CS-335 / 337 Assignment-4

Niraj Mahajan - 180050069

Q-1.1) → Property-1 (Proved in class)

If K_1, K_2 are kernels, then $\alpha_1 K_1 + \alpha_2 K_2$ is also a kernel $\forall \alpha_1, \alpha_2 \geq 0$

→ Property-2 (Proved in class)

If K_1, K_2 are kernels, then $K_1 \cdot K_2$ is also a kernel.

→ Property-3 If K is a kernel, K^d is also a kernel, $\forall d \in \mathbb{N}$

Proof by Induction:-
Base Case, $d=1$ $K^1 = K$ (pressumption) is a kernel.

Induction Hypothesis:- K^d is a kernel.
Induction Step: To prove K^{d+1} is a kernel,
$K^{d+1} = K \cdot K^d$

Using Property 2, since K, K^d are valid kernels, then K^{d+1} is a valid kernel
Hence proved.

⇒ Now, we need to prove

$$K(x,y) = \exp\left(-\frac{1}{2\sigma^2}\|x-y\|^2\right) = \exp\left(\frac{-1}{2\sigma^2} x^T x\right).$$
$$.\exp\left(\frac{x^T y}{\sigma^2}\right)$$
$$.\exp\left(\frac{-1}{2\sigma^2} y^T y\right).$$

In the RHS, just consider the middle term.

$$K'(x,y) = \exp\left(\frac{x^T y}{\sigma^2}\right)$$
$$= \sum_{n=0}^{\infty} \frac{(x^T y)^n}{n! \cdot \sigma^{2n}}$$

Now, since $K(x,y) = x^T y$ is a valid kernel, using property 1 and property 3, we get $K'(x,y)$ is also a valid kernel.

Hence, there exist a $\phi(x) : R^n \to H$.

s.t $K'(x,y) = \phi(x)^T \phi(y)$.

Let us define $\phi_{new} = \phi(x) . \exp\left(\frac{-1}{2\sigma^2} x^T x\right)$.

↳ scalar.

∴ using this ϕ_{new}, we can create a kernel

$$\therefore \phi_{new}(x)^T \phi_{new}(y) = \phi(x)^T \phi(y) \cdot \exp\left(\frac{-1}{2\sigma^2} x^T x\right) \cdot \exp\left(\frac{-1}{2\sigma^2} y^T y\right).$$

$$= \exp\left(\frac{x^T y}{\sigma^2}\right) \cdot \exp\left(\frac{-1}{2\sigma^2} x^T x\right) \cdot \exp\left(\frac{-1}{2\sigma^2} y^T y\right).$$

$$= \exp\left(\frac{-1}{2\sigma^2} \left(\|x-y\|^2\right)\right).$$

$$= K(x, y)$$

Hence Proved, the Gaussian (rbf) kernel can be expressed in terms of $\phi^T(x) \cdot \phi(y)$, Hence, rbf kernel is a valid kernel.

Q-1.2b).(i). Best $\sigma = 1$ (minimum errors at $\sigma=1$).

(ii) As we ~~increase~~ decrease the σ, we basically are shrinking the neighborhood of points for performing our prediction.

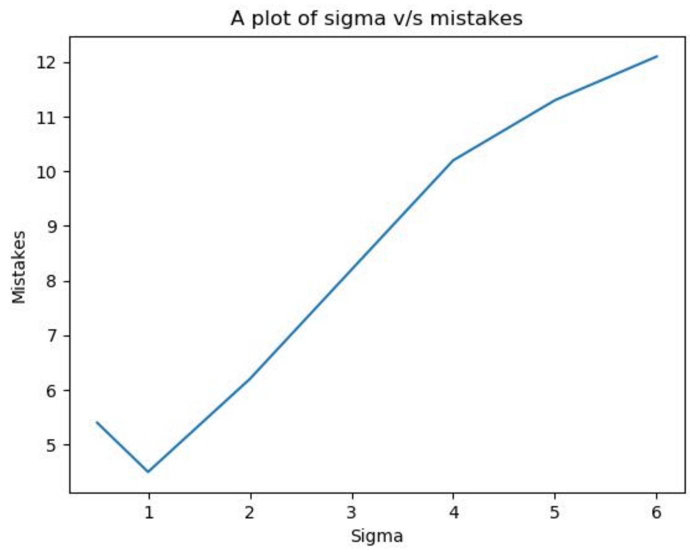
Upto $\sigma=1$, the errors decrease as we are shrinking the neighborhood to consider just a small meaningful set of neighbours.

But below $\sigma=1$, our errors increase as ~~we begin~~ our interval shrinks down tremendously and we begin to overfit on our data.

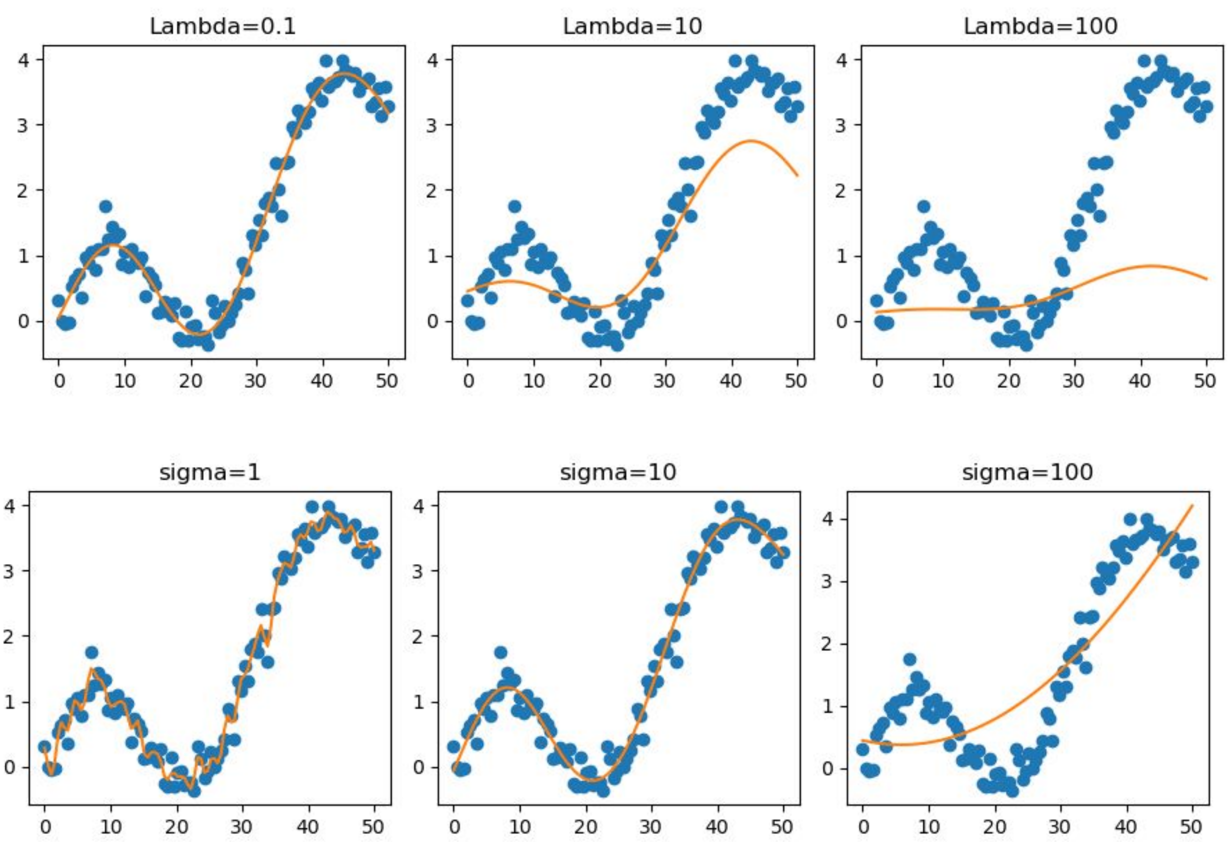
[Plot on the next page].

Assignment 4 Plots: Niraj Mahajan: 180050069

Q1.2 b) (ii)



Q1.2) c)



Q.1.2) c) ii). [Plots were on the previous page].

→ Variation of sigma.

↳ Sigma basically represents the neighborhood that we take around our target point. So, a larger sigma implies a bigger neighbourhood.

→ For sigma = 1, the neighborhod is really small, and as we can see, the model is overfitting on the data (High Variance)

→ For sigma = 100, the neighborhood is really large and the model is under-fitting on the data (High bias).

↳ For sigma = 10, we observe that the model fits just perfectly to the data.

→ Variation of lambda.

→ Higher the lambda, more will be the regularisation penalty, and the variance of the model will decrease (more bias).

↳ At lambda = 0.1, the model fits the data just perfectly, but as we increase lambda, the regularisation penalty forces the model to decrease it's degrees of freedom, and hence, for higher lamba's the curve that is fit by our model is more ~~[illegible]~~ ~~[illegible]~~ rigid / more linear as compared to lower lambdas.

Q-2.1]. Given $K(x, x')$ is valid kernel.

Then there has to exist $\phi : R^m \to H$ s.t

$$K(x, x') = \phi(x)^T \phi(x') \quad -① $$

(i) Consider a function $\phi_{new} : R^m \to H$.

s.t

$$\phi_{new}(x) = \phi(g(x)).$$

$$\therefore \phi_{new}(x)^T \phi_{new}(y) = \phi(g(x))^T \phi(g(y)).$$

Using eq^n-①

$$= K(g(x), g(y)).$$

Hence, $K(g(x), g(y))$ can be represented in terms of inner space product of some ϕ_{new}.

Hence $K_{new}(x,y) = K(g(x), g(y))$ is a valid kernel.

ii) Let $g(x) = \sum_{i=0}^{n} a_i x^i$

Property-1: If K_1, K_2 are valid kernels. then $\alpha_1 K_1 + \alpha_2 K_2$ are valid kernels $\forall \alpha_1, \alpha_2 \geq 0$

[Proved in class]

Property-2. If K_1 & K_2 are valid kernels, then $K_1 \cdot K_2$ is also a valid kernel (proved in class).

Property 3 :- If K_1 is a valid kernel, then $(K_1)^d$ is a valid kernel $\forall d \in N$.

(Proved in part 1).

Using property 1,3 we can say that for ~~any i ∈ {0,...n-1}~~ any $i \in \{0, 1, \cdots n-1\}$.

$a_i K(x,x')^i$ is a valid kernel.

Hence summation over all $i \in N$, $i < n$

$\sum_{i=0}^{i<n} a_i K(x,x')^i$ is also a valid kernel

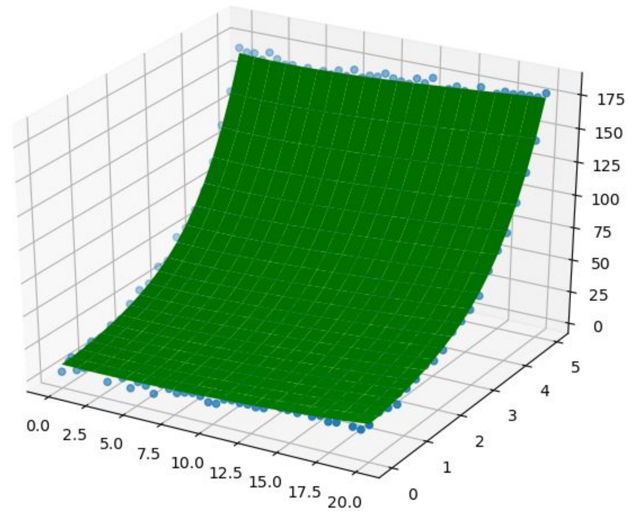
Hence proved. $q(K(x,x'))$ is also a valid kernel

[Q.2.2]

My kernel = $K(x,y) = (1 + x^T y)^4$
This gives an error < 7000 which is desired. [Plot attached on the next page]

Assignment 4 Plots: Niraj Mahajan: 180050069

Q2.2)



Q-3.1] - Data = $\{x^1, x^2 \ldots x^n\}$.

$C_1 = \{x^1, x^2 \ldots x^m\}$ $C_2 = \{x^{m+1}, x^{m+2}, \ldots x^n\}$.

Let ~~the~~ cluster centre be O_1, O_2 respectively s.t,

$$O_1 = \frac{\sum_{i=1}^{m} x^i}{m} \qquad O_2 = \frac{\sum_{i=m+1}^{n} x^i}{(n-m)}$$

~~Let~~ ~~there~~ ~~ex~~

We want to prove existence for a plane $a^T x + b = 0$

s.t $a^T x + b < 0$ $\forall x \in C_1$ $\quad$ $a^T x + b > 0$ $\forall x \in C_2$. — ①

Now, we also know that for any $x \in C_1$ since our solution of O_1, O_2 is optimal,

$$\|x - O_1\|^2 < \|x - O_2\|^2$$

and for any $x \in C_2$ — ②

$$\|x - O_2\|^2 > \|x - O_1\|^2.$$

In order to find a plane, we need to find all such points that are equidistance from O_1, O_2.

$\therefore$ For any x lying on our plane.

$$||x - O_1||^2 = ||x - O_2||^2$$

$$(x - O_1)^T (x - O_1) = (x - O_2)^T (x - O_2)$$

$$\therefore \cancel{x^T x} - x^T O_1 - O_1^T x + O_1^T O_1 = \cancel{x^T x} - x^T O_2 - O_2^T x + O_2^T O_2$$

Note: $O_1^T x = x^T O_1$ and $O_2^T x = x^T O_2$

$$-2x^T O_1 + O_1^T O_1 = -2x^T O_2 + O_2^T O_2.$$

$$2(x^T(O_1 - O_2)) = O_1^T O_1 - O_2^T O_2$$

$$x^T(O_1 - O_2) = \frac{||O_1||^2 - ||O_2||^2}{2}.$$

$$\therefore x^T(O_1 - O_2) - \left(\frac{||O_1||^2 - ||O_2||^2}{2}\right) = 0.$$

where $O_1 = \frac{\sum_{i=1}^{m} x^i}{m}$ and $O_2 = \frac{\sum_{i=m+1}^{n} x^i}{n-m}$

Hence proved, there exists a plan. $a^T x + b$ which separates the two clusters.

~~Q-3.2] (i). Image 1 Cubes~~

~~K=2 → Since we have two clusters, we can just make out the position of cubes, but their orientation / lighting is unclear.~~

~~K=5 → Since the cubes image has a very small number of pixels,~~

Q-3.2] As the number of clusters increase, the
ii) image becomes more and more similar to the original image. At K=2, we just have 2 clusters and this gives a very crude segmentatn ie. cube or no cube (image-1)
• foreground vs background (image 2, 3)

As K increases to 5, more aspects in the image becom clear, like orientatn of cubes in image 1, race track in image 3.

At K=10, the 1st image is nearly restored while the others give a somewhat 'noisy' segmation of the image components.

iii). Some Images (like image-1) have nearly discrete colours, (and fewer colours), and hence, having less clusters can preserve the information in the images.

But some images (like image-2) have continuous and a huge spectrum of colours. Hence it becomes difficult to preserve information in these images with less clusters and we need more clusters for preserving image-data.

Assignment 4 Plots: Niraj Mahajan: 180050069

Q 3.2) (ii)