Homework 3

Komal Niraula (N16417290)

kn2505@nyu.edu

Question 1

Answer:

Given, $k(x,\tilde{x}) = \phi(x)^\top \phi(\tilde{x})$ is a Mercer kernel, where $\phi(x)$ is a feature mapping that maps an input x from the input space to a higher-dimensional feature space.

$S = \{x_1, x_2, \dots, x_n\}$ is a sample of n inputs.

The kernel matrix $K \in R^{n\times n}$ is defined as:

$$K_{i,j} = k(x_i, x_j) = \phi(x_i)^\top \phi(x_j)$$

The quadratic form $c^\top Kc$ can be expanded as: $c^{\top K} c = \sum_{i,j=1}^{n} c_i c_j K_{i,j}$

$$= \sum_{i,j}^{n} c_i c_j \phi(x_i)^\top \phi(x_j)$$

$$c^\top Kc = \left(\sum_{i=1}^{n} c_i \phi(x_i)\right)^\top \left(\sum_{j=1}^{n} c_j \phi(x_j)\right)$$

$$\boldsymbol{c}^\top \boldsymbol{K}\boldsymbol{c} = |\sum_{i=1}^{n} c_i \phi(x_i)\,|^2$$

Since this expression represents a squared norm, this is always non-negative.

$$|\sum_{i=1}^{n} c_i \phi(x_i)\,|^2 \;\geq\; 0$$

This shows that $\boldsymbol{c}^\top \boldsymbol{K}\boldsymbol{c}$ **≥ 0**, proving **K** to be positive semi-definite.

a.

k_1 and k_2 are both Mercer Kernel, meaning their corresponding kernel matrices K_1 and K_2 are PSD. Since $k_1(x, \tilde{x})$ and $k_2(x, \tilde{x})$ are Mercer kernels, there exist feature maps $\phi_1(x)$ and $\phi_2(x)$ such that:

$$k_1(x, \tilde{x}) = \langle \phi_1(x), \phi_1(\tilde{x}) \rangle, \quad k_2(x, \tilde{x}) = \langle \phi_2(x), \phi_2(\tilde{x}) \rangle$$

We can express the new kernel $k(x, \tilde{x})$ as:

$$k(x,\ \tilde{x})\ =\ \alpha \langle\ \phi_1(x),\ \phi_1(\tilde{x})\ \rangle\ +\ \beta \langle\ \phi_2(x),\ \phi_2(\tilde{x})\ \rangle$$

$$k(x, \tilde{x}) = \left\langle \sqrt{\alpha}\phi_1(x), \sqrt{\alpha}\phi_1(\tilde{x}) \right\rangle + \left\langle \sqrt{\beta}\phi_2(x), \sqrt{\beta}\phi_2(\tilde{x}) \right\rangle.$$

Let combined feature map be $\phi(x)$. Then: $\phi(x) = \left(\sqrt{\alpha}\phi_1(x), \sqrt{\beta}\phi_2(x)\right)$

Thus, $k(x, \tilde{x}) = \langle \phi(x), \phi(\tilde{x}) \rangle$

$k(x, \tilde{x})$ is now expressed as a dot product in a feature space and is a valid Mercer kernel. Furthermore, for non-negative scalers $\alpha \geq 0$ and $\beta \geq 0$, the kernel is positive semi-definite.

b.

As in question a, we express $k_1(x, \tilde{x})$ and $k_2(x, \tilde{x})$ in terms of their respective feature maps:

$$k_1(x, \tilde{x}) = \langle \phi_1(x), \phi_1(\tilde{x}) \rangle, k_2(x, \tilde{x}) = \langle \phi_2(x), \phi_2(\tilde{x}) \rangle$$

Their product:

$$k(x, \tilde{x}) = \langle \phi_1(x), \phi_1(\tilde{x}) \rangle \cdot \langle \phi_2(x), \phi_2(\tilde{x}) \rangle$$

Expressing as an inner product in a new feature space:

$$k(x,\ \tilde{x})\ =\ \langle\ \phi_1(x)\ \otimes\ \phi_2(x),\ \phi_1(\tilde{x})\ \otimes\ \phi_2(\tilde{x})\ \rangle$$ ($\otimes$ denotes the tensor product)

since $k(x, \tilde{x})$ is now expressed as a dot product in a valid feature space, it is a valid Mercer kernel. Additionally, since both $k_1(x, \tilde{x})$ and $k_2(x, \tilde{x})$ are positive semi-definite, their product is also positive semi-definite.

c.

There exists a feature map $\phi_1(x)$ such that: $k_1(x, \tilde{x}) = \langle \phi_1(x), \phi_1(\tilde{x}) \rangle$

Let f be a polynomial of degree d with positive coefficients:

$f\big(k_1(x, \tilde{x})\big) = a_0 + a_1 k_1(x, \tilde{x}) + a_2 k_1(x, \tilde{x})^2 + \cdots + a_d k_1(x, \tilde{x})^d$, where $a_i \geq 0$ for all i.

Each term $k_1(x, \tilde{x})^i$ is the inner product of feature maps corresponding to the tensor product of $\phi_1(x)$.

Specifically: $k_1(x, \tilde{x})^i = \langle \phi_1(x)^{\otimes i}, \phi_1(\tilde{x})^{\otimes i} \rangle$, where $\otimes$ represents the tensor product. Thus, we can express $f\big(k_1(x, \tilde{x})\big)$ as a sum of inner products: $f\big(k_1(x, \tilde{x})\big) = \langle \psi(x), \psi(\tilde{x}) \rangle$, where $\psi(x)$ is the combined feature map obtained by concatenating all the tensor products of $\phi_1(x)$.

Since this is an inner product in some feature space, $k(x, \tilde{x}) = f\big(k_1(x, \tilde{x})\big)$ is a Mercer kernel that is still positive semi-definite.

d.

$k_1(x, \tilde{x})$ is a Mercer kernel, meaning there exists a feature map $\phi_1(x)$ such that:

$$k_1(x, \tilde{x}) = \langle \phi_1(x), \phi_1(\tilde{x}) \rangle$$

$$\exp\big(k_1(x, \tilde{x})\big) = 1 + k_1(x, \tilde{x}) + \frac{k_1(x, \tilde{x})^2}{2!} + \frac{k_1(x, \tilde{x})^3}{3!} + \cdots \qquad \text{[Exponential function]}$$

Here, each term $k_1(x, \tilde{x})^n$ is the inner product of higher-order tensor products of the feature maps $\phi_1(x)$.

$$k_1(x, \tilde{x})^n = \langle \phi_1(x)^{\otimes n}, \phi_1(\tilde{x})^{\otimes n} \rangle$$

Thus, we can express the exponential function as a sum of inner products:

$$\exp\big(k_1(x, \tilde{x})\big) = \sum_{n=0}^{\infty} \frac{1}{n!} \langle \phi_1(x)^{\otimes n}, \phi_1(\tilde{x})^{\otimes n} \rangle$$

Since this is an infinite sum of positive semi-definite kernels, the sum itself is a positive semi-definite kernel. So, the exponential function preserves the positive semi-definite of a matrix.

Therefore, $k(x, \tilde{x}) = \exp\big(k_1(x, \tilde{x})\big)$ is a Mercer kernel.

B.

The given kernel is a Gaussian Kernel and φ(x) represents a feature map. For the Gaussian kernel, the feature map $\varphi(x)$ does not have a simple closed-form representation. It maps x to an infinite-dimensional space that can be thought of as representing all polynomial combinations of features.

The exact formula for $\varphi(x)$ is implicit due to the infinite-dimensional nature of the Gaussian kernel. This is why the Gaussian kernel is useful in practice as it avoids explicitly mapping inputs into an infinite-dimensional space by computing only the inner product $K(x,y)$ directly through the kernel function.

Question 2:

Answer:

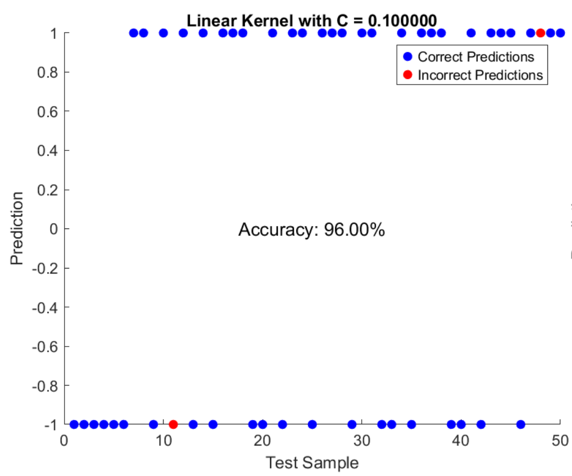
The problem2 folder contains the coding part and images generated.

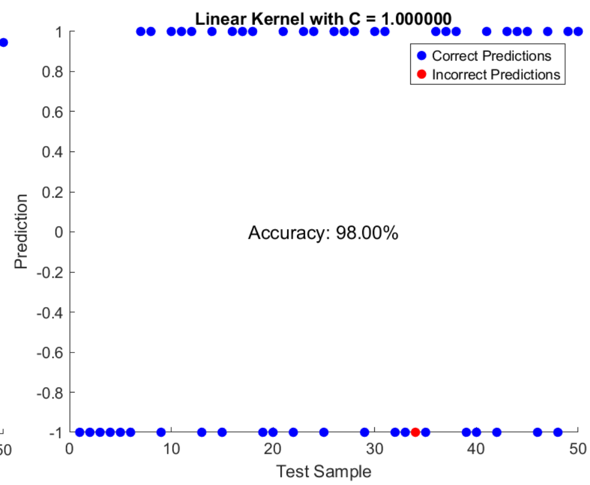
The solution is coded in problem2.m and figure of different kernels prediction are present in 'output_figure' folder.

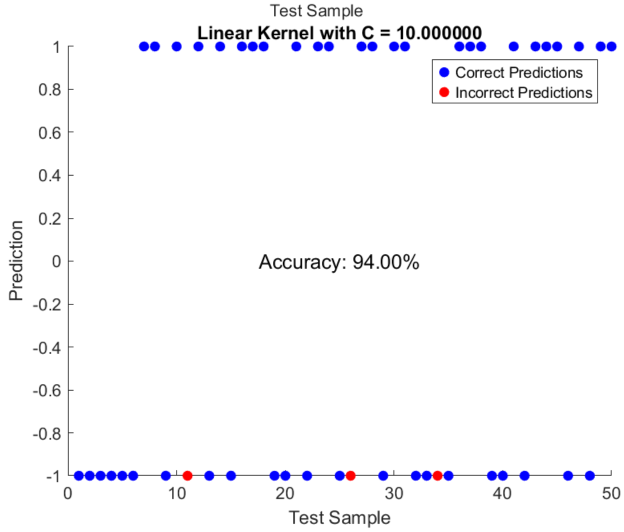
The datasets are randomly split into half for the creation of test and training set. Moreover, the y value (output) 0 is converted to -1, so that the output of SVM with be +1 and -1.

Trying out different combinations

Linear kernel:

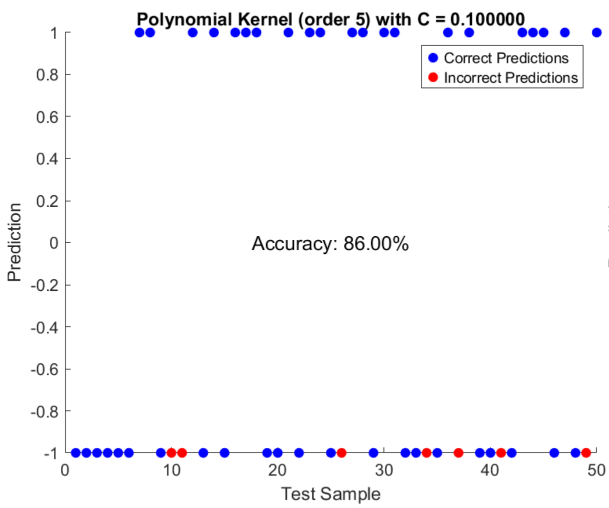


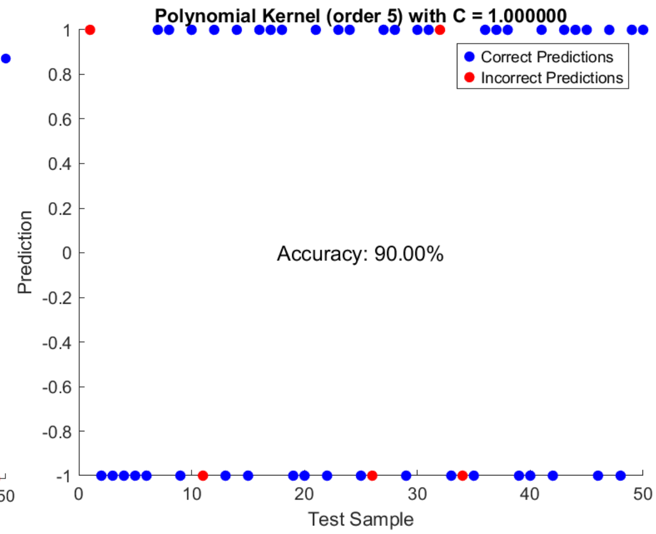


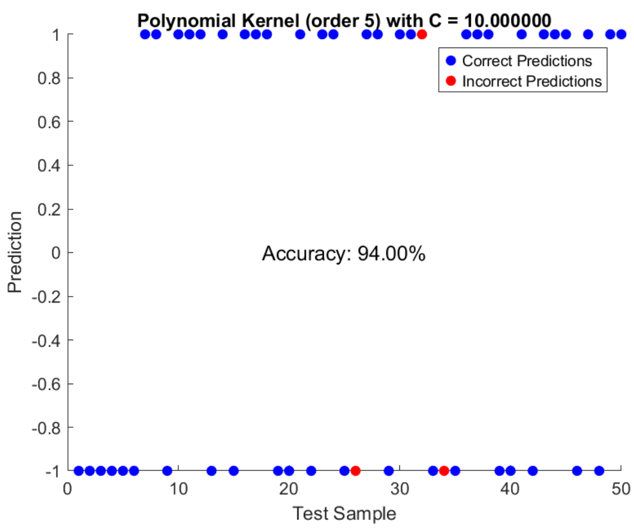


The linear kernel with C = 1 has the largest accuracy, i.e. 98%, while the C = 10 seems to have lowest accuracy of 94% in comparison here. So, we need to test both higher and lower values and identify the best C value for any given dataset.

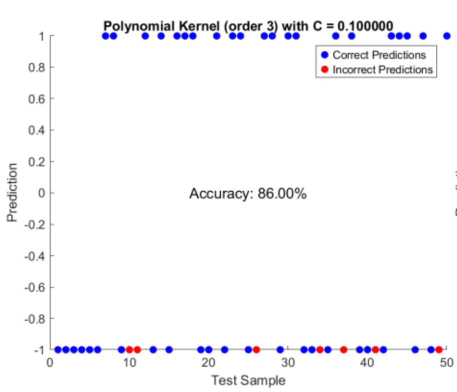
Similarly, using polynomial kernel:

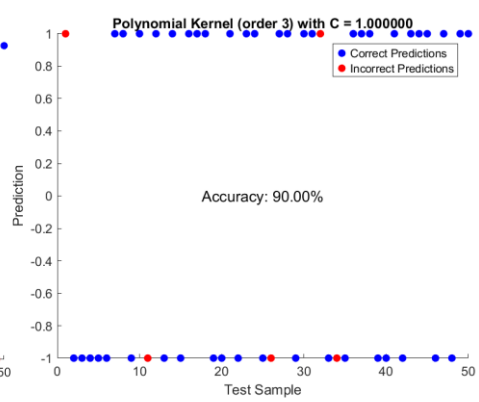


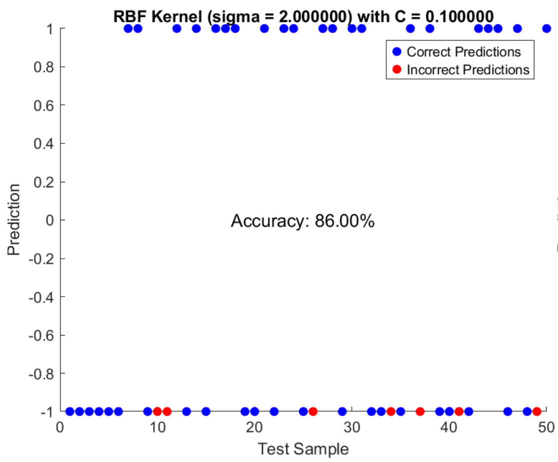


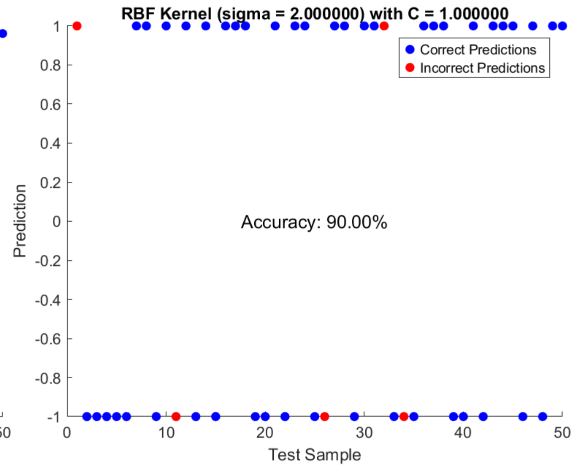


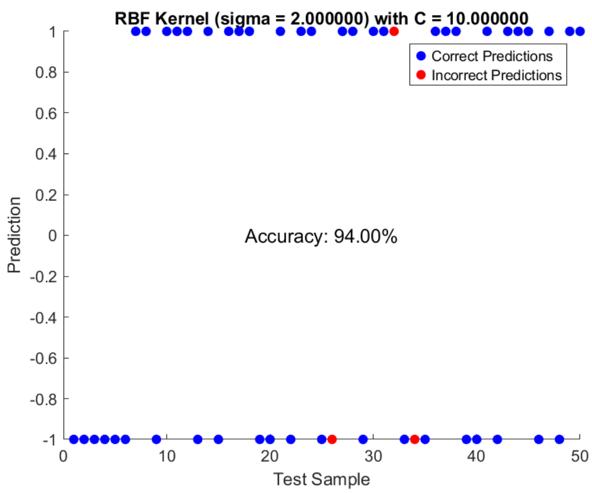
Based on the generated output, we can say that accuracy increases when the C value is increased. Also, in the experiment, the accuracy doesn't change when we change the order of polynomials (1 to 5). It only changes when the C value is changed. However, this is what's observed in the given dataset, need more experiments on different datasets to confirm.



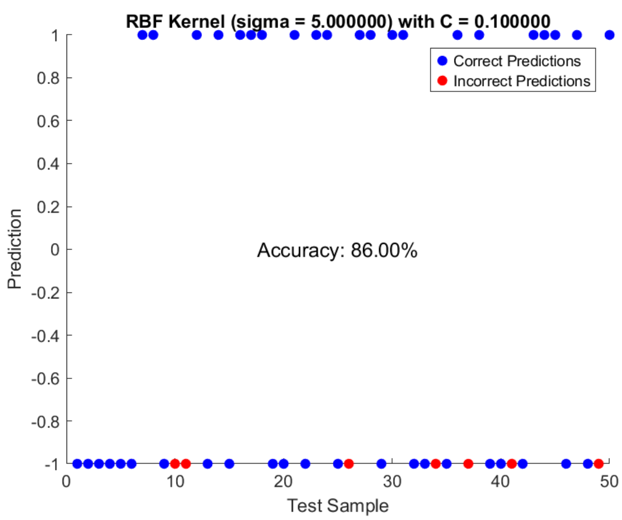


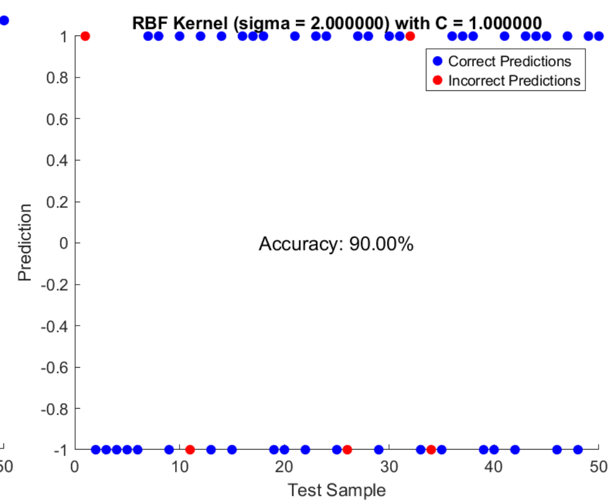






Even in the RBF kernel, C value seems to play important role than other parameters. Like in polynomial kernel, here also the accuracy seems to increase with the increase in C value.





Overall, the highest accuracy was reached by linear kernel i.e. 98% when C value is 1.

Question 3:

Answer:

$X_1, X_2, \ldots\ldots, X_n$, are independent and identically distributed (i.i.d.) samples.

The likelihood function L(α) is the product of the individual PDFs evaluated at each sample x_i for i = 1,2,.......,n:

$$L(\alpha) = \prod_{i=1}^{n} f(x_i|\alpha) = \prod_{i=1}^{n} \alpha^{-2} x_i e^{-x_i/\alpha}$$

$$L(\alpha) = \alpha^{-2n} \prod_{i=1}^{n} x_i \cdot e^{-\sum_{i=1}^{n} x_i/\alpha}$$

To maximize the likelihood, we take the log-likelihood:

$$l(\alpha) = \ln\big(L(\alpha)\big) = \ln\left(\alpha^{-2n} \prod_{i=1}^{n} x_i e^{-\sum_{i=1}^{n} x_i/\alpha}\right)$$

$$= -2n\ln(\alpha) + \sum_{i=1}^{n} \ln(x_i) - \frac{1}{\alpha}\sum_{i=1}^{n} x_i$$

$$\mathrm{l}(\alpha) = -2n\ln(\alpha) - \frac{1}{\alpha}\sum_{i=1}^{n} x_i$$

$$\frac{d\mathrm{l}(\alpha)}{d\alpha} = -\frac{2n}{\alpha} + \frac{1}{\alpha^2}\sum_{i=1}^{n} x_i = 0$$

$$-2n\alpha + \sum_{i=1}^{n} x_i = 0$$

$$\widehat{\alpha} = \frac{1}{2n}\sum_{i=1}^{n} x_i$$

Given sample values $x_1 = 0.25, x_2 = 0.75, x_3 = 1.50, x_4 = 2.50, x_5 = 2.0,$

The sum: $\sum_{i=1}^{5} x_i = 0.25 + 0.75 + 1.50 + 2.50 + 2.0 = 7.0$

The number of samples is 5, so the estimate for α is:

$$\widehat{\alpha} = \frac{1}{2 \times 5} \times 7 = \frac{7}{10} = 0.7$$

Thus, the maximum likelihood estimates for (α) is $\widehat{\alpha} = 0.7$