Machine Learning: Assignment 8
Harshita Agarwala

December 14, 2017

1 Problem 1

A soft margin SVM model allows errors to be present. This means that there might be data points inside the margin or the data may not be linearly separable.
Hence, with errors included, the soft-margin SVM might misclassify some data points of the training set.

2 Problem 2

The cost function in a soft-margin SVM is given by: $f_0 = \frac{1}{2}w^T w + C\sum\limits_{i=1}^{N} \xi_i$
Now $\xi_i \geq 0$ for every x_i in the training sample is the distance of how far the margin is violated. These distances have to be penalized in the model to reduce the number of misclassifications and data points inside the margin. For this the sum of these distances is added in the cost function and hence the new cost function would be reduce $||w||$ and the sum of the distances.
Now, $C > 0$ ensures that the $\xi_i' s$ are penalized. If $C < 0$ then the $\xi_i' s$ are rewarded and if $C = 0$ then the model would be a hard-margin SVM.

3 Problem 3

Now, from Kernel construction rules, we know a kernel of the form $K(x, y) = x^T By$ is a kernel if B is a symmetric and positive-semidefinite matrix. Taking B=I, the identity matrix, we have $K_1(x, y) = x^T y$ is a kernel. Also,
$K_2(x, y) = \phi(x)^T\phi(y)$ where $\phi(x) = \sqrt{c}$ is also a kernel.
Hence, $K_1(x, y) + K_2(x, y) = x^T y + c = K_3(x, y)$ is also a kernel.
Using construction rules again, we know that the product of kernels is a kernel. Hence, $(K_3(x, y))^d = (x^T y + c)^d$ is also a kernel.

4 Problem 4

We should not directly apply the feature transformation to the dataset as the feature space is infinite-dimensional and it would be impossible to carry out the numerical calculations using it. It would be better if we use a kernel instead of the basis function directly.

5 Problem 5

$$\phi_\infty(x) = \begin{bmatrix} e^{-x^2/2\sigma^2} \\ e^{-x^2/2\sigma^2}\frac{x}{\sigma} \\ \frac{1}{\sqrt{2}}e^{-x^2/2\sigma^2}(\frac{x}{\sigma})^2 \\ \vdots \\ \frac{1}{\sqrt{i!}}e^{-x^2/2\sigma^2}(\frac{x}{\sigma})^i \\ \vdots \end{bmatrix}$$

Taking $K(x,y) = \phi_\infty(x)^T\phi_\infty(y)$ we get,

$$\phi_\infty(x)^T\phi_\infty(y) = \begin{bmatrix} e^{-x^2/2\sigma^2} & e^{-x^2/2\sigma^2}\frac{x}{\sigma} & \frac{1}{\sqrt{2}}e^{-x^2/2\sigma^2}(\frac{x}{\sigma})^2 & \cdots & \frac{1}{\sqrt{i!}}e^{-x^2/2\sigma^2}(\frac{x}{\sigma})^i & \cdots \end{bmatrix} \begin{bmatrix} e^{-y^2/2\sigma^2} \\ e^{-y^2/2\sigma^2}\frac{y}{\sigma} \\ \frac{1}{\sqrt{2}}e^{-y^2/2\sigma^2}(\frac{y}{\sigma})^2 \\ \vdots \\ \frac{1}{\sqrt{i!}}e^{-y^2/2\sigma^2}(\frac{y}{\sigma})^i \\ \vdots \end{bmatrix}$$

$$\begin{aligned} \phi_\infty(x)^T\phi_\infty(y) &= e^{-(x^2+y^2)/2\sigma^2} + e^{-(x^2+y^2)/2\sigma^2}\frac{xy}{\sigma^2} + \frac{1}{2}e^{-(x^2+y^2)/2\sigma^2}\Big(\frac{xy}{\sigma^2}\Big)^2 + \cdots + \frac{1}{i!}e^{-(x^2+y^2)/2\sigma^2}\Big(\frac{xy}{\sigma^2}\Big)^i + \cdots \\ &= e^{-(x^2+y^2)/2\sigma^2}\Big[1 + \frac{xy}{\sigma^2} + \frac{1}{2}\Big(\frac{xy}{\sigma^2}\Big)^2 + \cdots + \frac{1}{i!}\Big(\frac{xy}{\sigma^2}\Big)^i + \cdots\Big] \\ &= e^{-(x^2+y^2)/2\sigma^2}\sum_{i=0}^{\infty}\frac{1}{i!}\Big(\frac{xy}{\sigma^2}\Big)^i \\ &= e^{-(x^2+y^2)/2\sigma^2}e^{xy/\sigma^2} \qquad \text{(Using Taylor series of } e^x\text{)} \\ &= exp\Big[-\frac{1}{2\sigma^2}(x^2+y^2-2xy)\Big] \\ &= exp\Big[-\frac{(x-y)^2}{2\sigma^2}\Big] \end{aligned}$$

Therefore we have obtained the Gaussian Kernel, $K(x,y) = exp\Big[-\frac{(x-y)^2}{2\sigma^2}\Big]$
If we regularize the value of σ, then we need not worry about overfitting.

6 Problem 6

Yes, as $\sigma \rightarrow 0$ it is possible to linearly separate the set of data points. However, this leads to overfitting and a bad generalization model.

7 Problem 7

If we map x to a feature space using the basis function ϕ then the distance between points/vectors x, y in feature space becomes $||\phi(x) - \phi(y)||_2$. Now let, $K(x,y) = \phi(x)^T\phi(y)$.

$$\begin{aligned} ||\phi(x) - \phi(y)||_2 &= \sqrt{(\phi(x) - \phi(y))^2} \\ &= \sqrt{(\phi(x) - \phi(y))^T(\phi(x) - \phi(y))} \\ &= \sqrt{\phi(x)^T\phi(x) + \phi(y)^T\phi(y) - 2\phi(x)^T\phi(y)} \\ &= \sqrt{K(x,x) + K(y,y) - 2K(x,y)} \end{aligned}$$

Therefore, the distance in terms of kernels is given by: $[K(x,x) + K(y,y) - 2K(x,y)]^{1/2}$