Karan Agarwalla – 180050045

Problem 1

Problem 1.1

$$K_\sigma(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}} = e^{-\frac{(x-y)^T(x-y)}{2\sigma^2}} = e^{-\frac{\|x\|^2+\|y\|^2-2x^Ty}{2\sigma^2}} = e^{-\frac{\|x\|^2}{2\sigma^2}} e^{\frac{x^Ty}{\sigma^2}} e^{-\frac{\|y\|^2}{2\sigma^2}}$$

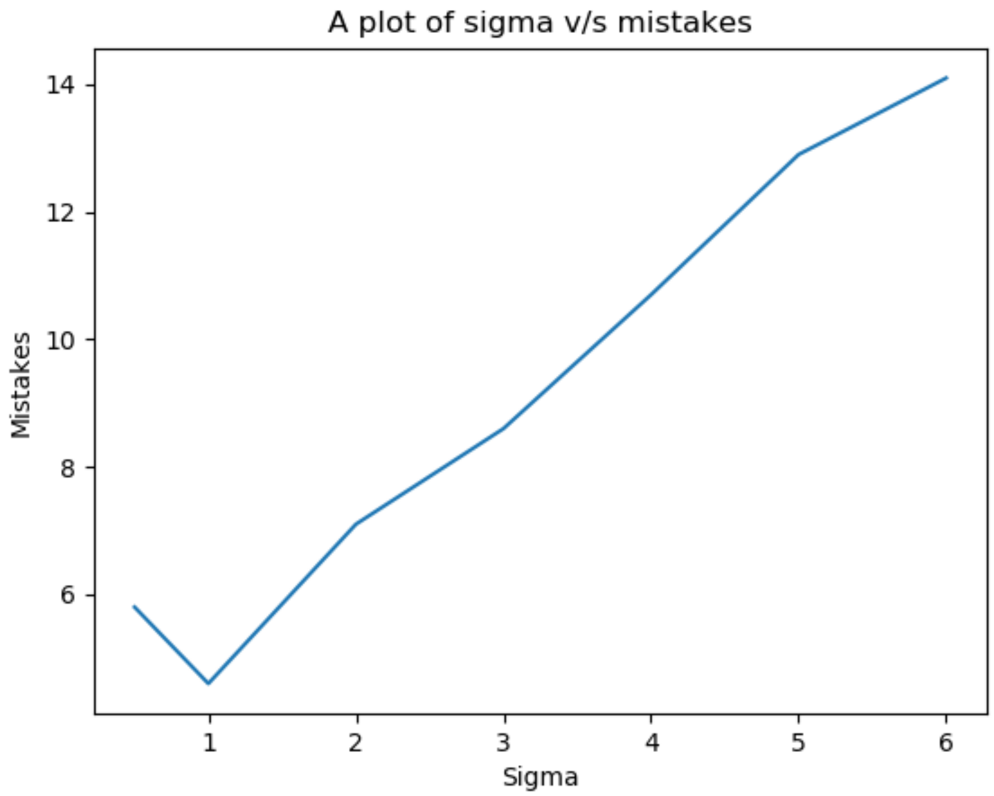
Now, $e^{\frac{x^Ty}{\sigma^2}} = \sum_{i=0}^{\infty} \frac{(x^Ty)^i}{i!\sigma^{2i}}$

Now we proved in class that $(x^Ty)^d$ is a valid kernel for $\forall d \in W$. Also, if K_1 and K_2 are valid kernels then K_1K_2 and aK_1+bK_2 are valid kernels for a>=0, b>=0. As x^Ty is a valid kernel, then by first of the results $(x^Ty)^d$ is a valid kernel. Then $e^{\frac{x^Ty}{\sigma^2}}$ is valid kernel as it is a positive linear combination of valid kernels.

If $K(x,y)$ is a valid kernel then $K'(x,y) = f(x)K(x,y)f(y)$ is also a valid kernel. Let $K(x,y) = \phi(x)^T\phi(y)$ then $K'(x,y) = f(x)\phi(x)^T\phi(y)f(y) = \phi'(x)^T\phi'(y)$ where $\phi'(x) = f(x)\phi(x)$. Hence $K'(x,y)$ is a valid kernel.

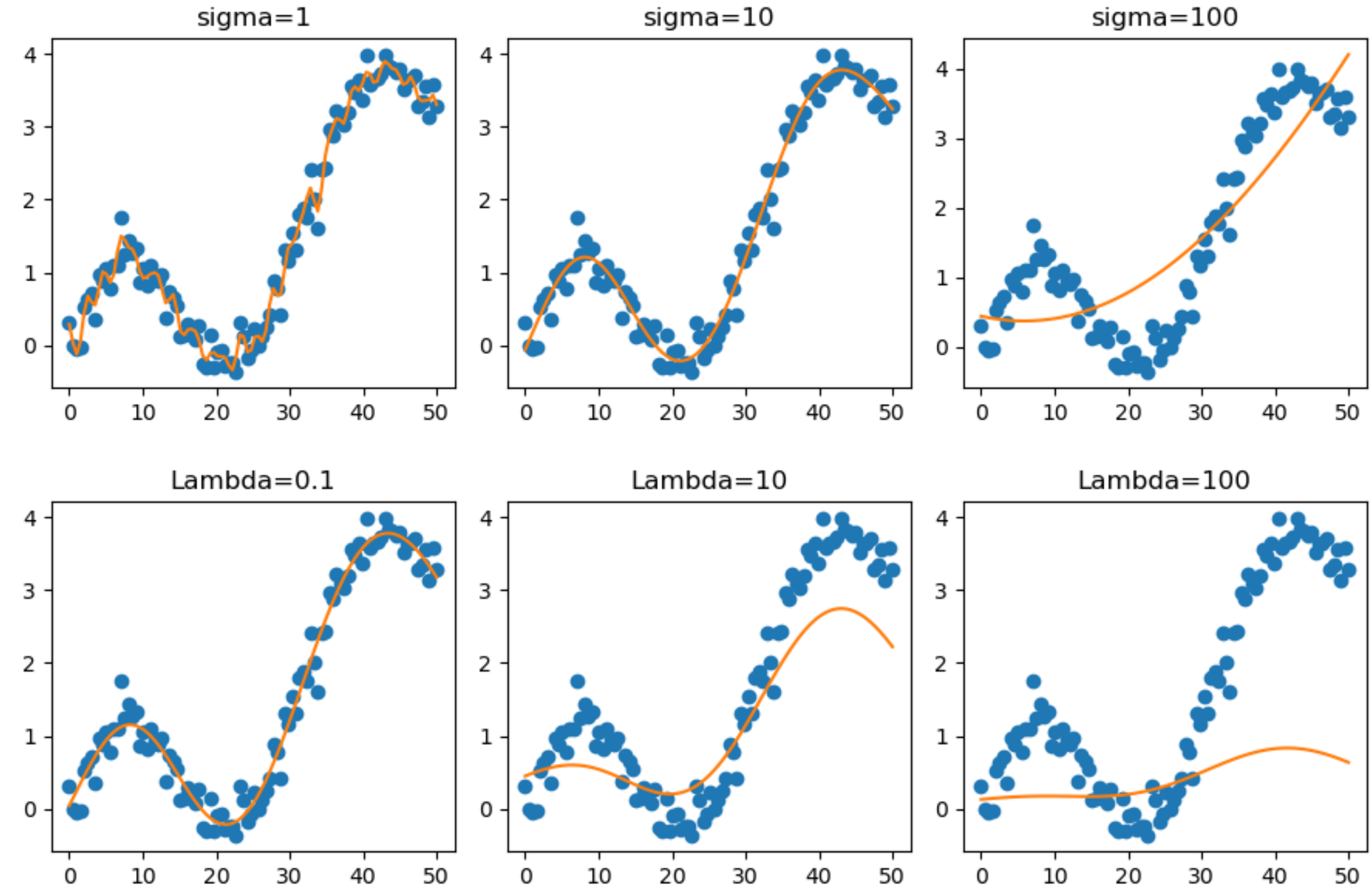
In our case $f(x) = e^{-\frac{\|x\|^2}{2\sigma^2}}$. Therefore $K_\sigma(x,y)$ is a valid kernel.

Problem 1.2
(b)
(ii)
The best value of $\sigma = 1$ as it has minimum number of mistakes.



(iii) The number of mistakes decreases initially then increases further. At low values of σ (low spread) overfitting occurs causing more mistakes. As σ increases (high spread) underfitting occurs causing more mistakes.

(c) (ii)



We observe overfitting at lower values of σ (low spread) and underfitting at higher value of σ (high spread). The spread is determined by σ of the distribution. One observes that at $\sigma = 1$ the curve tries to overfit the data and at $\sigma = 100$ it underfits the data.

Also, higher the value of λ higher is the regularisation term resulting in underfitting of data. Also, as value of λ increases, the curvature of the curve decreases (determined by ratio of maximum and minimum eigenvalues).

Problem 2

Problem 2.1

(i) Given: $K(x, x')$ is valid kernel where $x \in R^m$ and $g: R^m \rightarrow R^m$.

To show: $K\big(g(x), g(x')\big)$ is a valid kernel

Proof:

Since $K(x, x')$ is positive definite kernel, $\exists\ \phi: R^m \rightarrow H$ such that

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

Consider $\phi_g: R^m \rightarrow H$ such that $\phi_g(x) = \phi(g(x))$

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

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

(ii) Given: $K(x, x')$ is valid kernel where $x \in R^m$ and a polynomial with non $-$ negative coefficents q.

To show: $q(K(x, x'))$ is a valid kernel

Proof: $q(x) = \sum a_i x^i \ where \ \forall\ i\ a_i \geq 0$

$q\big(K(x,x')\big) = = \sum a_i K(x,x')^i \ where \ \forall\ i\ a_i \geq 0$

If K_1 and K_2 are valid kernels then $K = aK_1 + bK_2$ is a valid kernel for $a >= 0, b >= 0$. Let $K_1(x,x') = \phi_1(x)^T\phi_1(x')$ and $K_2(x,x') = \phi_2(x)^T\phi_2(x')$. Then $K(x,x') = \phi(x)^T\phi(x')$ where $\phi(x) = [\sqrt{a}\phi_1(x)\ ;\ \sqrt{b}\phi_2(x)]$.

If K_1 and K_2 are valid kernels then $K = K_1K_2$ is a valid kernel for $a >= 0, b >= 0$. Let $K_1(x,x') = \phi_1(x)^T\phi_1(x')$ and $K_2(x,x') = \phi_2(x)^T\phi_2(x')$. Then $K(x,x') = \phi(x)^T\phi(x')$ where $\phi(x)$ is vector of size $n * l$ where $\phi_1(x)$ is of size n and $\phi_2(x)$ is of size l. Then $\phi(x)_i = \phi_1(x)_{\frac{i}{l}}\phi_2(x)_{i\%l}$ where division is integer division.

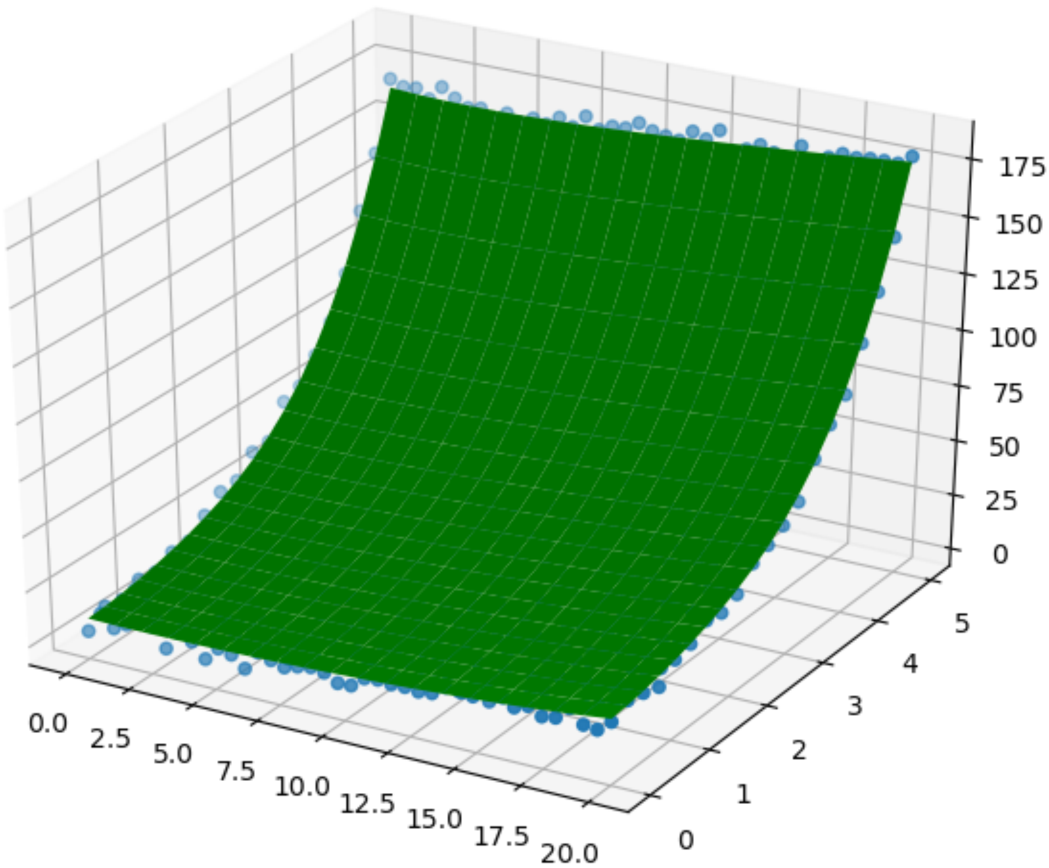
Define $K_i(x,x') = K(x,x')^i$.

$K_i(x,x')$ is a valid kernel as it is multiplication of valid kernels and hence, by second property it is a valid kernel.

Then $q(K(x,x'))$ is positive linear combination of valid kernels, hence is a valid kernel.

Problem 2.2

$K(x,x') = (1 + x^Tx')^4$ is a polynomial kernel with d = 4.



Error of the fit: 6935.36888

Problem 3

Problem 3.1

Given: Optimal $2-$class clustering of $x^1, \ldots, x^n$ where $x^i \in R^d$.
To show: $\exists$ a hyperplane $a.x + b = 0$ *where* $a \in R^d, b \in R$ that seperates the two classes
Proof: Let us show that one of the required hyperplanes is the perpendicular bisector plane between the two cluster centers $\mu_1 = \frac{\sum_{i=1}^{m} x^i}{m}$ and $\mu_2 = \frac{\sum_{i=m+1}^{n} x^i}{n-m}$. Now consider

$$V_i = \| x^i - \mu_1 \|^2 - \| x^i - \mu_2 \|^2$$

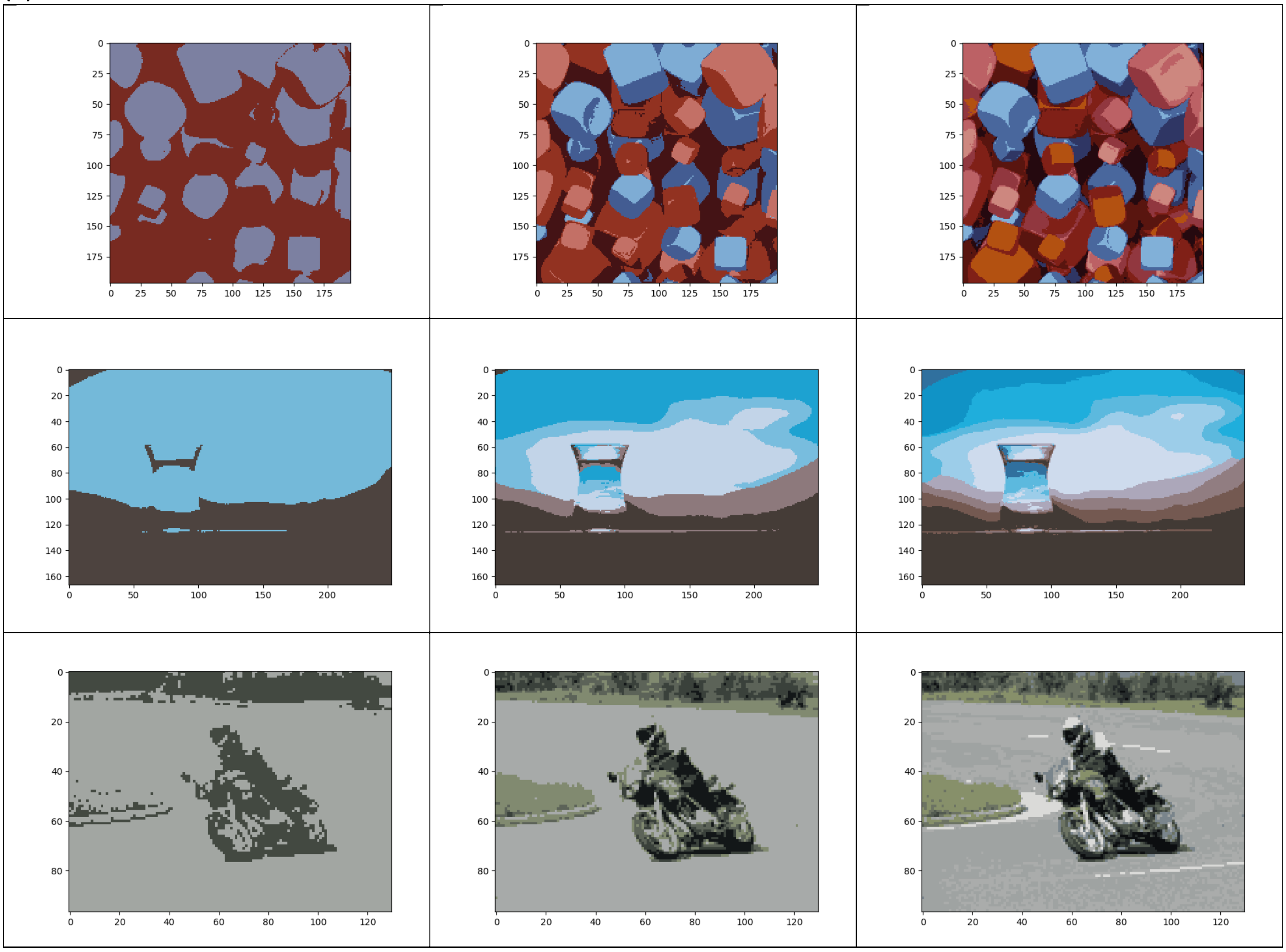
Then by definition of clustering, if $V_i < 0$ then it lies in cluster 1 else it lies in cluster 2. Also $V_i = 0$ is the perpendicular bisector plane.

$$V_i = \left(\| x^i \|^2 - 2\mu_1^T x^i + \| \mu_1 \|^2\right) - \left(\| x^i \|^2 - 2\mu_2^T x^i + \| \mu_2 \|^2\right)$$
$$= 2(\mu_2 - \mu_1)^T x^i + \| \mu_1 \|^2 - \| \mu_2 \|^2$$

By property of planes, we know if and only if two points x and y lie on the same side of plane, $V(x)V(y) > 0$. Now for all points in cluster 1, $V_i < 0 =>$ all points in cluster 1 lie on same side of plane defined by $V = 0$. A similar argument holds for cluster 2. Now for two points in cluster 1 and cluster 2, say x *and* y, $V(x)V(y) < 0 =>$ they lie on opposite sides of the plane. Hence the plane defined by $V_i = 0$ is one such plane where $a = 2(\mu_2 - \mu_1)$ and $b = \| \mu_1 \|^2 - \| \mu_2 \|^2$.

Problem 3.2

(ii)



With increase in number of cluster centers, we observe increase in detail in the image. This can be seen from the fact that with increase in number of cluster centers, the more related pixels can be assigned to a cluster, resulting in better detail.

(iii) The pixels are assigned the color of the centroid. So with increase in number of cluster centroids, number of colors in the final image increases resulting in more detail. Hence images having fewer dominant colors(like image 3) look fine with smaller number of clusters as compared with others(like images 1 & 2) needing more clusters.