Assignment 2

Adarsh Sai - AI20BTECH11001

1 SUPPORT VECTOR MACHINES:

Support Vector Machines: (4 marks) In the derivation for the Support Vector Machine, we assumed that the margin boundaries are given by w.x+b = +1 and w.x+b = -1. Show that, 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. (You can show this for the hard-margin SVM without any slack variables.

Solution: On replacing +1 and -1 with γ and $-\gamma$ respectively, the SVM dual becomes

$$\max_{\overline{\alpha}\geq 0} \min_{\overline{w},b} \frac{1}{2}\left\|\overline{w}\right\|^2 - \sum_i \alpha_i \left[\left(\overline{w}\cdot\overline{x_i}+b\right)y_i - \gamma\right] \tag{1.0.1}$$

$$\max_{\overline{\alpha}\geq 0} \min_{\overline{w},b} \frac{1}{2}\left\|\overline{w}\right\|^2 - \sum_i \alpha_i \left[\left(\overline{w}\cdot\overline{x_i}+b\right)y_i - 1\right] - \sum_i \alpha_i\left(1-\gamma\right) \tag{1.0.2}$$

Solving for optimal w,b as a function of α

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \tag{1.0.3}$$

$$\sum_i \alpha_i y_i = 0 \tag{1.0.4}$$

Substituting these in (1.0.2) gives

$$\max_{\overline{\alpha}\geq 0, \sum_i \alpha_i y_i = 0} \sum_i \gamma\alpha_i - \frac{1}{2}\sum_{i,j} y_i y_j \alpha_i \alpha_j \left(\overline{x_i}\cdot\overline{x_j}\right) \tag{1.0.5}$$

The optimal $\overline{\alpha}$ is scaled by γ. So optimal **w** and b are also scaled by γ. Then the equation of the maximum margin hyperplane is given by

$$\gamma\mathbf{w}\cdot\mathbf{x} + \gamma b = 0 \tag{1.0.6}$$

$$\mathbf{w}\cdot\mathbf{x} + b = 0 \tag{1.0.7}$$

(1.0.7) is same as the hyperplane for the case of +1 and -1.
$\therefore$ The maximum margin hyperplane is unchanged.

2 SUPPORT VECTOR MACHINES:

Support Vector Machines: (4 marks) Consider the half-margin of maximum-margin SVM defined by ρ, i.e. $\rho = \frac{1}{||w||}$. Show that ρ is given by:

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

where α_i are the Lagrange multipliers given by the SVM dual (as on Slide 30 of the SVM lecture uploaded on Piazza). (Hint: The answer involves just 3-4 steps, if you are thinking of something longer, re-think!)
Solution:

$$\frac{1}{\rho^2} = ||\mathbf{w}||^2 = \mathbf{w}\mathbf{w}^\top \tag{2.0.1}$$

The optimal $\mathbf{w}$ is given by

$$\mathbf{w} = \sum_{i}^{N} \alpha_i y_i \mathbf{x_i} \tag{2.0.2}$$

$$\sum_{i}^{N} \alpha_i y_i = 0 \tag{2.0.3}$$

Let k be the number of support vectors. WLOG assume $\alpha_1, ..., \alpha_k$ be the Lagrangian multipliers of the support vectors. From definition $\alpha_{k+1} = .. = \alpha_N = 0$. So the optimal $\mathbf{w}$ can be written as

$$\mathbf{w} = \sum_{i}^{k} \alpha_i y_i \mathbf{x_i} \tag{2.0.4}$$

$$\mathbf{w}^\top = \sum_{i}^{k} \alpha_i y_i \mathbf{x_i}^\top \tag{2.0.5}$$

For support vectors

$$y_i[\mathbf{w}\mathbf{x_i}^\top + b] = 1 \tag{2.0.6}$$

$$\mathbf{w}\mathbf{x_i}^\top = \frac{1}{y_i} - b \tag{2.0.7}$$

$$\mathbf{w}\left(\alpha_i y_i \mathbf{x_i}^\top\right) = \alpha_i - b\left(\alpha_i y_i\right) \tag{2.0.8}$$

$$\text{Summing from i=1 to i=k} \tag{2.0.9}$$

$$\mathbf{w}\left(\sum_{i}^{k} \alpha_i y_i \mathbf{x_i}^\top\right) = \sum_{i=1}^{k} \alpha_i - b\sum_{i=1}^{k} \alpha_i y_i \tag{2.0.10}$$

$$\mathbf{w}\mathbf{w}^\top = \sum_{i=1}^{k} \alpha_i \tag{2.0.11}$$

$$= \sum_{i=1}^{N} \alpha_i \tag{2.0.12}$$

3 KERNALS:

Kernels: (5 marks) Let k_1 and k_2 be valid kernel functions. Comment about the validity of the following kernel functions, and justify your answer with proof or counter-examples as required:

1) $k(x, z) = k_1(x, z) + k_2(x, z)$
2) $k(x, z) = k_1(x, z)k_2(x, z)$
3) $k(x, z) = h(k_1(x, z))$, where h is a polynomial function with positive co-efficients.
4) $k(x, z) = \exp\left(\frac{-\|x-z\|_2^2}{\sigma^2}\right)$

Solution:

1) $k(x, z) = k_1(x, z) + k_2(x, z)$
   Since k_1, k_2 are kernel functions they can be decomposed as
$$k_1(x, z) = \Phi_1(x) \cdot \Phi_1(z) \tag{3.0.1}$$
$$k_2(x, z) = \Phi_2(x) \cdot \Phi_2(z) \tag{3.0.2}$$
   Let $\Phi(x), \Phi(z)$ be defined as
$$\Phi(x) = \begin{pmatrix} \Phi_1(x) & \Phi_2(x) \end{pmatrix} \tag{3.0.3}$$
$$\Phi(z) = \begin{pmatrix} \Phi_1(z) & \Phi_2(z) \end{pmatrix} \tag{3.0.4}$$
   Then the dot product of $\Phi(x), \Phi(z)$ is given by
$$\Phi(x) \cdot \Phi(z) = \begin{pmatrix} \Phi_1(x) & \Phi_2(x) \end{pmatrix} \cdot \begin{pmatrix} \Phi_1(z) & \Phi_2(z) \end{pmatrix} \tag{3.0.5}$$
$$= \Phi_1(x) \cdot \Phi_1(z) + \Phi_2(x) \cdot \Phi_2(z) \tag{3.0.6}$$
$$= k_1(x, z) + k_2(x, z) \tag{3.0.7}$$
$$= k(x, z) \tag{3.0.8}$$
   $\implies k(x, z)$ can be decomposed. So it is a kernel function.
2) $k(x, z) = k_1(x, z) k_2(x, z)$
   The gram matrix $\mathbf{K}$ for $k(x, z)$ is element by element product of $\mathbf{K_1}$ and $\mathbf{K_2}$. Since $\mathbf{K_1}$ and $\mathbf{K_2}$ are symmetric and positive definite, $\mathbf{K}$ is also symmetric and positive definite. Therefore $k(x, z)$ is a kernel function.
3) $k(x, z) = h(k_1(x, z))$
   $k(x, z)$ is sum of kernel functions. From 3.1 and 3.2 it follows that $k(x, z)$ is also a kernel function.
4) $k(x, z) = exp(k_1(x, z))$
$$e^x = \lim_{n\to\infty}\left(1 + \frac{x}{1!} + \frac{x^2}{2!} + \ldots.. + \frac{x^n}{n!}\right) \tag{3.0.9}$$
   Since (3.0.9) is a polynomial with positive coefficients. So from 3.3 we can say $k(x, z)$ is also a kernel function.
5) $k(x, z) = exp\left(\frac{-\|x-z\|^2}{\sigma^2}\right)$

$$
\begin{aligned}
k(x,z) &= exp\left(\frac{-\|x\|^2 - \|z\|^2 + 2x^\top z}{\sigma^2}\right) && (3.0.10)\\
&= exp\left(\frac{-\|x\|^2}{\sigma^2}\right) exp\left(\frac{-\|z\|^2}{\sigma^2}\right) exp\left(\frac{2x^\top z}{\sigma^2}\right) && (3.0.11)\\
&= exp\left(\frac{-\|x\|^2}{\sigma^2}\right) exp\left(\frac{-\|z\|^2}{\sigma^2}\right) \Phi(x)\cdot\Phi(z) && (3.0.12)\\
&= \left(exp\left(\frac{-\|x\|^2}{\sigma^2}\right)\Phi(x)\right)\cdot\left(exp\left(\frac{-\|z\|^2}{\sigma^2}\right)\Phi(z)\right) && (3.0.13)\\
&= \Phi'(x)\cdot\Phi'(z) && (3.0.14)
\end{aligned}
$$

$\implies$ $k(x,z)$ is also a kernel function.

Programming Questions

4 SVMs

1) Accuracy: 0.9787735849056604
   Number of support vectors: 28
2) a) Accuracy using first 50 samples: 0.9811320754716981
      Number of support vectors: 2
   b) Accuracy using first 100 samples: 0.9811320754716981
      Number of support vectors: 4
   c) Accuracy using first 200 samples: 0.9811320754716981
      Number of support vectors: 8
   d) Accuracy using first 800 samples: 0.9811320754716981
      Number of support vectors: 14
3) a) FALSE
   b) TRUE
   c) FALSE
   d) FALSE

4) Train error is least for $C = 10^6$.
   Test error is least for C = 100

$$\text{Train error C} = 0.01 : 0.0038436899423446302$$
$$\text{Test error C} = 0.01 : 0.02358490566037741$$
$$\text{Train error C} = 1 : 0.004484304932735439$$
$$\text{Test error C} = 1 : 0.021226415094339646$$
$$\text{Train error C} = 100 : 0.0032030749519538215$$
$$\text{Test error C} = 100 : 0.018867924528301883$$
$$\text{Train error } C = 10^4 : 0.002562459961563124$$
$$\text{Test error } C = 10^4 : 0.02358490566037741$$
$$\text{Train error } C = 10^6 : 0.0006406149903908087$$
$$\text{Test error } C = 10^6 : 0.02358490566037741$$

5 SVMs (contd)

1) **Standard run:**

train error: 0.0
test error: 0.024000000000000002
Number of SV's: 1084

2) **Kernel Variations:**

   a) **RBF:**

train error: 0.0
test error: 0.5
Number of SV's: 6000

   b) **Polynomial:**

train error: 0.0004999999999999449
test error: 0.020000000000000018
Number of SV's: 1332