just a scaling constant. As we can see changing the scaling contant does not change the decision boundary and as a consequence of unchanged decision boundary **the maximum margin hyperplane remains unchanged**.

Hence proven.

Method 2: Due to the introduction of gamma, the change in definition of margin can be observed as,

- for $y_i > 0$
$$y_i[\mathbf{w^T}\phi(\mathbf{x_i}) + b] = +\gamma \tag{6}$$
- for $y_i < 0$
$$y_i[\mathbf{w^T}\phi(\mathbf{x_i}) + b] = -\gamma \tag{7}$$

The maximum margin can be computed as,

$$\frac{+\gamma * (+1)[\mathbf{w^T}\phi(\mathbf{x_i}) + b]}{||w||} + \frac{-\gamma * (-1)[\mathbf{w^T}\phi(\mathbf{x_i}) + b]}{||w||} = \frac{2\gamma}{||w||} \tag{8}$$

Now the dual can be written as,

$$\mathbf{L_p} = \max_{a_i \geq 0} \min_{\mathbf{w},b} \frac{||w||^2}{2\gamma^2} + \sum_i (a_i(\gamma - (y_i[\mathbf{w^T}\phi(\mathbf{x_i}) + b]))) \tag{9}$$

Differentiating eqn(4) w.r.t $\mathbf{w}$ in order to find the solution,

$$\frac{\partial L}{\partial \mathbf{w}} = \frac{||w||}{\gamma^2} - \sum_i a_i y_i \mathbf{x_i} = 0$$

$$\mathbf{w} = \gamma^2 \sum_i a_i y_i \mathbf{x_i} \tag{10}$$

Differentiating eqn(4) w.r.t b in order to find the solution,

$$\frac{\partial L}{\partial \mathbf{b}} = -\sum_i a_i y_i = 0$$

$$\sum_i a_i y_i = 0 \tag{11}$$

Substituting the value of $\mathbf{w}$ from eqn(5) in eqn(4) subject to the constraint in eqn(5) and $a_i > 0$,

$$\begin{aligned}
\mathbf{L_p} &= \frac{(\gamma^2 \sum_i a_i y_i \mathbf{x_i})^2}{2\gamma^2} + \sum_i a_i(\gamma - (y_i[(\boldsymbol{\gamma^2} \sum_{\mathbf{i}} \mathbf{a_i y_i x_i})\phi(\mathbf{x_i}) + b])) \\
&= \frac{\gamma^4}{2} \sum_i \sum_j a_i a_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} + \gamma \sum_i a_i - \frac{\gamma^2}{2} \sum_i \sum_j a_i a_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} \\
&= \gamma \sum_i a_i + \gamma^2 (\frac{\gamma^2}{2} - 1) \sum_i \sum_j a_i a_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} && (12) \\
& && (13)
\end{aligned}$$

In eqn(7), if we substitute $\gamma = \mathbf{1}$ we will get back the solution which we obtained in class,

$$\mathbf{L_p} = \sum_i a_i - \frac{1}{2} \sum_i \sum_j a_i a_j y_i y_j \mathbf{x_i} \cdot \mathbf{x_j} \tag{14}$$

Hence proven.

2. **Support Vector Machines:**

Statement: Consider the half-margin of maximum-margin SVM defined by ρ i.e. $\rho = \frac{1}{||w||}$

To prove: $\frac{1}{\rho^2} = \sum_{i=1}^{N} \alpha_i$
where α_i are the Lagrange multipliers given by the SVM dual.

Proof:
From the statement:

$$\rho = \frac{1}{||w||}$$

$$\Rightarrow \frac{1}{\rho^2} = ||w||^2 \tag{1}$$

We have the primal as follows:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} \alpha_i \{y_i(\mathbf{w^T}\phi(\mathbf{x_i}) + b) - 1\} \tag{2}$$

For the maximum margin of the solution the following condition must be satisfied:

$$\sum_{i=1}^{N} \alpha_i \{y_i(\mathbf{w^T}\phi(\mathbf{x_i}) + b) - 1\} = 0 \tag{3}$$

As a result the 2nd term(responsible for points other than the ones on the margins) from eqn(2) vanishes and the primal can now be written as:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}||w||^2 \tag{4}$$

We know that the dual can be written as:

$$\tilde{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i\alpha_j y_i y_j \mathbf{x_i}\cdot\mathbf{x_j}$$

$$\tilde{L}(\mathbf{w}, b, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}||w||^2 \tag{5}$$

5Now using the primal from eqn(4) and the Dual from eqn(5), we can write:

$$\frac{1}{2}||w||^2 = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}||w||^2$$

$$||w||^2 = \sum_{i=1}^{N} \alpha_i$$

On comparing the above with eqn(6),

$$\frac{1}{\rho^2} = \sum_{i=1}^{N} \alpha_i \tag{6}$$

Hence proven.

3. **Kernels :**

   (a) $k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z}) + k_2(\mathbf{x}, \mathbf{z})$
   Let us define the feature space of k_1 and k_2 as ϕ_1 and ϕ_2.
   Let ϕ be the concatenation of the above mentioned ferature maps, defined by,

$$\phi(\mathbf{x}) = (\phi_1(\mathbf{x},), \phi_2(\mathbf{x}))$$

$$\phi(\mathbf{z}) = (\phi_1(\mathbf{z}), \phi_2(\mathbf{z}))$$

Hence ϕ is the feature space of k.
To understand this more clearly,

$$\begin{aligned}
\phi(\mathbf{x}) \cdot \phi(\mathbf{z}) &= ((\phi_1(\mathbf{x},), \phi_2(\mathbf{x})) \cdot (\phi_1(\mathbf{z}), \phi_2(\mathbf{z})) \\
&= \phi_1(\mathbf{x}) \cdot \phi_1(\mathbf{z}) + \phi_2(\mathbf{x}) \cdot \phi_2(\mathbf{z}) \\
&= k_1(\mathbf{x}, \mathbf{z}) + k_2(\mathbf{x}, \mathbf{z}) \\
&= k(\mathbf{x}, \mathbf{z})
\end{aligned}$$

Hence the given kernel function **k** ***is a valid kernel function***.

(b) $k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z}) k_2(\mathbf{x}, \mathbf{z})$
Let ϕ be the concatenation of the above mentioned ferature maps, defined by,

$$\phi(\mathbf{x}) = (\phi_1(\mathbf{x},), \phi_2(\mathbf{x}))$$

$$\phi(\mathbf{z}) = (\phi_1(\mathbf{z}), \phi_2(\mathbf{z}))$$

Hence ϕ is the feature space of k.
It is important for us to understand that ϕ_1 and ϕ_2 can have different dimensions, for example:

$$\begin{aligned}
\phi_1(\mathbf{x}) &= (\phi_1^1(\mathbf{x}), \phi_1^2(\mathbf{x})) \\
\phi_2(\mathbf{x}) &= (\phi_2^1(\mathbf{x}), \phi_2^2(\mathbf{x}), \phi_2^3(\mathbf{x}))
\end{aligned}$$

then $\phi(\mathbf{x})$ must contain **all six values** after multiplication operation, i.e.

$$\phi(\mathbf{x}) = (\phi_1^1(\mathbf{x})\phi_2^1(\mathbf{x}), \phi_1^2(\mathbf{x})\phi_2^1(\mathbf{x}), \phi_1^1(\mathbf{x})\phi_2^2(\mathbf{x}), \phi_1^2(\mathbf{x})\phi_2^2(\mathbf{x}, \phi_1^1(\mathbf{x})\phi_2^3(\mathbf{x}, \phi_1^2(\mathbf{x}\phi_2^3(\mathbf{x})$$

Once we understand this we can now write:

$$\begin{aligned}
\phi(\mathbf{x}) \cdot \phi(\mathbf{z}) &= \sum_m \phi_m(\mathbf{x})\phi_m(\mathbf{z}) \\
&= \sum_i \sum_j \phi_{1i}(\mathbf{x})\phi_{2j}(\mathbf{x})\phi_{1i}(\mathbf{z})\phi_{2j}(\mathbf{z}) \\
&= (\sum_i \phi_{1i}(\mathbf{x})\phi_{1i}(\mathbf{z}))(\sum_j \phi_{2j}(\mathbf{x})\phi_{2j}(\mathbf{z})) \\
&= \phi_1(\mathbf{x}) \cdot \phi_1(\mathbf{z}) + \phi_2(\mathbf{x}) \cdot \phi_2(\mathbf{z}) \\
&= k_1(\mathbf{x}, \mathbf{z}) k_2(\mathbf{x}, \mathbf{z}) \\
&= k(\mathbf{x}, \mathbf{z})
\end{aligned}$$

Hence the given kernel function **k** ***is a valid kernel function***.

(c) $k(\mathbf{x},\mathbf{z}) = h(k_1(\mathbf{x},\mathbf{z}))$ where h is a polynomial function with positive co-efficients

Since each polynomial term is a product of kernel with a positive coeffiient , the proof follows the proof in the previous solution, i.e. **product of two kernels is a valid kernel**. Hence the given kernel function **k** ***is a valid kernel function***.

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

We know that the exponential function can be expanded as a Taylor series as below:

$$exp(x) = \lim_{i \to \inf}(1 + x + ... + \frac{x^i}{i!})$$

We see that the above is nothing but a sum of polynomial terms.

The proof basically follows the **(a)** and **(c)** proofs. Hence the given kernel function **k** ***is a valid kernel function***

(e) $k(x,z) = exp(\frac{||\mathbf{x}-\mathbf{z}||_2^2}{\sigma^2})$

Let us define a valid kernel function (as proven in the previous part exponential is a valid kernel function),

$$k_1(\mathbf{x},\mathbf{z}) = exp(\frac{2(\mathbf{x}\cdot\mathbf{z})}{\sigma^2})$$

Let us define a feature space,

$$\phi_1(x) = exp(\frac{||\mathbf{x}||_2^2}{\sigma^2})$$

of a valid kernel function (as explpained in previous parts multiplication of valid kernels result in a valid kernel),

$$\begin{aligned} k_2(\mathbf{x},\mathbf{z}) &= \phi_1(\mathbf{x})\cdot\phi_1(\mathbf{z}) \\ &= exp(\frac{||\mathbf{x}||_2^2}{\sigma^2})exp(\frac{||\mathbf{z}||_2^2}{\sigma^2}) \\ &= exp(\frac{||\mathbf{x}||_2^2 + ||\mathbf{z}||_2^2}{\sigma^2}) \end{aligned}$$

Now we can write the following.

$$\begin{aligned} k(\mathbf{x},\mathbf{z}) &= exp(\frac{||\mathbf{x}-\mathbf{z}||_2^2}{\sigma^2}) \\ &= exp(\frac{(\mathbf{x}-\mathbf{z})^\mathbf{T}(\mathbf{x}-\mathbf{z})}{\sigma^2}) \\ &= exp(\frac{||\mathbf{x}||_2^2 + ||\mathbf{z}||_2^2 - 2(\mathbf{x}\cdot\mathbf{z})}{\sigma^2}) \\ &= exp(\frac{||\mathbf{x}_2^2|| + ||\mathbf{z}_2^2||}{\sigma^2})exp(\frac{2(\mathbf{x}\cdot\mathbf{z})}{\sigma^2}) \\ &= k_2(\mathbf{x},\mathbf{z})k_1(\mathbf{x},\mathbf{z}) \end{aligned}$$

Hence as **k** is a product of two valid kernels, the given kernel function ***is a valid kernel function***.

Questions: Programming

4. **SVMs :**

output.txt

(a)
```
Kernel: linear
Number of Support Vectors: 28
Test Accuracy: 0.9787735849056604
```

output.txt

(b)
```
Kernel: linear

No. of training samples: 50
Number of Support Vectors: 2
Test Accuracy: 0.9811320754716981

No. of training samples: 100
Number of Support Vectors: 4
Test Accuracy: 0.9811320754716981

No. of training samples: 200
Number of Support Vectors: 8
Test Accuracy: 0.9811320754716981

No. of training samples: 800
Number of Support Vectors: 14
Test Accuracy: 0.9811320754716981
```

output.txt

(c)
```
Kernel: poly
i: FALSE
ii: TRUE
iii: FALSE
iv: FALSE
```

output.txt

(d)
```
Kernel: rbf
C: 0.01
Train Error: 0.0038436899423446302
Test Error: 0.02358490566037741

C: 1
Train Error: 0.004484304932735439
Test Error: 0.021226415094339646

C: 100
Train Error: 0.0032030749519538215
Test Error: 0.018867924528301883

C: 10000
Train Error: 0.002562459961563124
Test Error: 0.02358490566037741

C: 1000000
Train Error: 0.0006406149903908087
Test Error: 0.02358490566037741

Min train error corresponds to C =  [1000000]
Min test error corresponds to C =  [100]
```

5. **SVMs (cont):**

output.txt

(a)
```
Kernel: linear
Number of Support Vectors: 1084
Train Error: 0.0
Test Error: 0.02400000000000002
```

output.txt

(b)
```
Kernel: rbf
Number of Support Vectors: 6000
Train Error: 0.0
Test Error: 0.5

Kernel: poly
Number of Support Vectors: 1332
Train Error: 0.0004999999999999449
Test Error: 0.020000000000000018

MIN TRAINING ERROR: [['RBF', 0.0]]
MIN TESTING ERROR: [['Poly', 0.020000000000000018]]
```