1.1

Theorem 1: If K is a kernel, rK is a kernel for $r>0$.

Proof: If ϕ is feature map for K, $\sqrt{r}\phi$ is feature map for rK.

Theorems from lectures:

2: If K_1, K_2 are kernels, K_1+K_2 is a kernel

3: If K_1, K_2 are kernels, K_1K_2 is a kernel

4: x^Ty is a kernel and $(x^Ty)^n$ is also a kernel

Theorem 5: If K is a kernel, then e^K is also a kernel

Proof: $e^K = 1 + \frac{K}{1!} + \frac{K^2}{2!} + \cdots$

It can be seen as a kernel using ③, ① and ②

From ④, ③ and ①, each term is a kernel because, K^n is obtained by multiplying K n times, which by ③ (or by ④ ⑦) makes K^n a kernel and by making r as $\frac{1}{n!}$ in ①, $\frac{K^n}{n!}$ is a kernel and by ②, summation of all such kernels is a kernel.

$\therefore e^K$ is a kernel

Theorem 7: If K is a kernel, K^n is a kernel

Proof: For $n=1$, $K^n = K$, which is a kernel

Let for $n=$ [redacted] $m-1$, K^n is a kernel

$K^m = K^1(K^{m-1})$

By ③, K^m is a kernel

$\therefore$ By induction, K^n is a kernel

Theorem 6: If $f: R^n \to R$, and K is a kernel, $f(x)K(x,y)f(y)$ is also a kernel

Proof: If ϕ is a feature map of K, $f(x)\phi(x)$ is the feature map for the new kernel

Now, consider $K(x,y) = e^{\left(\frac{-\|x-y\|^2}{2\sigma^2}\right)} = e^{\left(\frac{-\|x\|^2}{2\sigma^2}\right)} e^{\left(\frac{x^Ty}{\sigma^2}\right)} e^{\left(\frac{-\|y\|^2}{2\sigma^2}\right)}$

By ④ and ①, $\frac{1}{\sigma^2} x^Ty$ is a kernel by taking r as $\frac{1}{\sigma^2}$

By ⑤, $e^{\left(\frac{x^Ty}{\sigma^2}\right)}$ is a kernel

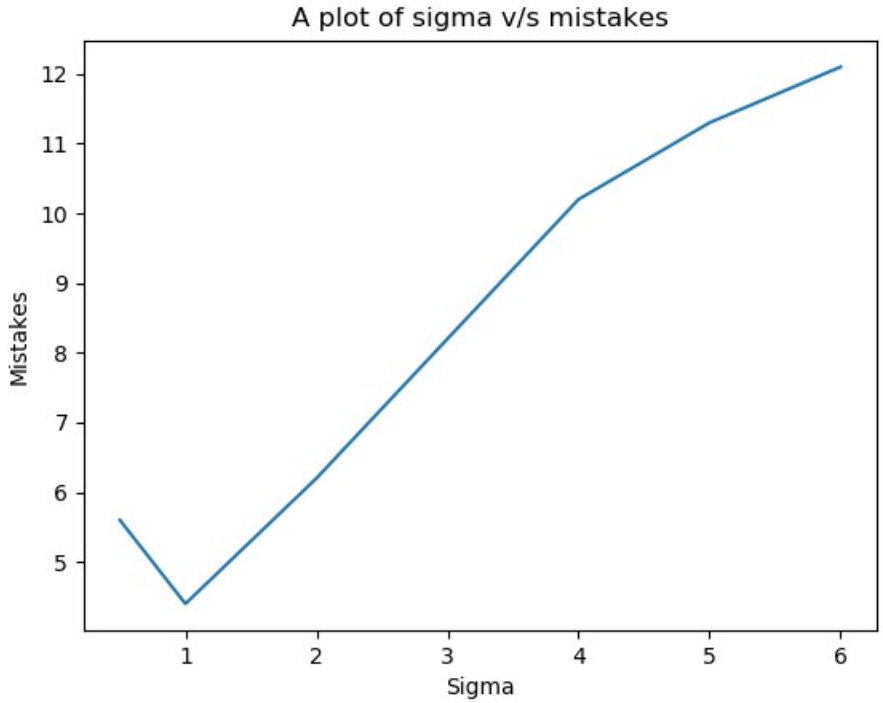
By ⑥, $e^{\left(\frac{-\|x\|^2}{2\sigma^2}\right)} e^{\left(\frac{x^Ty}{\sigma^2}\right)} e^{-\left(\frac{\|y\|^2}{2\sigma^2}\right)}$ is a kernel by taking

$f(x) = e^{\left(\frac{-\|x\|^2}{2\sigma^2}\right)}$

$\therefore K(x,y) = e^{\left(\frac{-\|x-y\|^2}{2\sigma^2}\right)}$ is a kernel

1.2

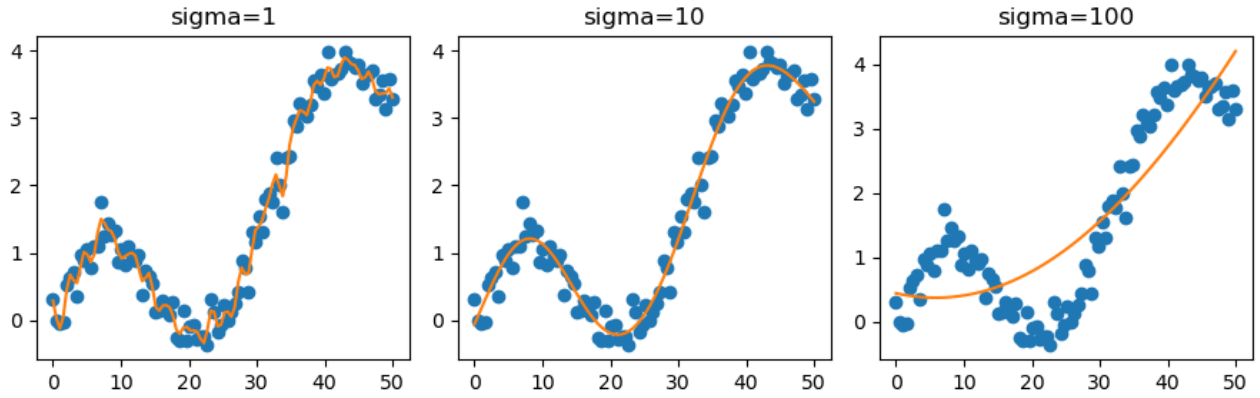
b) (ii)



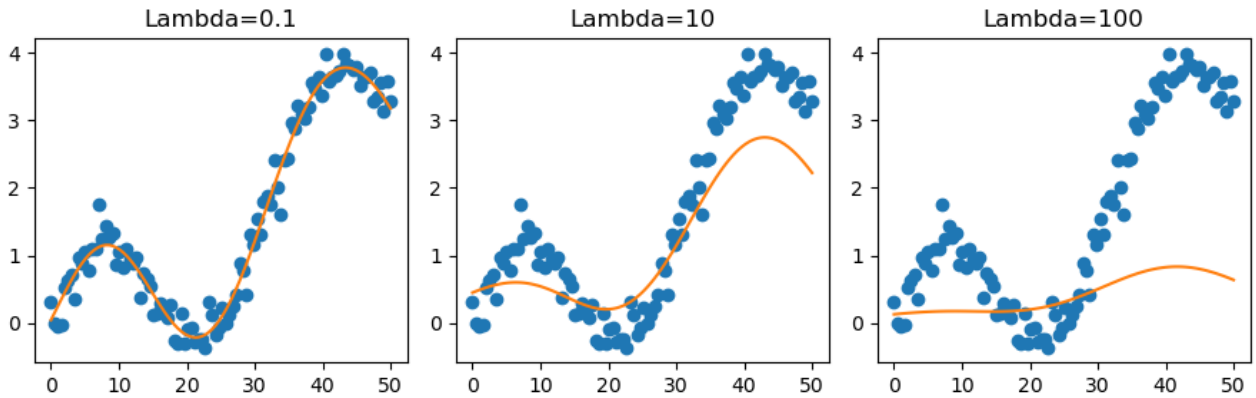
From the plot, the best value of sigma is 1.

(iii) The sigma in gaussian kernel controls the tolerance for similarity between two points. If sigma is low, the similarity between two points has to be very high or ||x-y||^2 must be low to make the kernel function value high. If sigma is high, even if similarity is less or ||x-y||^2 is high, the kernel function value becomes higher. So, as we increase sigma, we accomodate more and more points for a given point as being similar to it. So, until we reach the ideal value of sigma, number of mistakes reduces because initially we were overfitting, but once we cross the ideal value, the number of mistakes increase because underfitting occurs.

c) (ii)



The sigma in gaussian kernel controls the tolerance for similarity between two points. If sigma is low, the similarity between two points has to be very high or ||x-y||^2 must be low to make the kernel function value high. If sigma is high, even if similarity is less or ||x-y||^2 is high, the kernel function value becomes higher. So, as we increase sigma, we accomodate more and more points for a given point as being similar to it. As seen in the graph, until we reach the ideal value of sigma, we were overfitting because we considered only the closest points, but once we cross the ideal value, underfitting occurs because now we are also considering those points also which are not properly related.

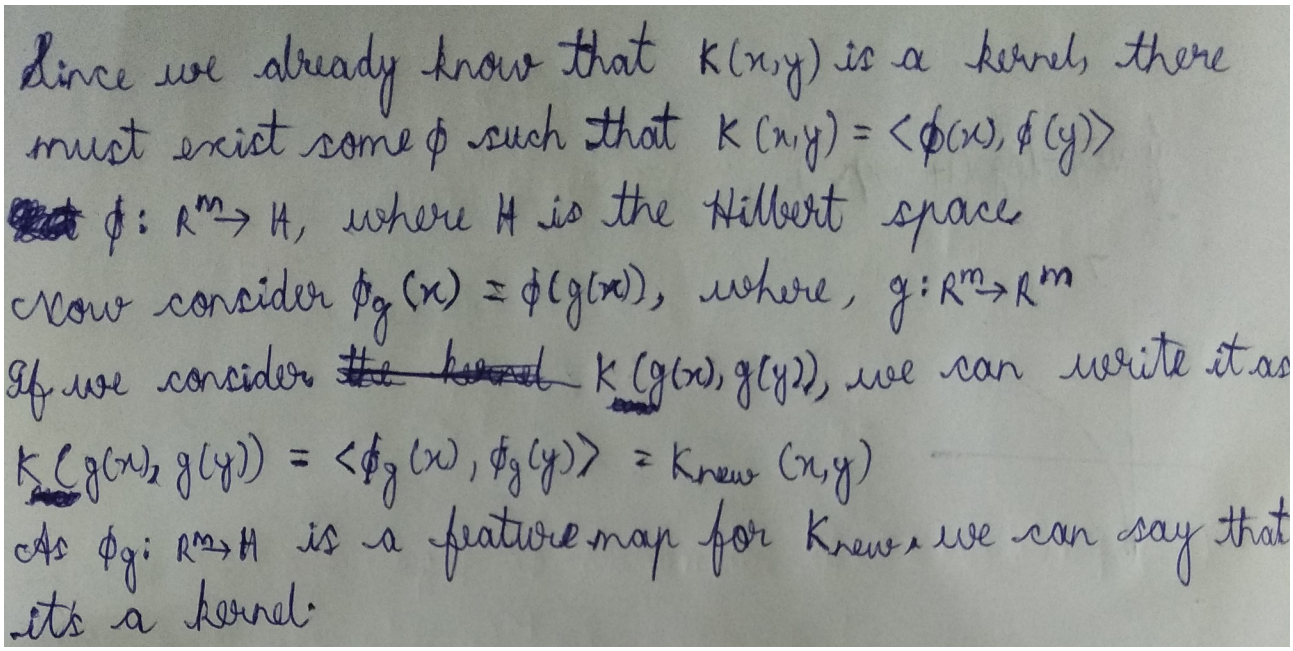


Since lambda is the parameter for imposing penalty on weights to avoid overfitting, as lambda increases, the penalty imposed increases which makes the values of weights smaller and closer to

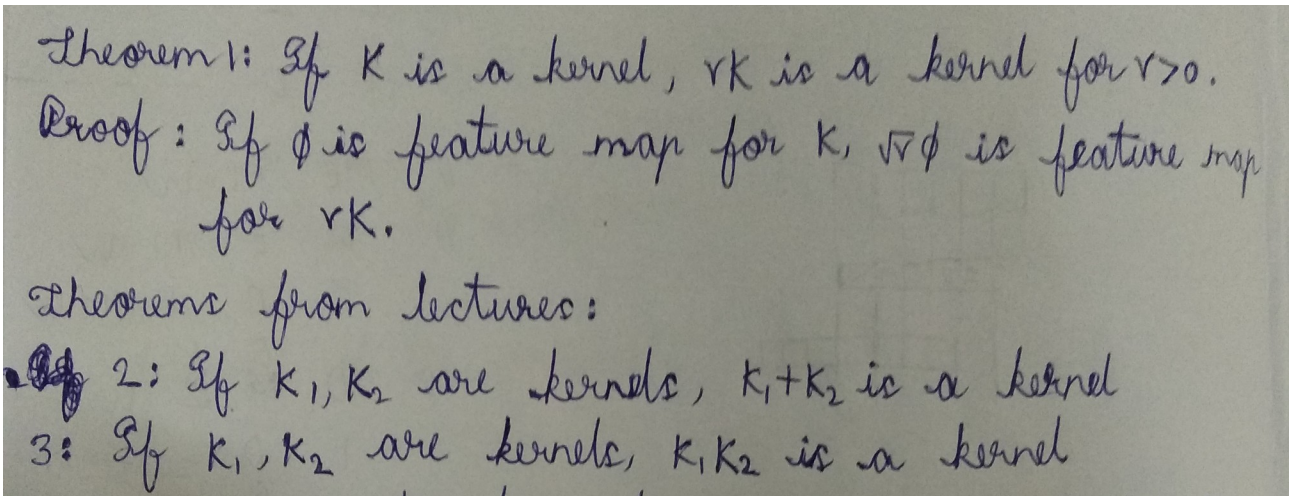
zero. As seen in the graph, as lambda increases after the optimal value, we tend to underfit the data. Also, decreasing lambda beyond the optimal value may cause overfitting due to large weights.

2.1

(i)



(ii)



Theorem 7: If K is a kernel, K^n is a kernel

Proof: For n=1, $K^n = K$, which is a kernel

Let for n=[redacted] m-1, K^n is a kernel

$K^m = K^1(K^{m-1})$

By ③, K^m is a kernel

∴ By induction, K^n is a kernel

Theorem 8: If q is a polynomial with non-negative coefficients, q(K) is a kernel given K is a kernel.

Proof: Every term of q(K) is of the form αK^n where, $\alpha > 0$.

By ⑦ and ⑤, αK^n is a kernel.

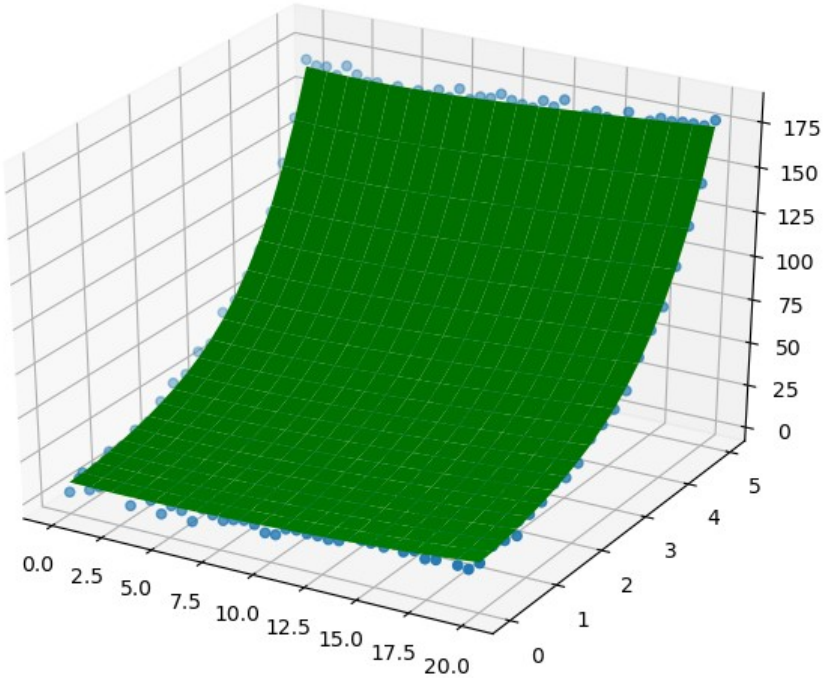
$q(K) = \sum \alpha K^n$

As each term is a kernel, by applying ① repetitively, the whole summation is a kernel.

As the whole summation is q(K), q(K) is a kernel.

∴ q(K) is a kernel

2.2 The kernel function used is (<x,y>+1)^4.



3.1

Let $\bar{x}_1$ be the first cluster mean (cluster centre)

$$\bar{x}_1 = \frac{x_1 + x_2 + \cdots + x_m}{m}$$

Let $\bar{x}_2$ be the second cluster mean (cluster centre)

$$\bar{x}_2 = \frac{x_{m+1} + x_{m+2} + \cdots + x_n}{n-m}$$

Let the separating hyperplane be $a \cdot x + b = 0$

Claim: $a = \bar{x}_1 - \bar{x}_2$

$$b = \frac{\|\bar{x}_2\|^2 - \|\bar{x}_1\|^2}{2}$$

~~and for $x_1, \ldots, x_m$, $a \cdot x$~~ but $a \cdot x + b \leq 0$

Proof: Let x belong to the first cluster ~~but~~ $a \cdot x + b \leq 0$

As x belongs to first cluster,

$\|x - \bar{x}_1\|^2 < \|x - \bar{x}_2\|^2$ (∵ no point is equidistant from two centres)

$$(x - \bar{x}_1)^T (x - \bar{x}_1) < (x - \bar{x}_2)^T (x - \bar{x}_2)$$

$$\|x\|^2 - 2\bar{x}_1^T x + \|\bar{x}_1\|^2 < \|x\|^2 - 2\bar{x}_2^T x + \|\bar{x}_2\|^2$$

$$2(\bar{x}_1^T - \bar{x}_2^T)x > \|\bar{x}_1\|^2 - \|\bar{x}_2\|^2$$

$$2(\bar{x}_1^T - \bar{x}_2^T)^T x > \|\bar{x}_1\|^2 - \|\bar{x}_2\|^2 \quad -①$$

$$a \cdot x + b \leq 0$$

$$(\bar{x}_1 - \bar{x}_2) \cdot x + \left(\frac{\|\bar{x}_2\|^2 - \|\bar{x}_1\|^2}{2}\right) \leq 0$$

$$(\bar{x}_1 - \bar{x}_2)^T x \leq \frac{\|\bar{x}_1\|^2 - \|\bar{x}_2\|^2}{2}$$

$$2(\bar{x}_1 - \bar{x}_2)^T x \leq \|\bar{x}_1\|^2 - \|\bar{x}_2\|^2 \quad -②$$

But ① and ② are contradictions of each other

∴ By proof by contradiction, if $x \in \{x_1, \ldots, x_m\}$, $a \cdot x + b > 0$ and similarly, if $x \in \{x_{m+1}, \ldots, x_n\}$, $a \cdot x + b < 0$

∴ The two clusters are separated by the hyperplane $a \cdot x + b = 0$, where,

$$a = \frac{\sum_{i=1}^{m} x_i}{m} - \frac{\sum_{j=m+1}^{n} x_j}{n-m}$$

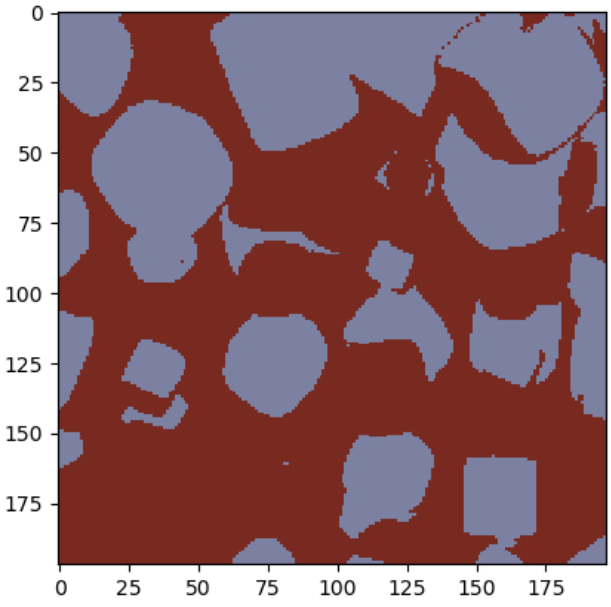
$$b = \frac{\left\| \frac{\sum_{j=m+1}^{n} x_j}{n-m} \right\|^2 - \left\| \frac{\sum_{i=1}^{m} x_i}{m} \right\|^2}{2}$$

3.2

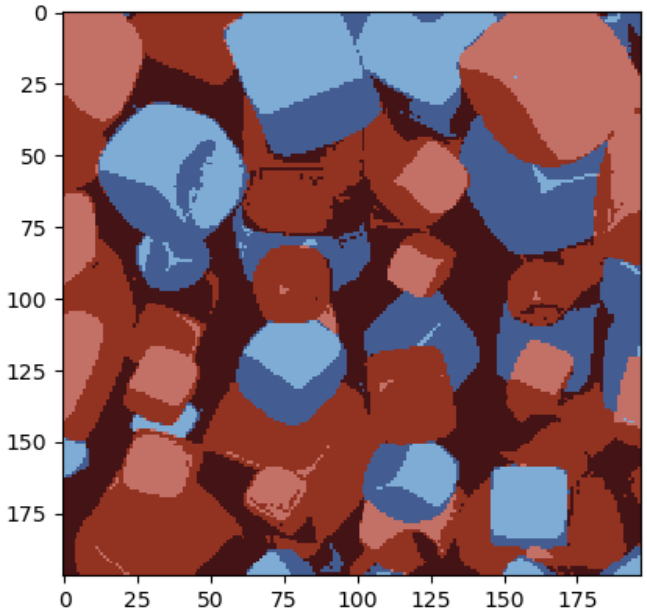
The images are shown in the order of increasing number of clusters 2,5,10.

(ii)

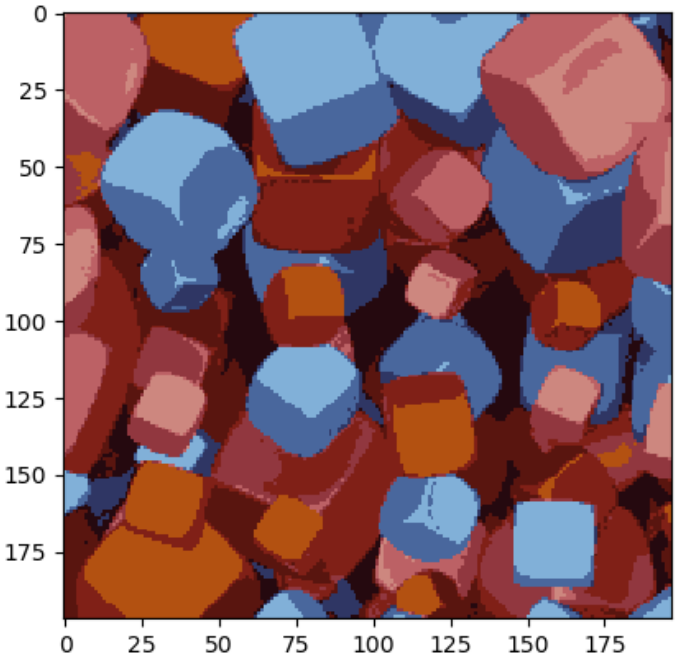
I.



For number of clusters = 2, the generated image is not even close to the actual image, since the number of clusters is low compared to the actual number of clusters in the image.

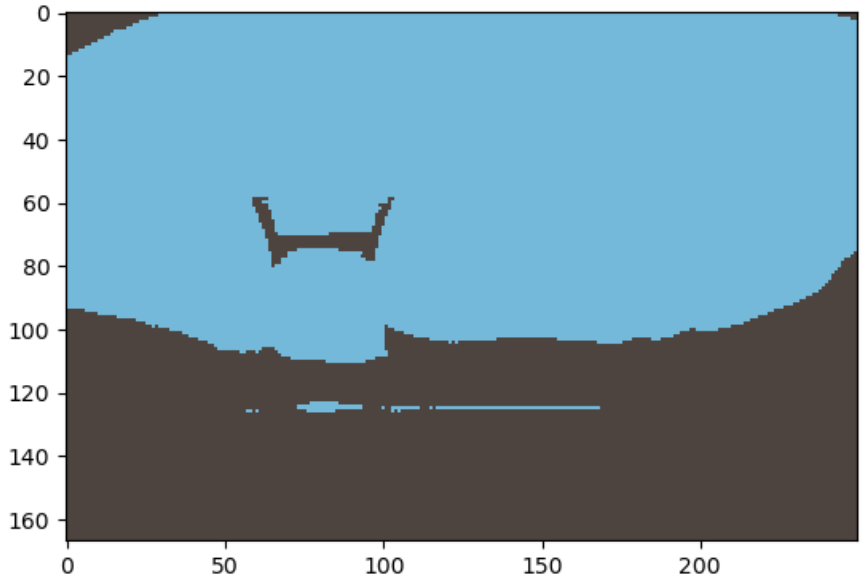


For number of clusters = 5, the generated image is somewhat close to the actual image, since the number of clusters is comparable to the actual number of clusters in the image. Here the structure of the image is captured but colours are not captured properly.

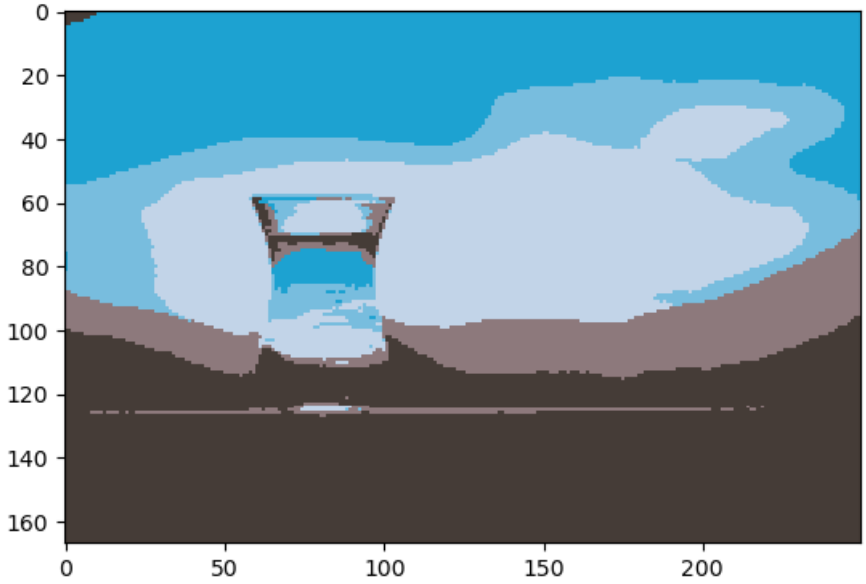


For number of clusters = 10, the generated image is very close to the actual image, since the number of clusters is almost equal to the actual number of clusters in the image.

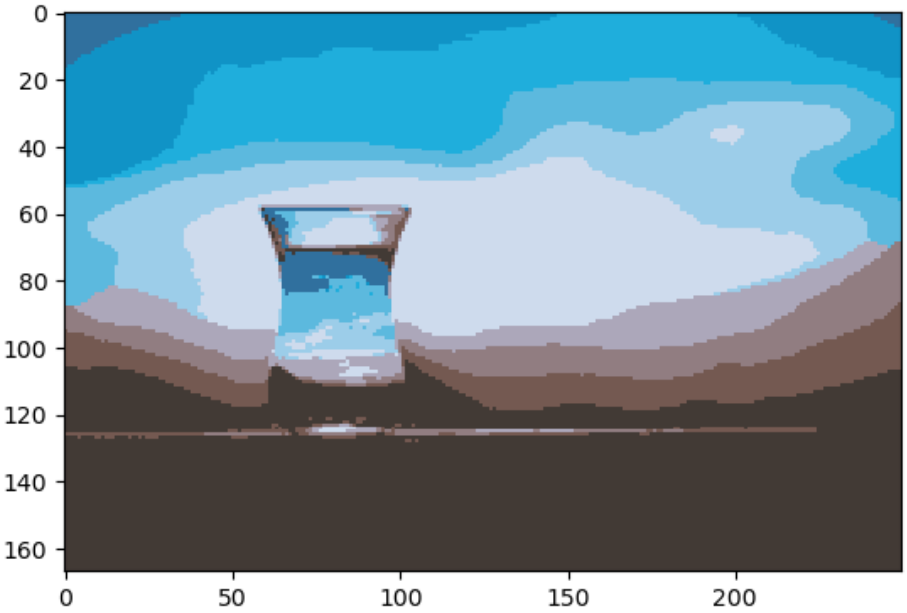
II.



For number of clusters = 2, the generated image is not even close to the actual image, since the number of clusters is low compared to the actual number of clusters in the image.

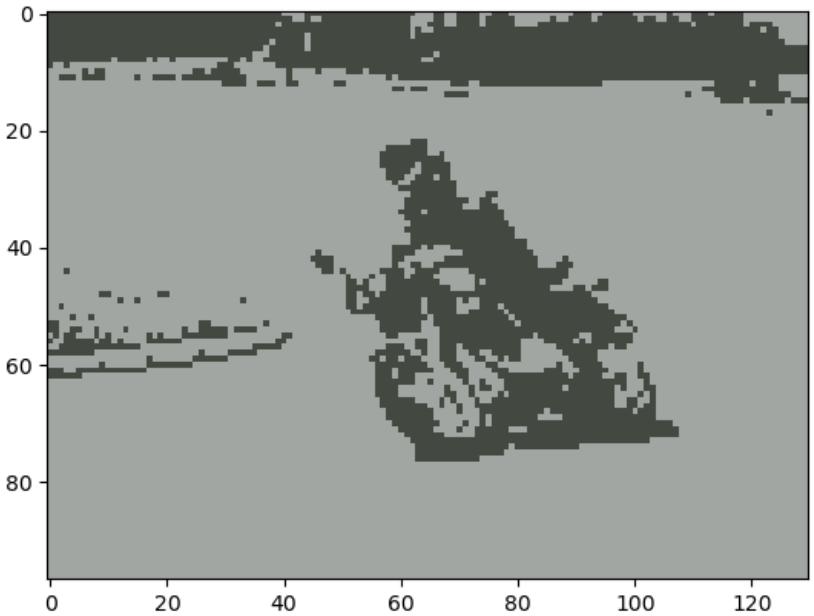


For number of clusters = 5, the generated image is very close to the actual image, since the number of clusters is almost equal to the actual number of clusters in the image.

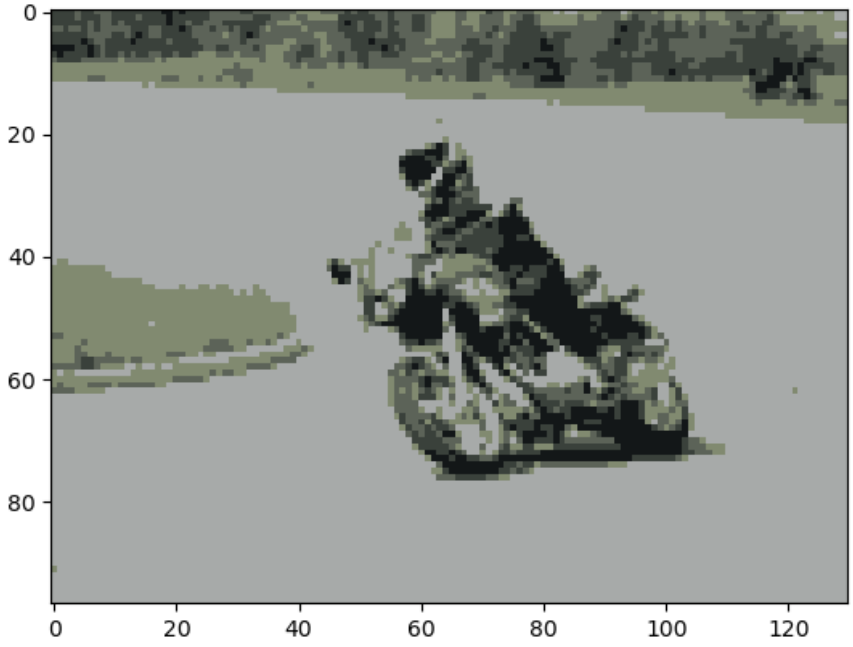


For number of clusters = 10, the generated image adds more detailing to the image with number of clusters as 5, since it has more clusters than the actual number of clusters needed for the image.

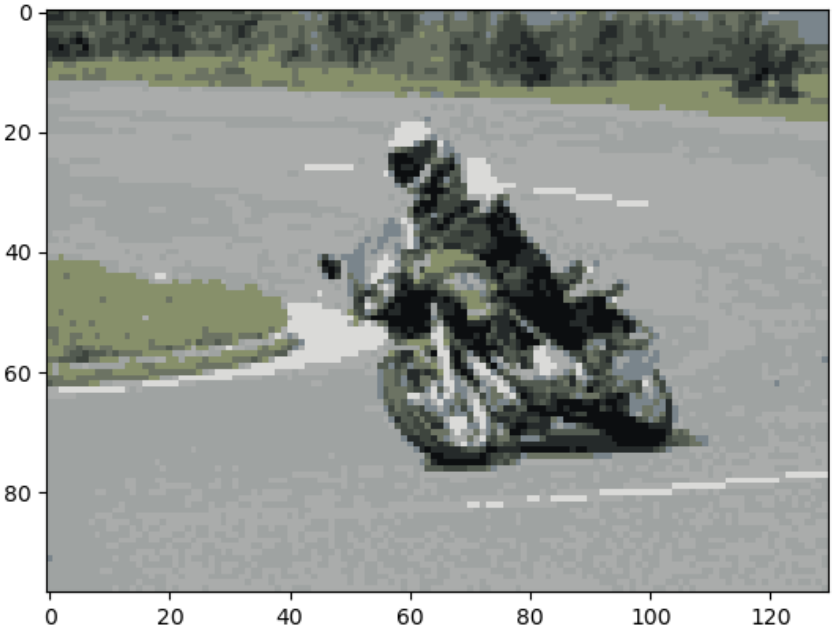
III.



For number of clusters = 2, the generated image is not even close to the actual image, since the number of clusters is low compared to the actual number of clusters in the image.



For number of clusters = 5, the generated image is somewhat close to the actual image, since the number of clusters is less compared to the actual number of clusters in the image.



For number of clusters = 10, the generated image is close but not very close to the actual image, since the number of clusters is less than the actual number of clusters in the image. Here the structure of the image is captured but colours are not captured properly.

(iii) The number of clusters needed to represent the image is almost represented by the actual number of colours and detail in the image(that is by the number of unique (R,G,B) pixels). In the first image, we have less number of colours but more detail, so we required 10 clusters to clearly represent the image. The second image has less colours and less detailing, so only 5 clusters were enough. In the last image, we have more colours and more detailing, so, even 10 clusters are not sufficient. Note that detailing also demands more clusters because, with increasing clusters, points become more closer to the cluster centers and with many clusters, many groups of points with different colours whose cluster center very closely resembles them can be formed.