1.1

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

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

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

$K_0(x,y) = x^Ty$ — Pd kernel

we know that if $K(x,y)$ is a Pd kernel so is $\gamma K(x,y)$ if $\gamma > 0$

so scaling $K_0(x,y)$ by $\frac{1}{\sigma^2}$ we get

$$K_1(x,y) = \frac{1}{\sigma^2}K_0(x,y) = \frac{x^Ty}{\sigma^2}$$

also note that if K is a Pd kernel so $\exp(K(x,y))$ as

$$\exp(K(x,y)) = \lim_{N\to\infty}\sum_{i=0}^{N}\frac{1}{i!}(K(x,y))^n$$

and $(K(x,y))^n$ is valid kernel because of Products property that $K(x,y) = K_1(x,y)K_2(x,y)$ is Pd kernel if K_1 and K_2 are Pd kernels

So! we Now get $K_2(x,y) = \exp(K_1(x,y))$

$$K_2(x,y) = \exp\left(\frac{x^Ty}{\sigma^2}\right)$$

similar steps for exp

also note that

if K_1 is a Pd kernel

then for some $f : X \to R$

$K_0(x, y) = f(x)\, K(x, y) f(y)$

is also a valid kernel

as we can use a feature map

$\phi'(x) \longmapsto f(x)\, \phi(x)$

finally

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

$K(x, y) = f(x)\, K_2(x, y)\, f(y)$

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

and $K_2(x, y) = \exp\left(\frac{x^T y}{\sigma^2}\right)$

hence $K(x, y)$ is a valid kernel

1.2(b) (ii) Number of mistakes is minimised for $\sigma = 1$

1.2 (b) (iii) ~~The nature of~~ Graph depicts an decreasing that increasing trend which is justified as follows

→ for smaller sigma the model is overfitting the data, so higher test error than $\sigma = 1$

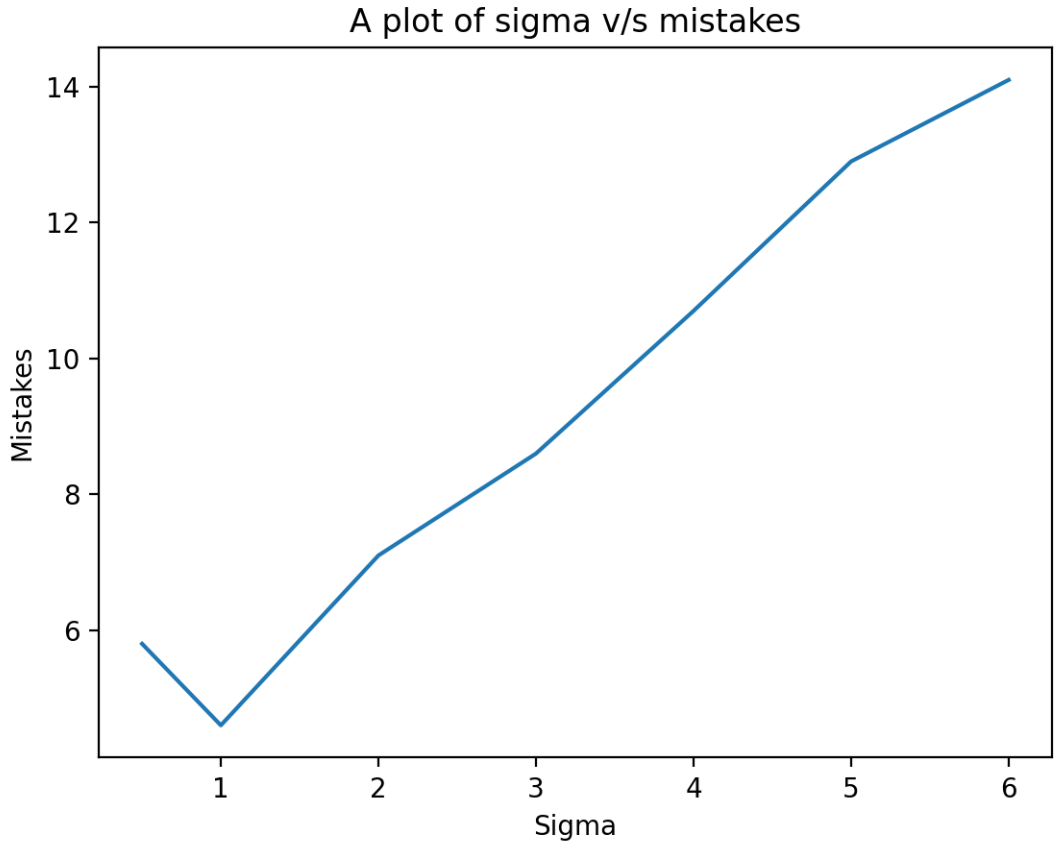
→ for larger sigma the model is underfitting the data, so higher test error than $\sigma = 1$

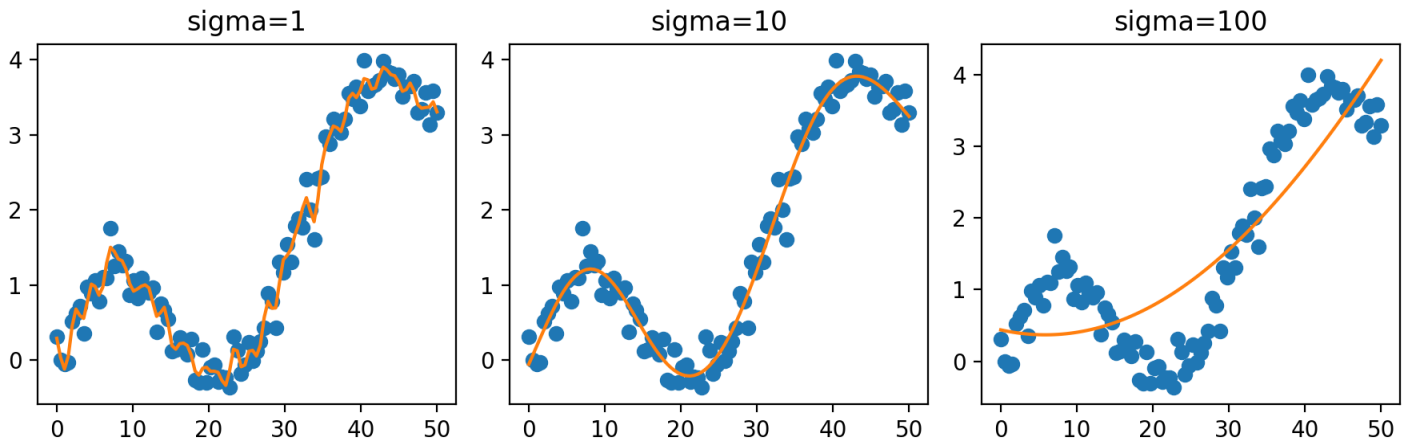
1.2 (c) (ii) for sigma variation

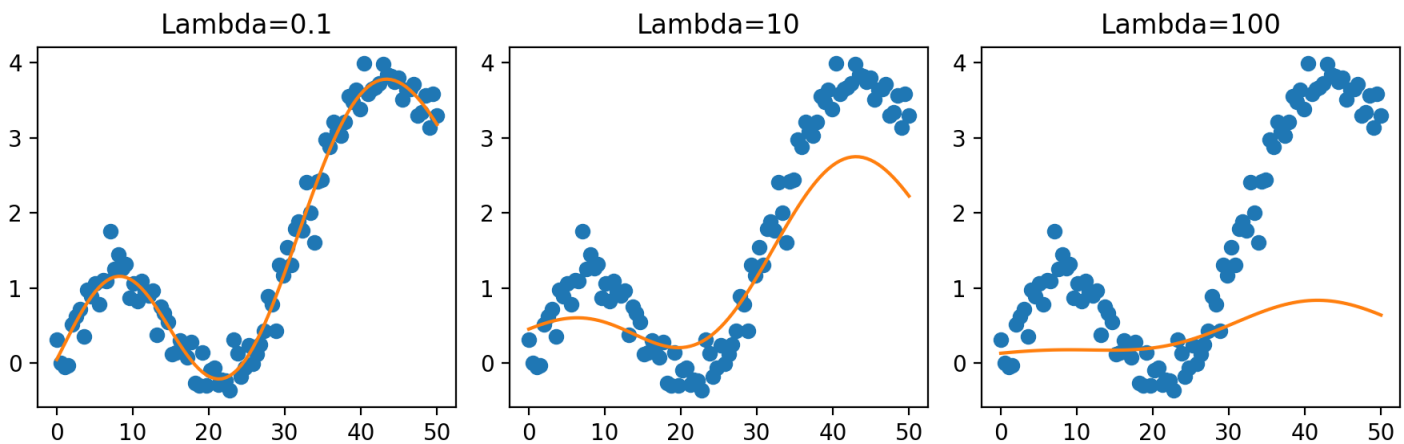
→ small sigma overfitting
→ large sigma underfitting
→ sigma = 10 good fit

for lambda variation

→ As lambda increases model is underfitting


A plot of sigma v/s mistakes
Mistakes
14
12
10
8
6
1
2
3
4
5
6
Sigma


sigma=1
sigma=10
sigma=100
0
1
2
3
4
0
10
20
30
40
50


Lambda=0.1
Lambda=10
Lambda=100
0
1
2
3
4
0
10
20
30
40
50

Sajal
Date : / /20
Page :

2.1 (i) $K(x_1, x_2)$ is a Pd Kernel

so ϕ exists such that $\phi : R^n \to H$ such that $K(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle$

and we know that $g(x) : R^n \to R^N$

so consider $\phi_g : R^n \to H$ such that

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

then $K_{new}(x_1, x_2) = K(g(x_1), g(x_2))$

$K_{new}(x_1, x_2) = \langle \phi_g(x_1), \phi_g(x_2) \rangle$

$K_{new}(x_1, x_2)$ is a Pd kernel

2.1 (ii) Let $q(x) = \sum_i a_i x^i$

where $a_i > 0$ for all i

Now $K(x_1, x_2)$ is a Pd Kernel

~~it is sufficient to show that~~

Now $q(K(x_1, x_2)) = \sum_i a_i (K(x_1, x_2))^i$

it is a known property that linear combination of valid kernel is also valid kernel so it is sufficient to show that $\gamma (K(x_1, x_2))^n$ is a valid kernel where $\gamma > 0$ and $n \in \mathbb{Z}^+$

Now we know that product of two Pd kernel is also a Pd kernel so generalising this property of Pd kernels we can say that

$$K_{new}(x_1, x_2) = K_1(x_1, x_2) K_2(x_1, x_2) \dots K_n(x_1, x_2)$$

Knew is a Pd kernel if K_i is Pd for all $i \in [1, N]$

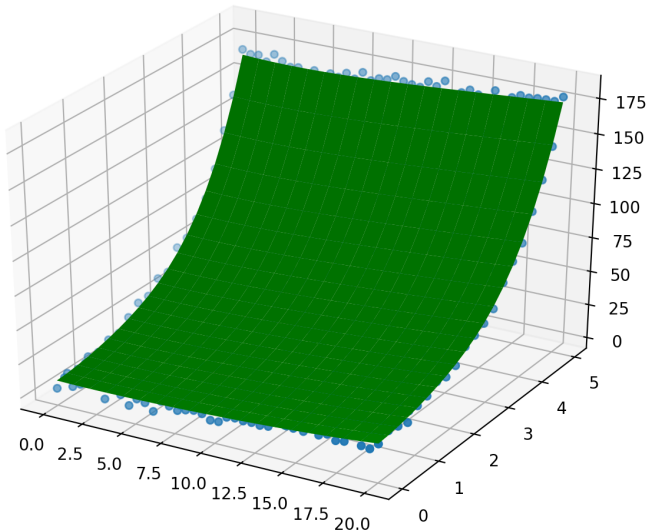
hence $(K(x_1, x_2))^n$ is a valid kernel from above stated property if $K(x_1, x_2)$ is valid Pd kernel

and using the other property that $K(x, y) = \gamma K_1(x, y)$ K is a Pd kernel if K is a Pd kernel and $\gamma > 0$

So hence $\gamma (K(x_1, x_2))^n$ is a Pd kernel for $\gamma > 0$ and $n \in \mathbb{Z}^+$

so $q(K(x_1, x_2))$ is a valid Pd kernel

2.2 my-kernel$(x, y) = (1 + x^T y)^4$


175
150
125
100
75
50
25
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
0
1
2
3
4
5

3.7

Centroid of first cluster, $\alpha_1 = \sum_{i=1}^{m} x_i / m$

" " second " , $\alpha_2 = \sum_{i=m+1}^{n} x_i / n-m$

For Points in first Cluster

$$\|x - \alpha_1\|^2 < \|x - \alpha_2\|^2$$

$$2(\alpha_2 - \alpha_1) \cdot x + (\alpha_1 \cdot \alpha_1 - \alpha_2 \cdot \alpha_2) < 0$$

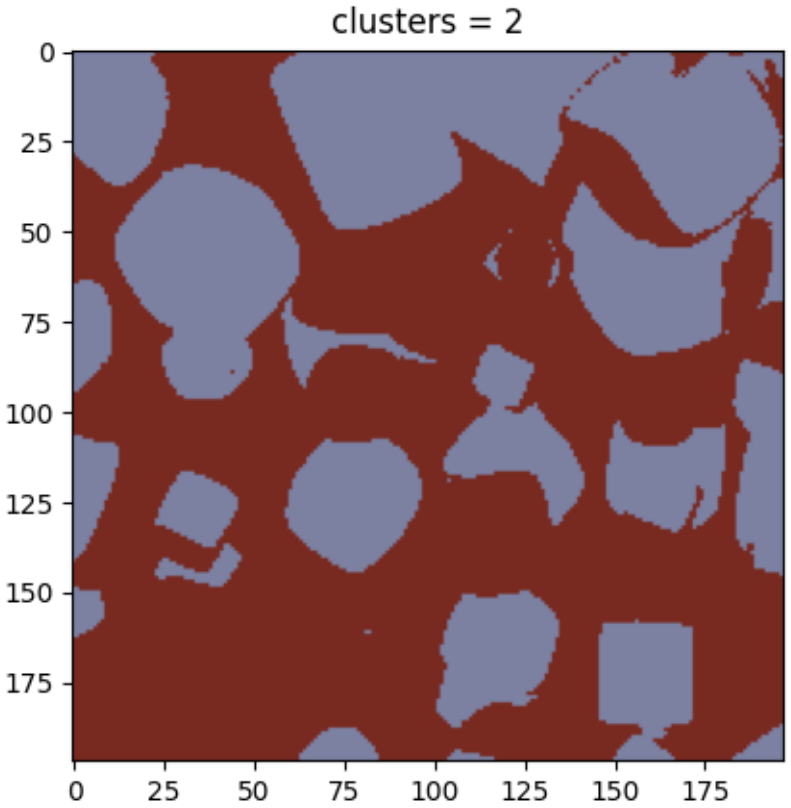
For Points in second Cluster

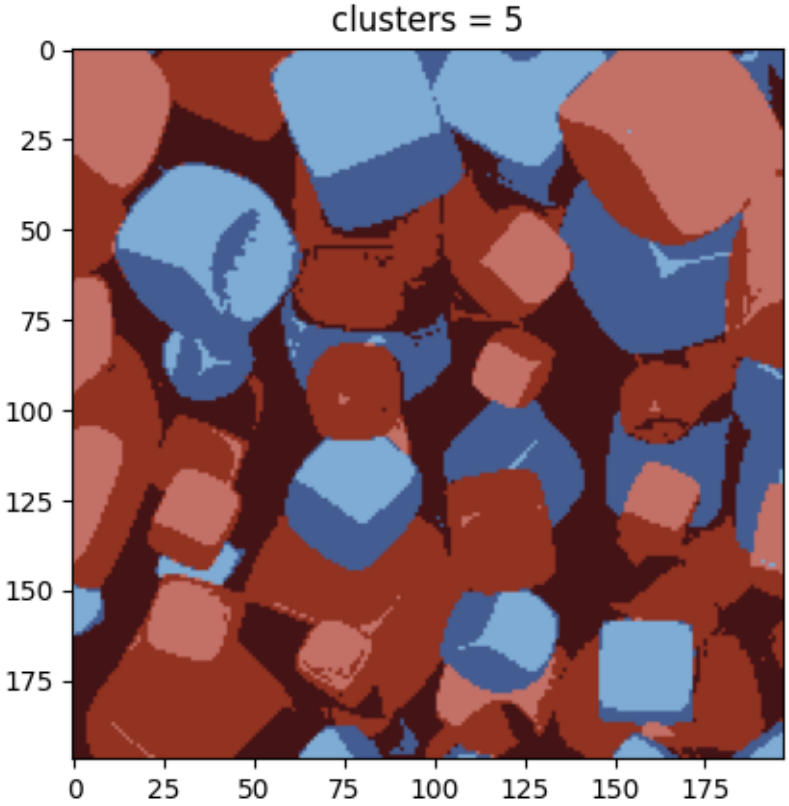
$$\|x - \alpha_1\|^2 > \|x - \alpha_2\|^2$$

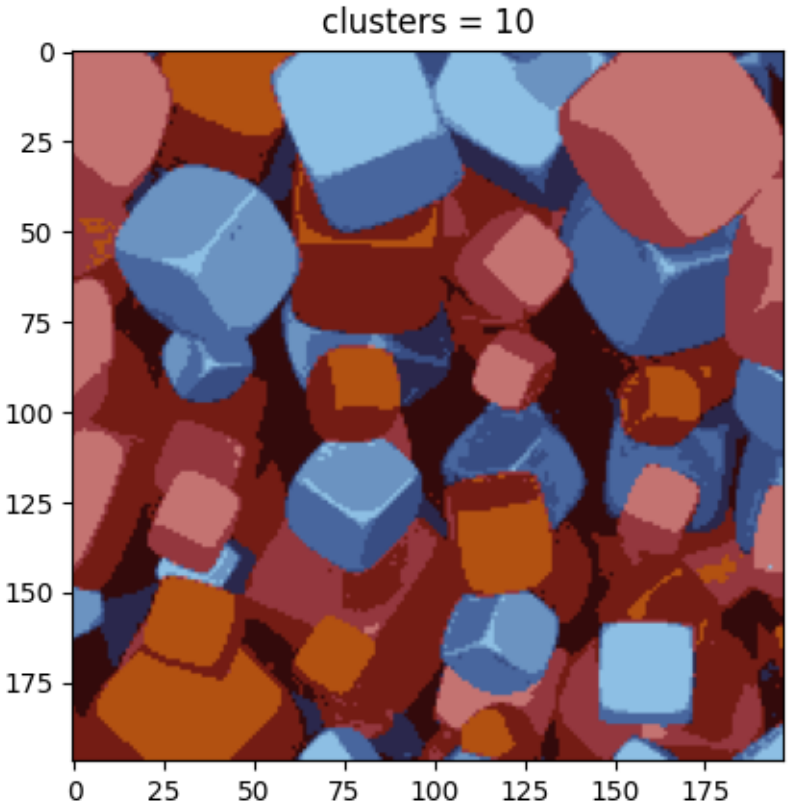
$$2(\alpha_2 - \alpha_1) \cdot x + (\alpha_1 \cdot \alpha_1 - \alpha_2 \cdot \alpha_2) > 0$$

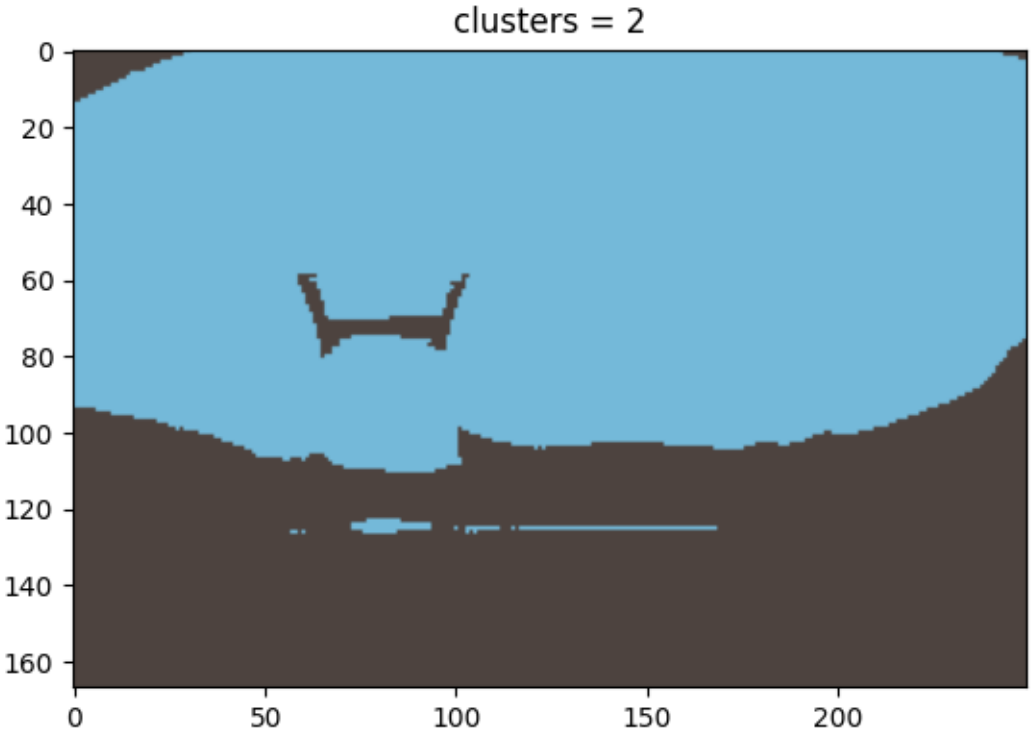
Hence $2(\alpha_2 - \alpha_1) \cdot x + (\alpha_1 \cdot \alpha_1 - \alpha_2 \cdot \alpha_2) = 0$

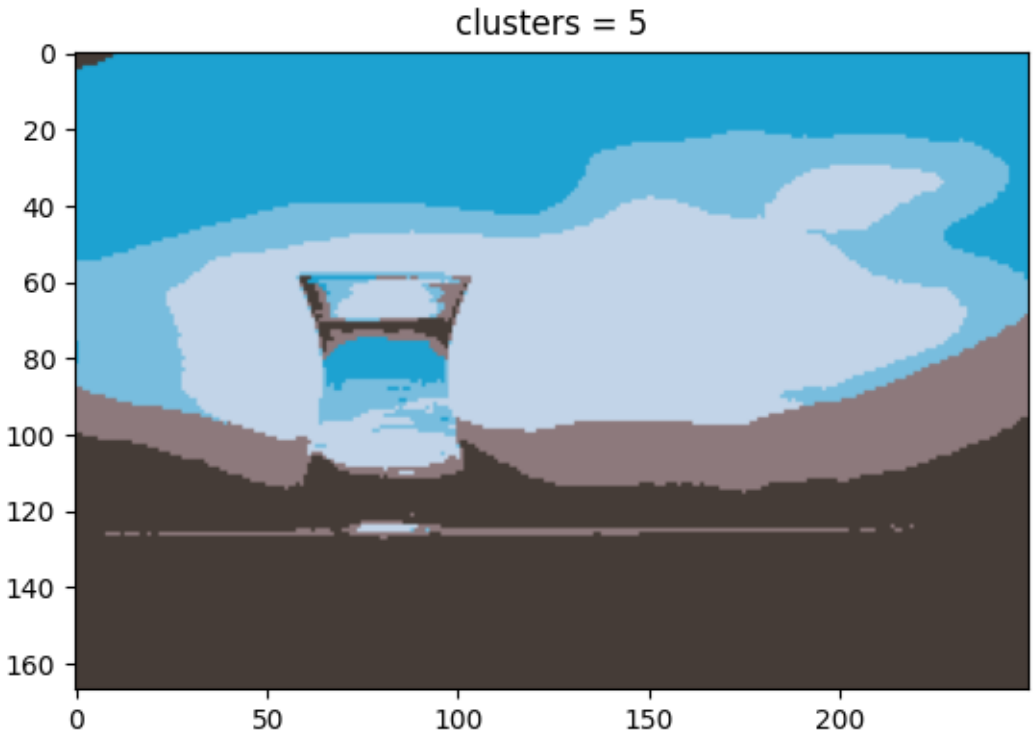
is the sheating plane

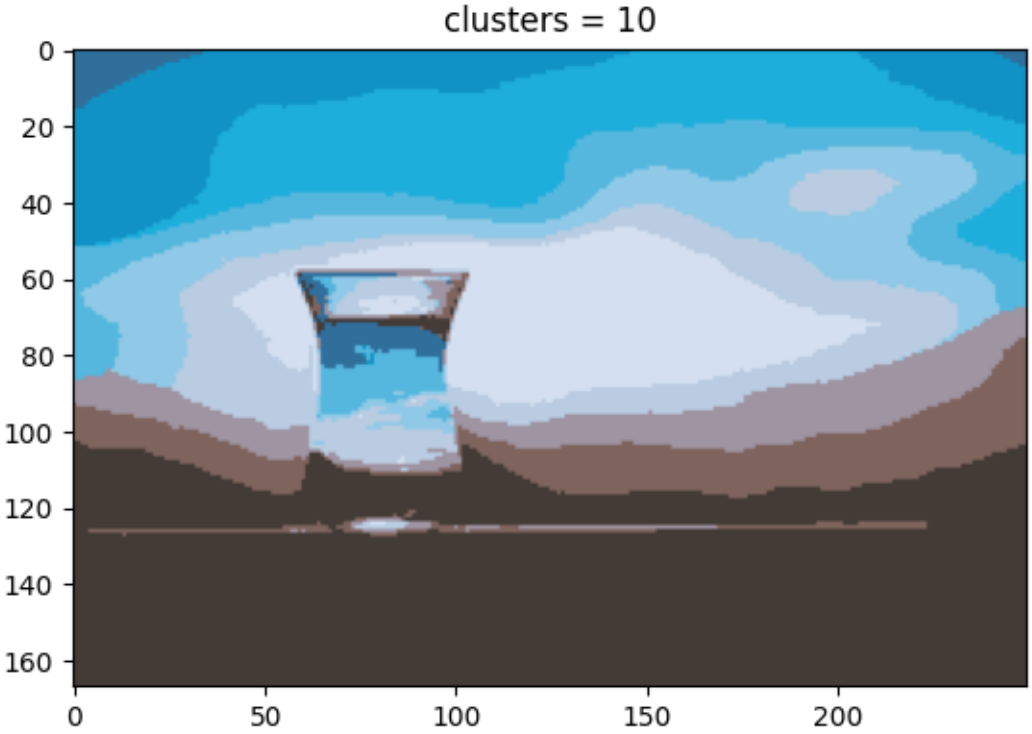

clusters = 2
0
25
50
75
100
125
150
175
0
25
50
75
100
125
150
175

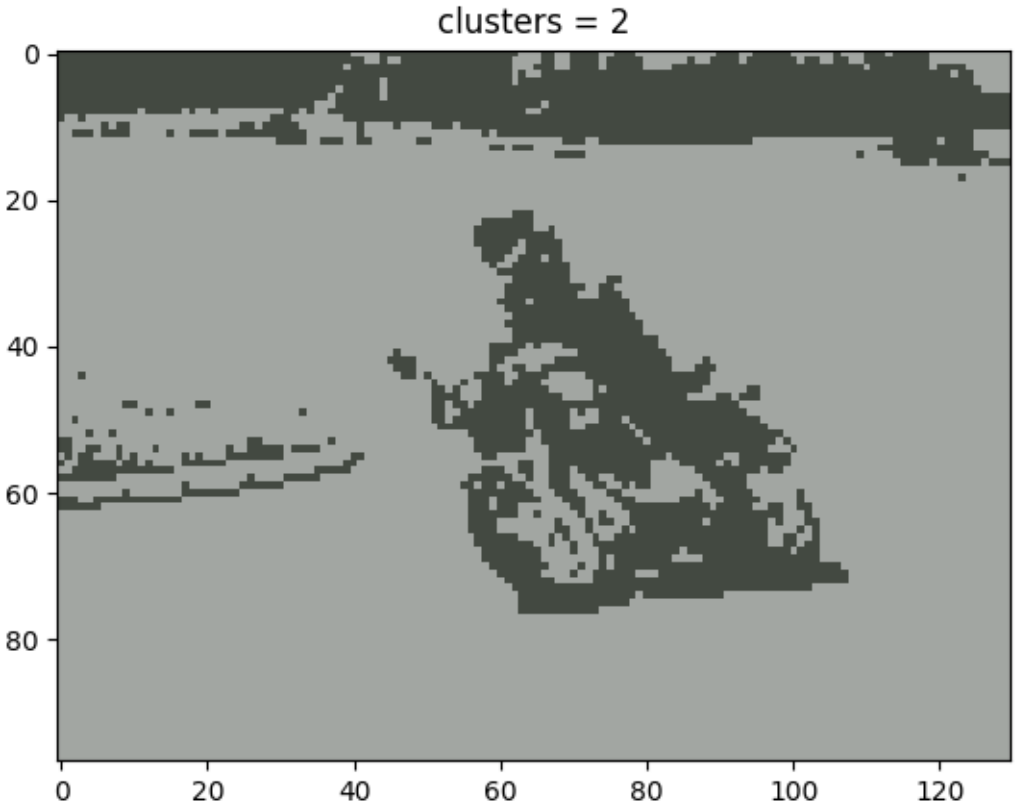

clusters = 5
0
25
50
75
100
125
150
175
0
25
50
75
100
125
150
175

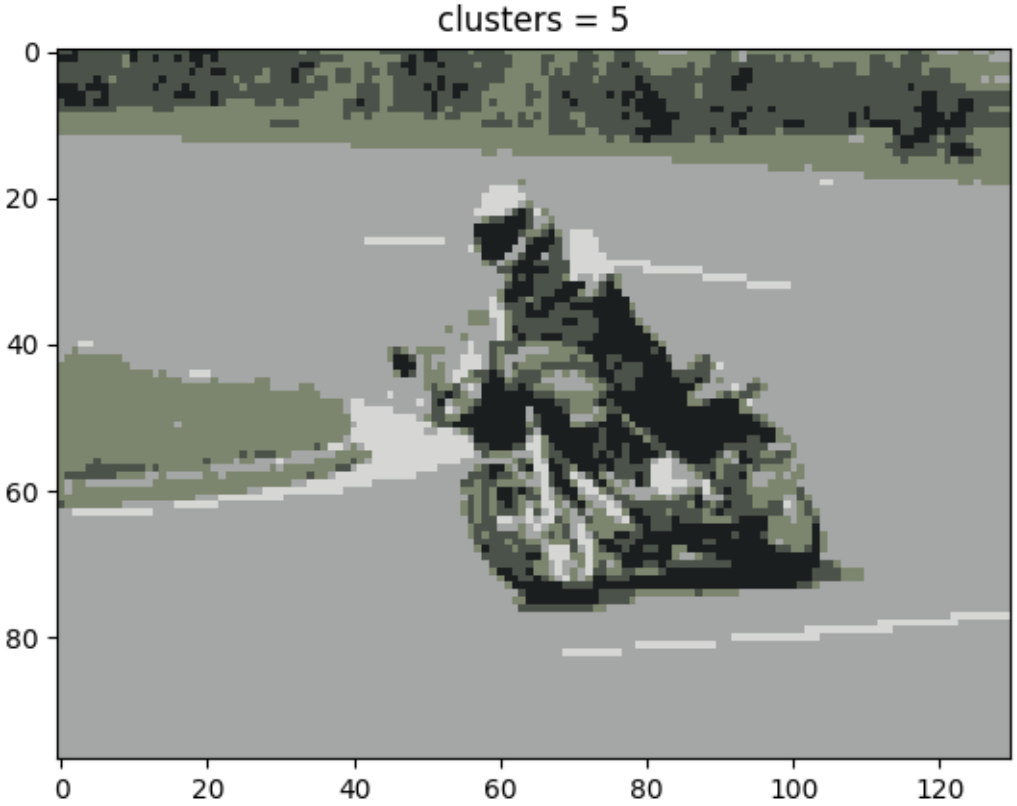

clusters = 10
0
25
50
75
100
125
150
175
0
25
50
75
100
125
150
175

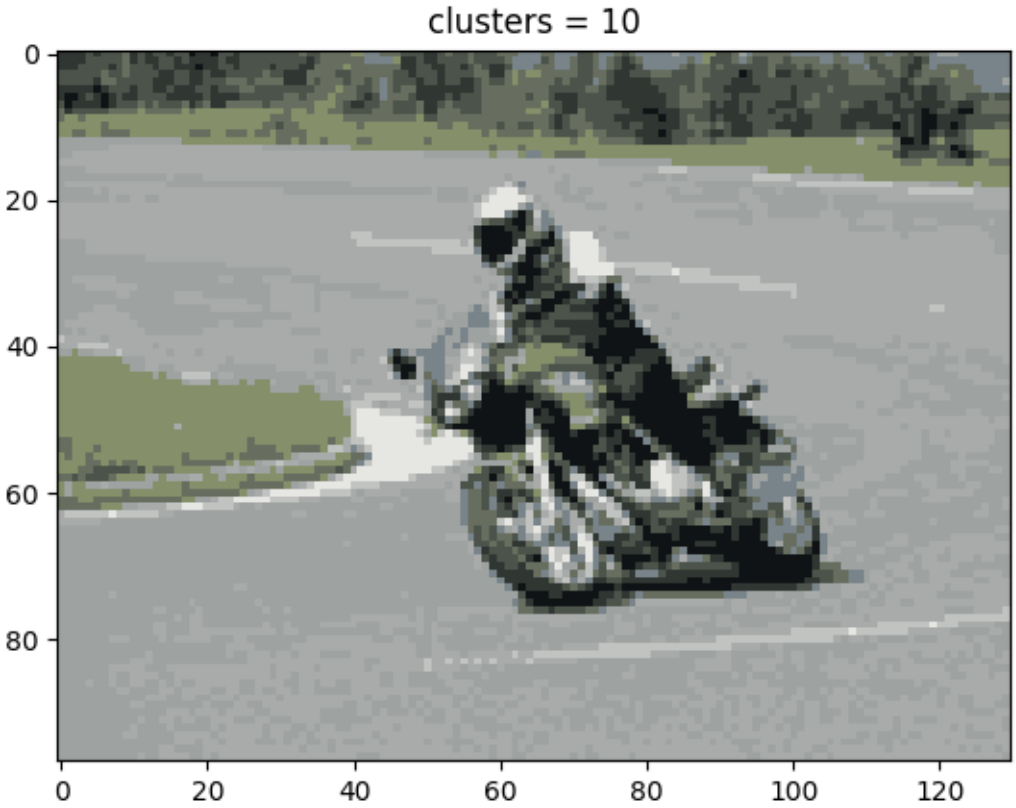

clusters = 2
0
20
40
60
80
100
120
140
160
0
50
100
150
200


clusters = 5
0
20
40
60
80
100
120
140
160
0
50
100
150
200


clusters = 10
0
20
40
60
80
100
120
140
160
0
50
100
150
200


clusters = 2
0
20
40
60
80
0
20
40
60
80
100
120


clusters = 5
0
20
40
60
80
0
20
40
60
80
100
120


clusters = 10
0
20
40
60
80
0
20
40
60
80
100
120

Sajal
Date : / /20
Page :

3.2 (iii) The amount of details present in the original image like numbers of colors, textures govern the minimum number of clusters required to retain most of details