Lab 4: AI - ML

Anish Deshpande:- 180100013

Part1:

Date :

Assignment / Lab 4

1.1 To prove: $K_\sigma : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R} : K_\sigma(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$

is a valid kernel.

Assumptions (proved in class): ① If $K_1(x,y)$ is a valid kernel, $K_2(x,y)$ is a valid kernel

$\therefore \alpha K_1(x,y) + \beta K_2(x,y) = K(x,y)$ is a valid kernel

$\alpha, \beta \geq 0$.

② The polynomial kernel is valid. i.e. $K(x,y) = x^T y$
$= \langle x,y \rangle$

$\Rightarrow K'(x,y) = (x^T y)^n = (\langle x,y \rangle)^n$ is a valid kernel.

Proof: $\|x-y\|_2^2$ can be written as an inner product.

$\|x-y\|_2^2 = \langle x-y, x-y \rangle$

But, by the linearity of the inner product:

$\langle x-y, x-y \rangle = \langle x,x \rangle - \langle x,y \rangle - \langle y,x \rangle + \langle y,y \rangle$

$\therefore K_\sigma(x,y) = \exp\left(\frac{-\|x\|^2 - \|y\|^2 + 2\langle x,y \rangle}{2\sigma^2}\right)$

$K_\sigma(x,y) = \exp\left(\frac{-(\|x\|^2 + \|y\|^2)}{2\sigma^2}\right) \exp\left(\frac{\langle x,y \rangle}{\sigma^2}\right)$

Let $\exp\left(\frac{-\|x\|^2}{2\sigma^2}\right)$ be a function from $\mathbb{R}^n \to \mathbb{R} : g(x)$

$\therefore K_\sigma(x,y) = g(x) \exp\left(\frac{\langle x,y \rangle}{\sigma^2}\right) g(y)$

Using the Taylor series expansion for e^x

$$K_\sigma(x,y) = g(x)\left(\sum_{i=0}^{\infty}\left(\frac{\langle x,y\rangle}{\sigma^2}\right)^i \times \frac{1}{i!}\right) g(y)$$

$$K_\sigma(x,y) = \sum_{i=0}^{\infty} \frac{g(x)\langle x,y\rangle^i g(y)}{\sigma^2 i!} \quad ; \quad \frac{1}{\sigma^2 i!} > 0 \ \forall i \in \cancel{\mathbb{Z}}^+$$

We know that $g(x) = \exp\left(-\frac{\|x\|^2}{2\sigma^2}\right)$ is square integrable

Taking the integral over both sides. variables.

$$\iint_{y\,x} \sum_{i=0}^{\infty} \frac{g(x)(\langle x,y\rangle)^i g(y)}{\sigma^2 i!} = \frac{1}{\sigma^2}\sum_{i=0}^{\infty}\frac{1}{i!}\iint_{x\,y} g(x)(\langle x,y\rangle)^i g(y)$$

We know that each term in the summation is ≥ 0 as $(\langle x,y\rangle)^i$ is a valid kernel $\forall i$ ②.

$\therefore$ Let $h(x)$ be any square integrable f^n.

$\therefore$ $K_\sigma(x,y)$ is valid iff $\int_x\int_y h(x) K_\sigma(x,y) h(y)\, dy\, dx \geq 0$

We have: $\cancel{\iint_{x\,y}} \ \frac{1}{\sigma^2}\sum_{i=0}^{\infty}\frac{1}{i!}\iint_{x\,y} h(x)g(x)(\langle x,y\rangle)^i g(y)h(y)$

$h(x)g(x) = g'(x)$ is also square integrable, and can take the value of any f^n by changing $h(x)$.

$\therefore$ We have: $\frac{1}{\sigma^2}\sum_{i=0}^{\infty}\frac{1}{i!}\iint_{x\,y} g'(x)(\langle x,y\rangle)^i g'(y)$

As the Polynomial kernel is valid.

$$\therefore \frac{1}{\sigma^2} \sum_{i=0}^{\infty} \frac{1}{i!} \int_x \int_y g'(x) (\langle x,y \rangle)^i g'(y) \geq 0$$

(Linearity of kernels), $\forall g'(x)$

$$\therefore \int_x \int_y h(x) K_\sigma(x,y) h(y) \geq 0$$

$\forall h(x)$, $h(x)$ is square integrable.

$\therefore$ $K_\sigma(x,y)$ is a valid kernel hence proved.

Proof that $h(x), g(x)$ square integrable $\Rightarrow$ $h(x)g(x)$ sq. integrable.

$(h(x) - g(x))^2 \geq 0$

$\therefore (h(x))^2 + (g(x))^2 \geq 2h(x)g(x)$

$$\therefore h(x)g(x) \leq \frac{1}{2} h(x)^2 + \frac{1}{2} g(x)^2$$

$\because$ $h(x)^2$, $g(x)^2$ are square integrable, i.e, bounded and finite over some interval

By linearity of square integrable f^ns:

$$\iint_{x\,y} \frac{1}{2} h(x)^2 \frac{1}{2} h(y)^2 \, dx\,dy + \iint_{x\,y} \frac{1}{2} g(x)^2 \frac{1}{2} g(y^2) \, dx\,dy \leq k$$

$$\therefore \int_x (h(x)g(x))^2 \, dx \leq k$$

$\therefore$ $h(x)g(x)$ is also square integrable

Alternately, we know that $K(x,y) = g(x)g(y)$ is a valid kernel $g(x) : \mathbb{R}^n \to \mathbb{R}$.

as the Gram matrix K is the outer product of

$\vec{v} = [g(x_1), \dots g(x_n)]'$ $\therefore$ $vv^T = K$ (gram matrix)

$\therefore$ Outer product of a vector is a positive semidefinite matrix (rank 1). $\therefore$

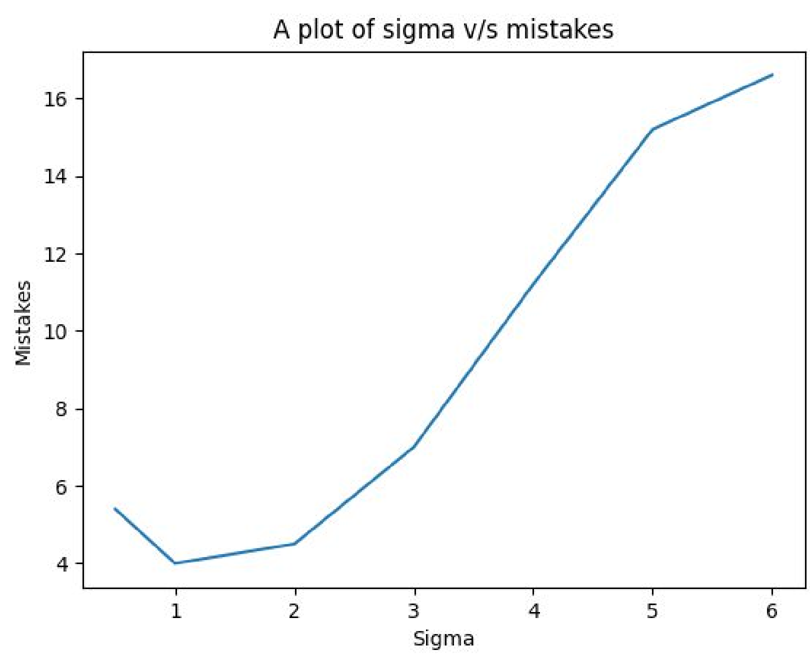
$$K = \sum_{i=0}^{\infty} g(x)g(y)(\langle x,y\rangle)^i$$ is a valid kernel

as $K(x,y) = K_1(x,y)\, K_2(x,y)$ is a valid kernel.

By linearity of kernels, the summation of kernels is also a kernel.

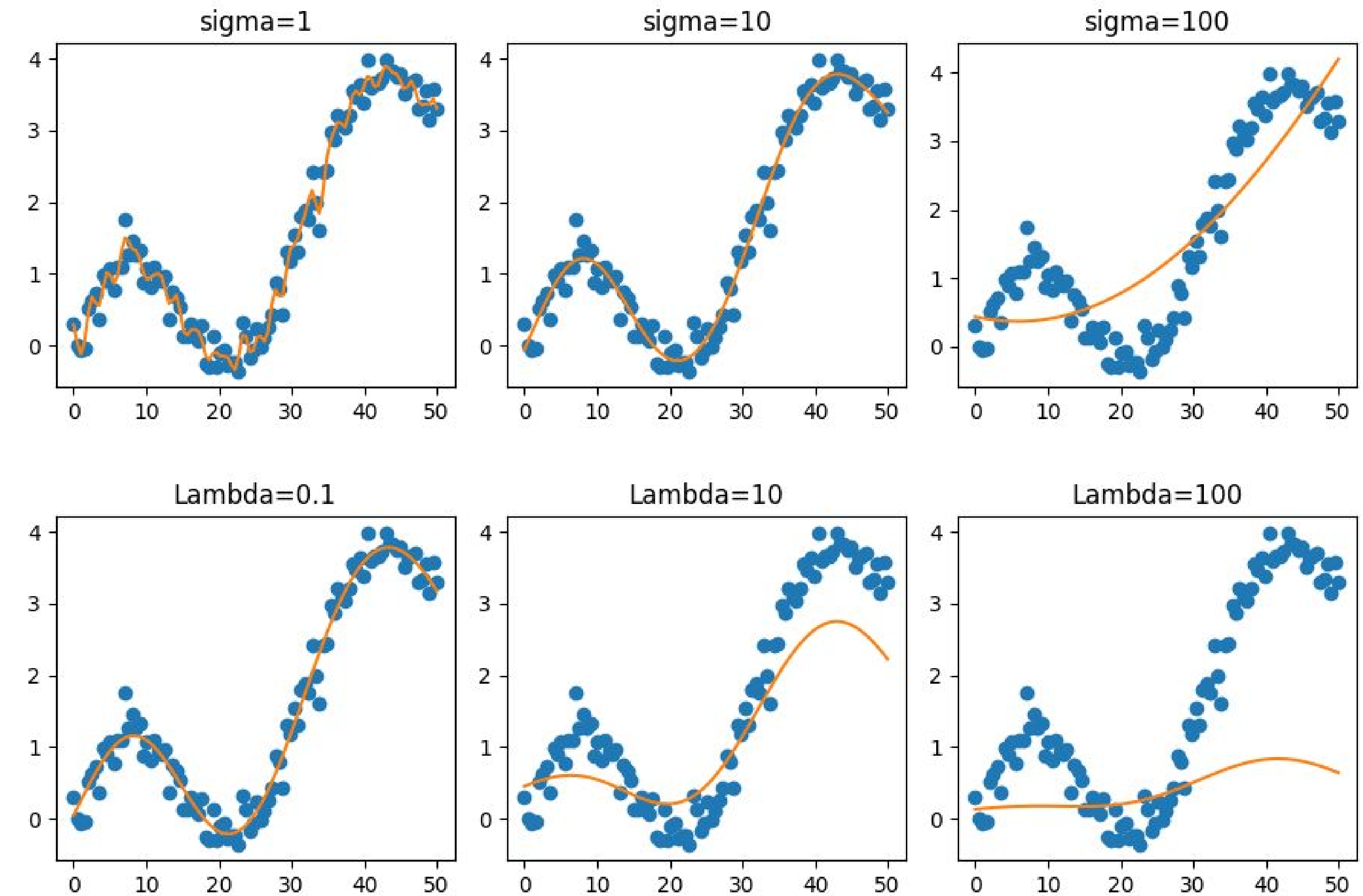
$$\therefore\ K = \frac{1}{\sigma^2}\sum_{i=0}^{\infty}\frac{1}{i!}\, g(x)g(y)\,(\langle x,y\rangle)^i$$ is a valid kernel.

1.2 b)



Sigma = 1 is the optimal sigma
This figure shows the trend between the number of mistakes and sigma. There is a minimum at sigma = 1.
With increasing sigma, the value of the evaluated gaussian kernel also increases. Hence, even when far from each other, two points will still have a high value as the output of K evaluated at them. This leads to a misrepresentation of similarity, and hence an increasing sigma will increase the errors of prediction. Effectively, we have blurred the boundaries between different classes by flattening out the gaussian/rbf kernel.

1.2 c)
The two figures shown below depict the variation of the fit of the curve with varying sigma and lambda (keeping the other constant)



Too high a value for either parameter ends up with the curve underfitting the data severely. Sigma = 10 and lambda = 0.1 seem to be the optimal values for ridge regression. Too high a sigma makes spread out data comparable (we lose the distinction between points), and too high a lambda means our regularisation penalty is high- a lot of weights will go to zero (ridge regression).
We also observe that sigma = 1 overfits the data, introducing unwanted artifacts. It mistakes noise for actual variation in the data.

Part 2:

2.1:

Date :

2.)

a) i) $K(x, x')$ is a valid kernel $\mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$

$g(x): \mathbb{R}^n \to \mathbb{R}^n$

To prove: $K(g(x), g(x'))$ is a valid kernel.

Say we have $K(x, x') = \phi^T(x)\phi(x')$

$\therefore K(g(x), g(x')) = \phi^T(g(x))\phi(g(x'))$

Take: $\phi(g(x))$

$\phi: \mathbb{R}^n \mapsto \mathbb{R}^k$ for some k.

$g: \mathbb{R}^n \mapsto \mathbb{R}^n$

$\therefore \phi(g(x)): \mathbb{R}^n \mapsto \mathbb{R}^n \mapsto \mathbb{R}^k$

$: \mathbb{R}^n \mapsto \mathbb{R}^k$

$\therefore$ This ~~g~~ composite function $\phi(g(x))$, ~~so~~ is in the same space as the original $\phi(x)$.

Say $\phi(g(x)) = \phi_g(x)$

$\therefore$ We have, in the same space

$K(g(x), g(x')) = \phi_g^T(x)\phi_g(x') = K'(x, x')$

$\therefore$ We have found a ϕ for which $K(g(x), g(x'))$ is a kernel function.

Hence $K(g(x), g(x'))$ is a valid kernel.

(It is simply a ~~f~~ different ~~one~~ kernel in the same feature space)

Date :

2.1

(i) Given: $K(x,x')$ is a valid kernel. q is a polynomial with non-negative coefficients.
(let its degree be 'n')

$$\therefore \quad q(K(x,x')) = a_0 + a_n (K(x,x'))^n + a_{n-1}(K(x,x'))^{n-1} + \ldots + a_1 K(x,x')$$

Lemma: The product of two kernels (valid)
$K(x_1, x_2') = K_1(x,x') K_2(x,x')$ is a valid kernel.

Say $\quad K_1(x,x') = \phi_1^T(x)\, \phi_1(x')$

$\quad K_2(x,x') = \phi_2^T(x)\, \phi_2(x')$

$$\therefore K_1(x,x') K_2(x,x') = \left(\sum_j \phi_{1j}(x_2)\, \phi_{1j}(x')\right)\left(\sum_j \phi_{2j}(x)\phi_{2j}(x')\right)$$

Converting this 'product of sum' to 'sum of product'

$$K_1 K_2 = \sum_i \sum_k \phi_{1i}(x)\phi_{2k}(x)\, \phi_{1i}(x')\phi_{2k}(x')$$

Let $\tilde{\Phi}(x) \doteq \phi_{1i}(x)\, \phi_{2k}(x)$

Consider the $\tilde{\Phi}(x) = \begin{bmatrix} \\ \phi_{1i}(x)\phi_{2k}(x) \\ \\ \end{bmatrix}_{i \times k}$ (a long vector)

This $\tilde{\Phi}(x)$ gives us $\quad K \doteq K_1 K_2 = \tilde{\Phi}^T(x)\, \Phi(x')$

Hence, we have found a representation for
$K = K_1(x,x') K_2(x,x')$ in terms of $\tilde{\Phi}$.

$\therefore \quad K(x,x')$ is a valid kernel

Date :

$\therefore$ If $K(x,x')$ is a valid kernel

$\therefore$ $(K(x,x'))^2$ is a valid kernel $(K'(x,x') = K(x,x')K(x,x')$

If $(K(x,x'))^{n-1}$ is a valid kernel, then

$K'(x,x') = (K(x,x'))^{n-1} K(x,x') = (K(x,x'))^n$

is a valid kernel (from previous proof)

$\therefore$ By mathematical induction $(K(x,x'))^n$ is a valid kernel $\forall n$.

$\therefore$ $q = C_n K_n + C_{n-1} K_{n-1} + C_{n-2} K_{n-2} + \dots$

$\therefore$ $q(K(x,x'))$ is a linear combination of valid kernels, with $C_i \geq 0$ $\forall i$.

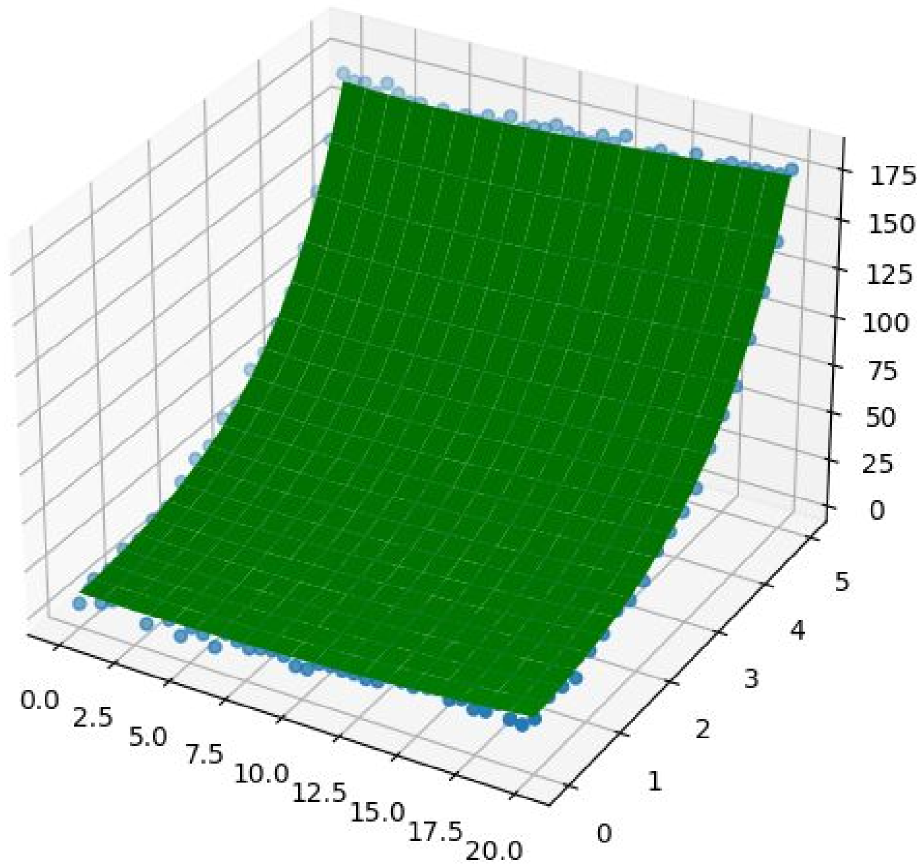
$\because$ K_1, K_2 valid $\Rightarrow$ $\alpha K_1 + \beta K_2$ valid $\forall \alpha, \beta \geq 0$

$\therefore$ $q(K(x,x'))$ is a valid kernel (The 'polynomial kernel').

Hence proved.

2.2:

The kernel I have selected as my custom kernel is a polynomial kernel. After visualising the data, one observes than when y is plotted against x, we see that the curve has 3 extrema. Also, it undulates like a polynomial, Hence, I tried a 4th degree polynomial (xTy)^4. However, this ‘passes through the origin’, which is an unneeded constraint our model. Therefore, to generalise, the kernel used is: (1+xTy)^4.



Part 3:

3.1:

3.1 Given: $\{x^1, x^2, \dots x^m\} \in$ cluster C_1

$\{x^{m+1}, \dots, x^n\} \in$ cluster C_2.

$x^i \in \mathbb{R}^d \quad d \geq 1$

$\therefore$ Consider cluster centres:

$\Rightarrow \quad c_1 = \frac{1}{m} \sum_{j=1}^{m} x^j \quad , \quad c_2 = \frac{1}{n+m-1} \sum_{j=m+1}^{n} x^j$

Consider the points c_1 and c_2 in space $\mathbb{R}^d$.

$\therefore$ If $x \in C_1$, then $\|x - c_1\|^2 < \|x - c_2\|^2$ — ①

$|||^{ly}$ If $x \in C_2$, then $\|x - c_2\|^2 < \|x - c_1\|^2$ — ②

Consider the hyperplane which is the perpendicular bisector of c_1 and c_2 ($\overline{c_1 c_2}$).

Let this hyperplane be H

$\therefore \quad \forall x \in H, \quad \|x - c_1\|^2 = \|x - c_2\|^2$

One form of the eq^n of H is:

$\|x - c_1\|^2 - \|x - c_2\|^2 = 0$

If $H(x) > 0$

$\Leftrightarrow \quad x \in C_2$ — ②

and If $H(x) < 0$

$\Leftrightarrow \quad x \in C_1$ — ①

$\therefore$ H is the required hyperplane, on either side of which lie the points belonging to

~~(Mid)~~ +

exactly one of the two clusters.

$\because$ H is the $\perp$ bisector of c_1 and c_2.

H passes through $\frac{\vec{c}_1+\vec{c}_2}{2}$. and H is normal

to $\vec{c}_1-\vec{c}_2$.

$\therefore \forall x \in H$: $\vec{x} - \left[\frac{\vec{c}_1+\vec{c}_2}{2}\right] \perp (\vec{c}_1-\vec{c}_2)$

$$\therefore \left[\vec{x} - \left(\frac{c_1+c_2}{2}\right)\right] \cdot (c_1-c_2) = 0$$

$$\therefore (\vec{c}_1-\vec{c}_2)\cdot\vec{x} - \left(\frac{\|c_1\|^2 - \|c_2\|^2}{2}\right) = 0$$

$$a\vec{x} + b = 0$$

$$\therefore a = \vec{c}_1-\vec{c}_2 \quad , \quad b = -\left(\frac{\|c_1\|^2 - \|c_2\|^2}{2}\right)$$

where $c_1 = \frac{1}{m}\sum_{j=1}^{m} x^j$

$$c_2 = \frac{1}{n-m} \sum_{j=m+1}^{n} x^j$$

Hence we have found the appropriate coefficients for the correct hyperplane H, which perfectly separates the optimal 2 cluster given

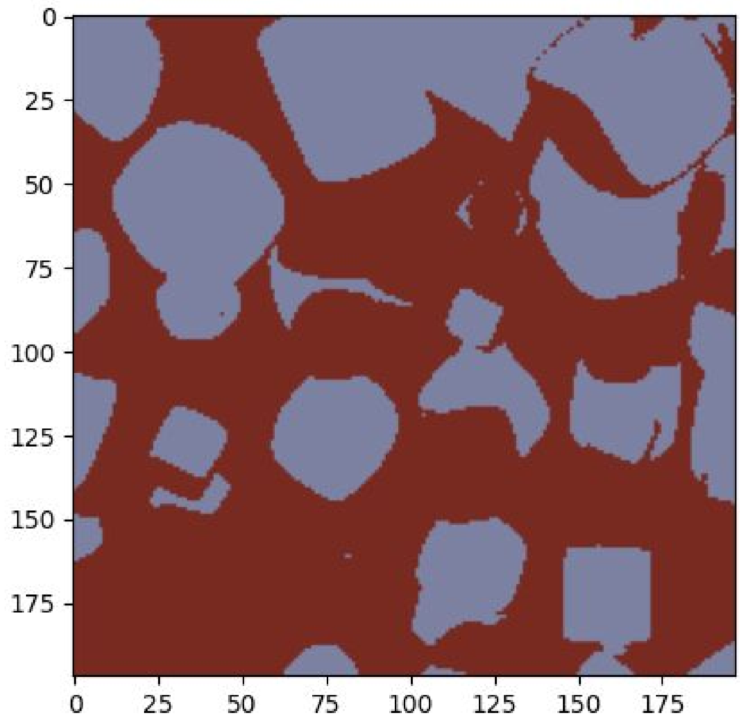
3.2:

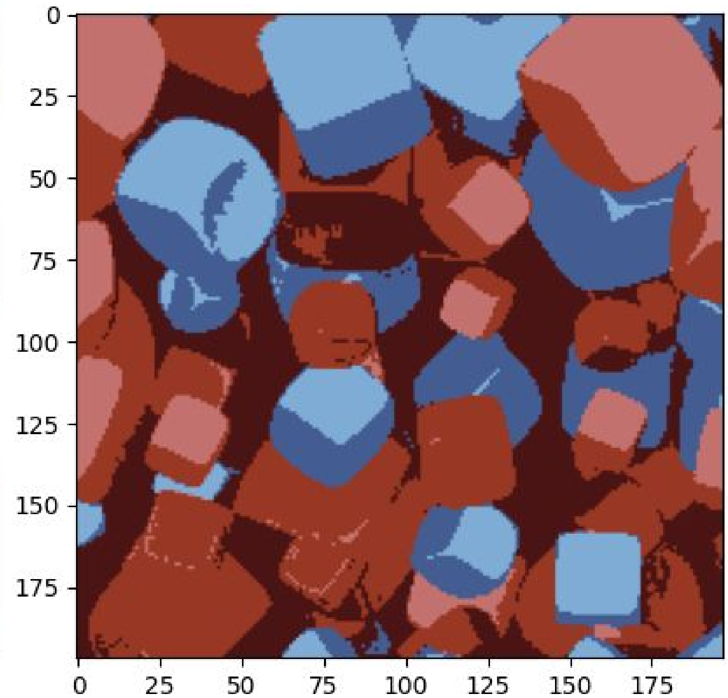
Here are images with k = 2,5,10

In each case, we see that the quality of the generated image increases with the number of clusters. This is because, with a higher k, we do not have to sacrifice details in the image like object edges or the complexity of an object.

Image 1:

Eg, here we can make out the cubes for k = 5, and we can distinguish between the different red cubes and blue cubes for k = 10. Due to the increased number of colours, there is also a perceived increase in sharpness.





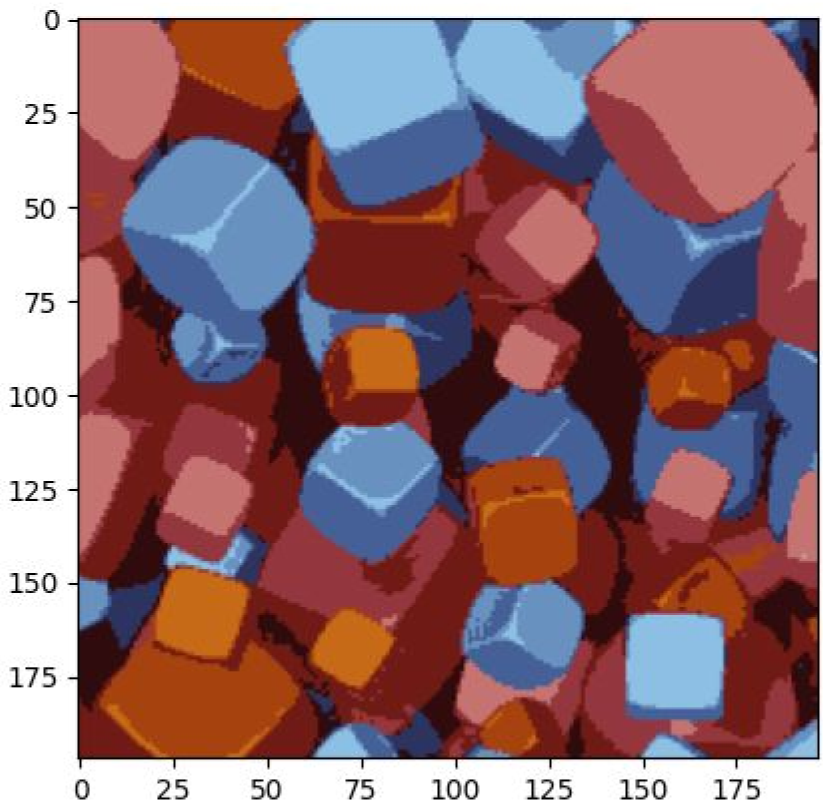
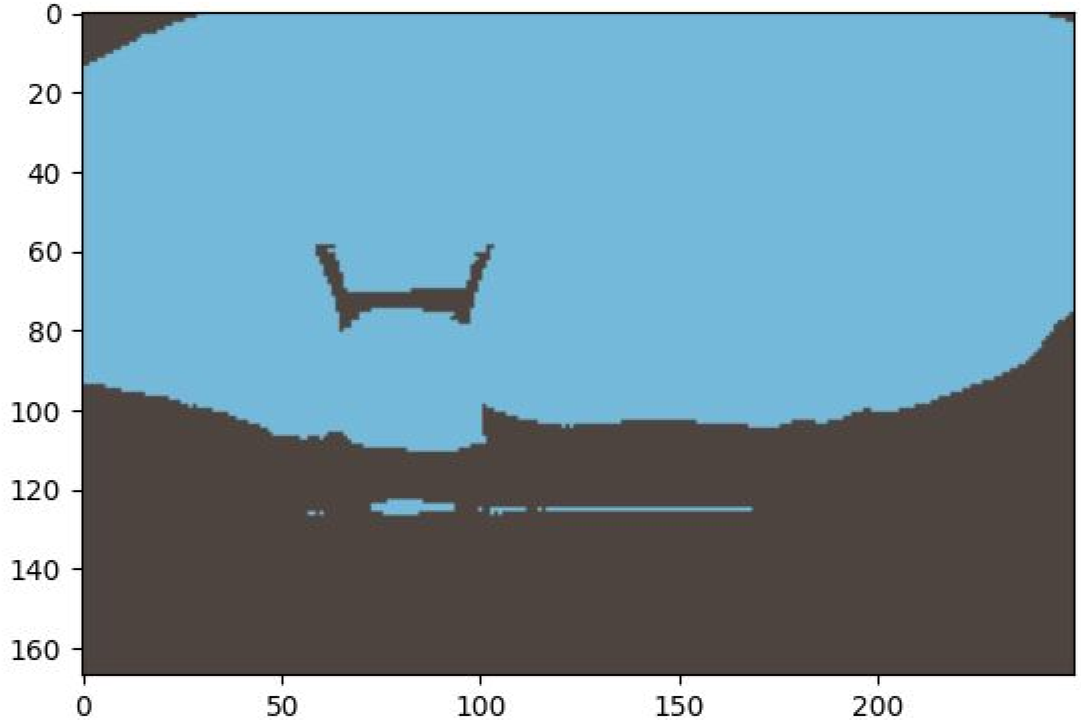
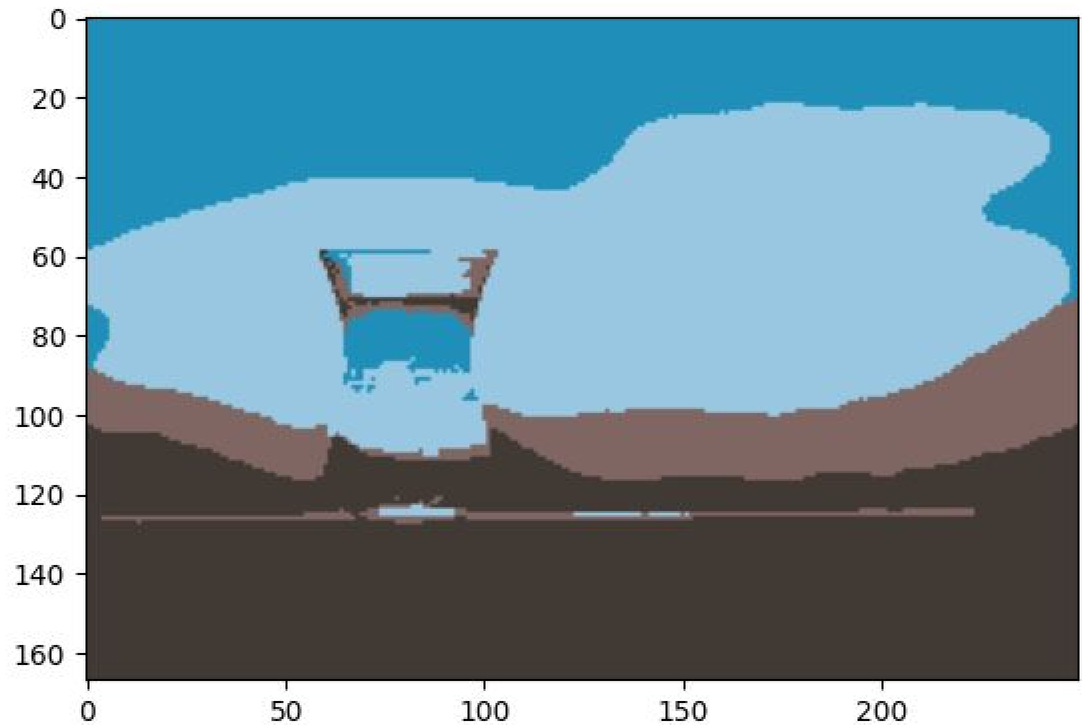


Image 2:
Here, k = 10 is required to see the image with its required details





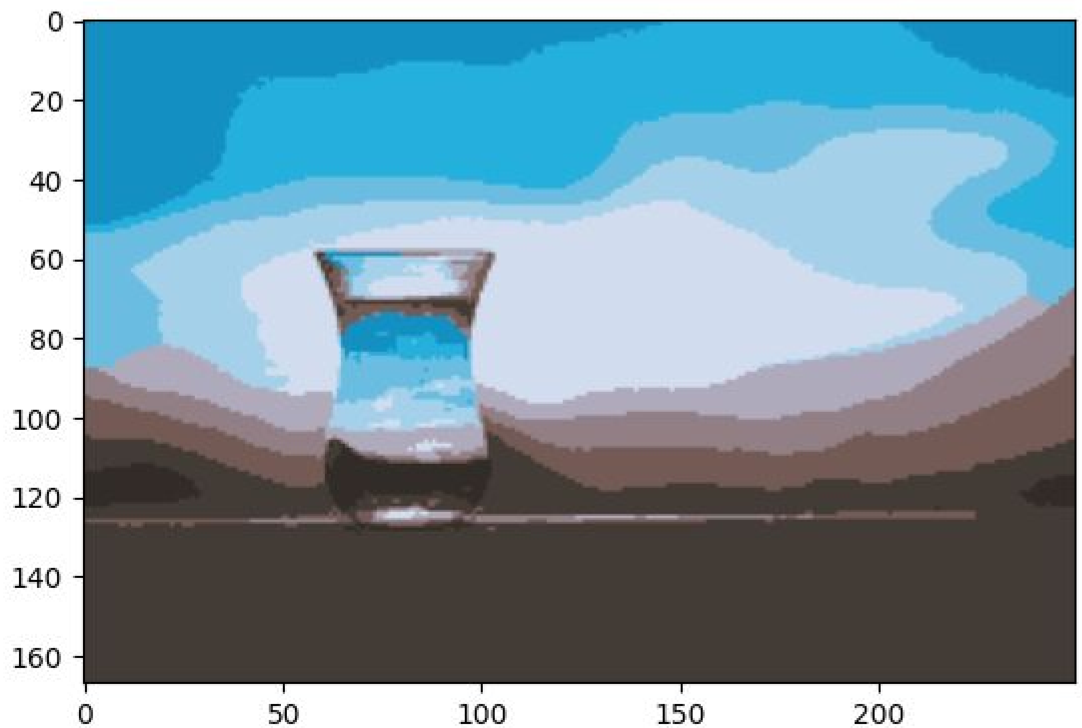
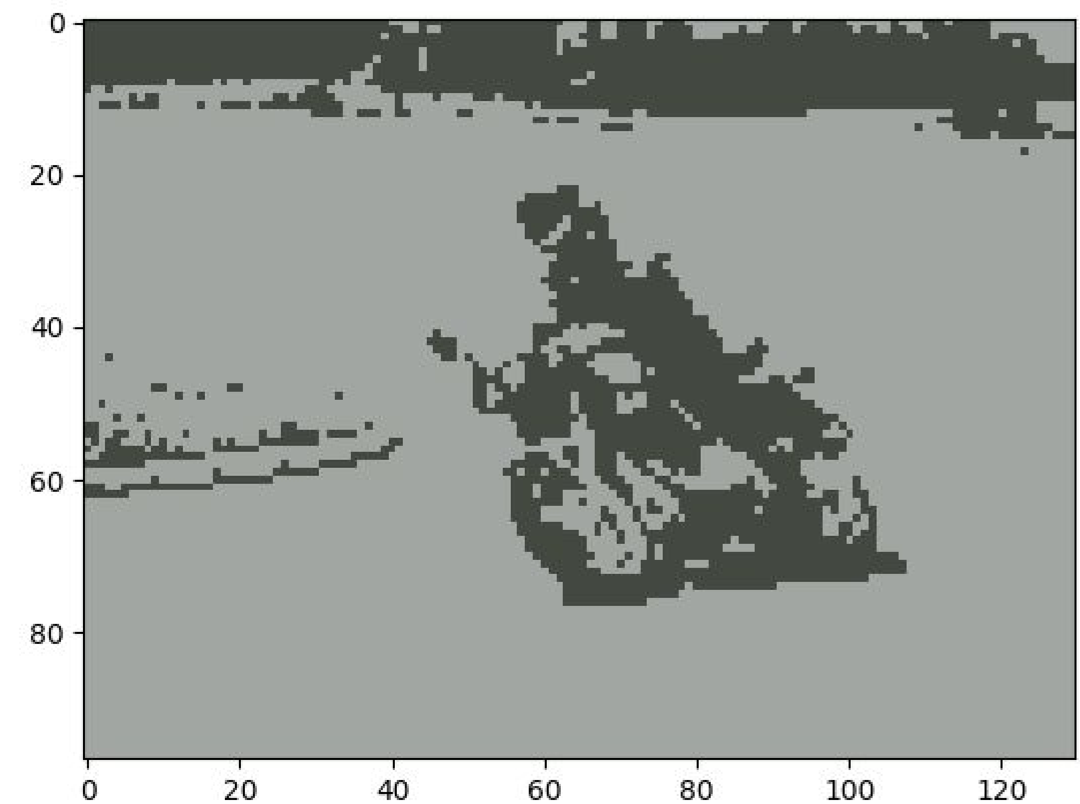
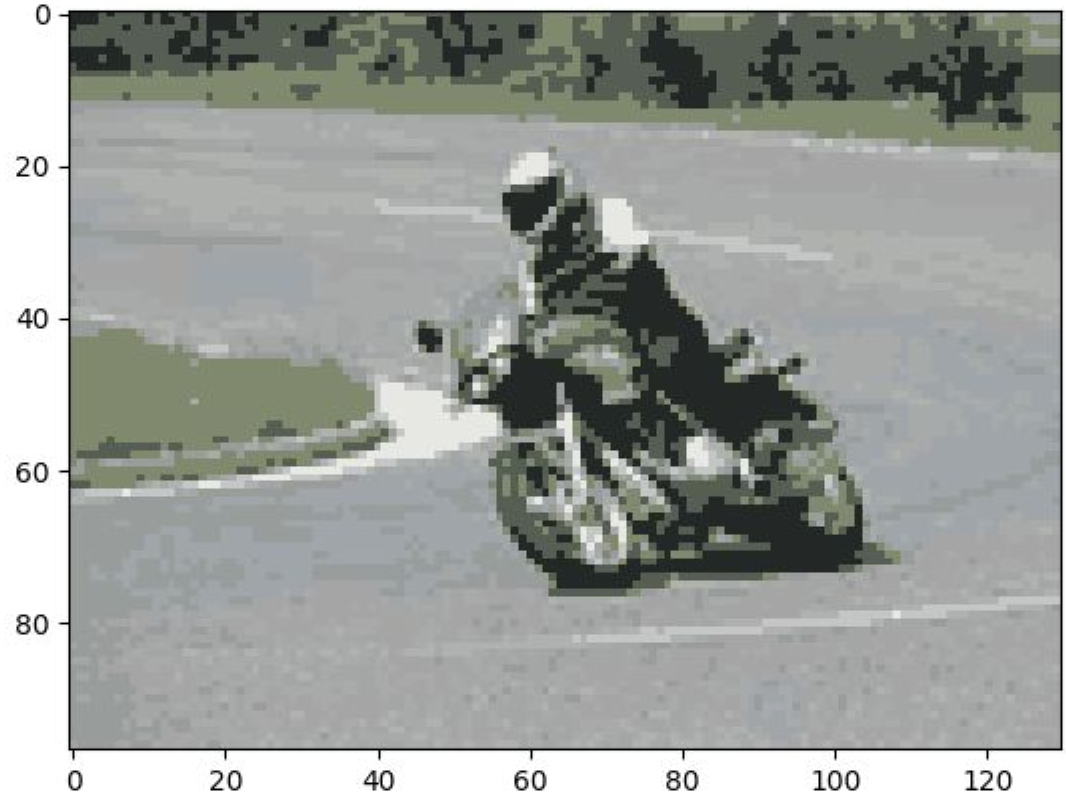


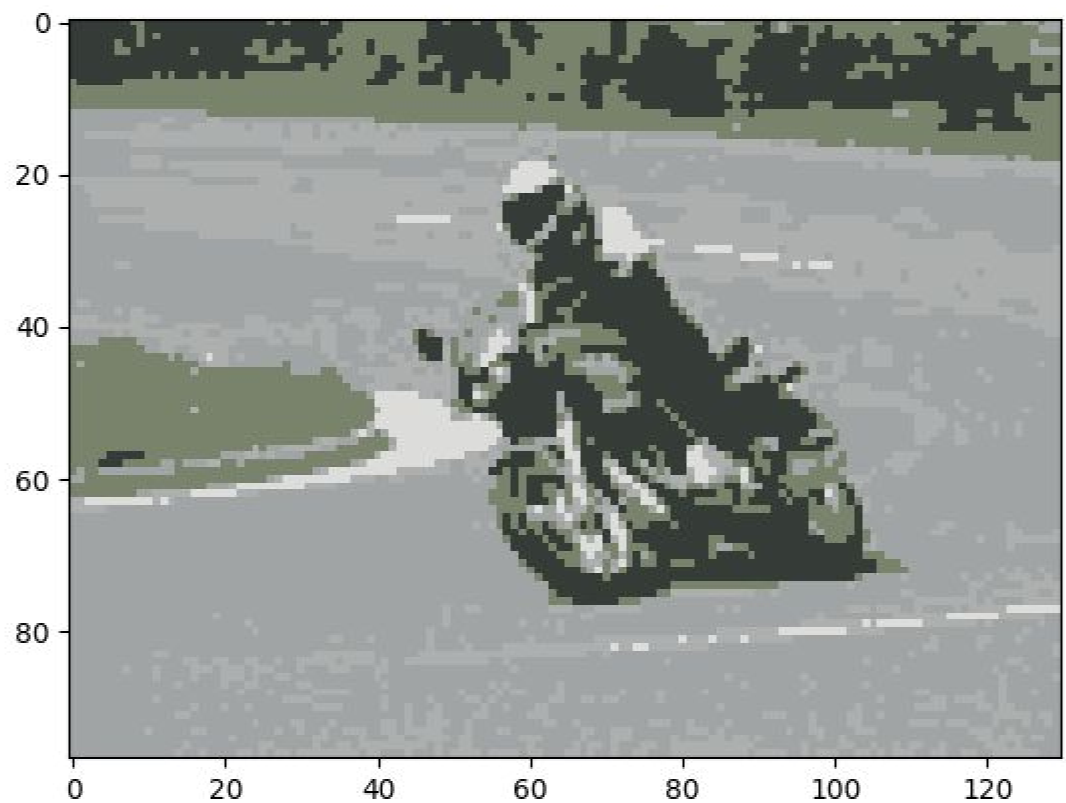
Image 3:
A stark difference between k=2 and k = 5, but not so much between k = 5 and k = 10
K = 2





k=5

k=10



Images 1 and 3 can make do with 5 clusters, while image 2 needs 10 to be able to be seen properly/replicate the original image to a reasonable degree. This is because image 2 has a transitioning gradient, and hence a wider range of colours. Clustering them into few clusters is not a good enough representation of the original image. It ruins subtle trends and differences in the image. Image 1 and 3 however, have more ‘blobs’ or patches of uniform colour. This makes clustering them, even with a few clusters, extremely natural and easy.