Assignment 4 — Machine Learning

Tarush Goyal

October 30, 2020

1 Kernel Methods

1.1

$$
\begin{aligned}
&K_\sigma(x, y) \\
&\quad = exp(-\frac{\|x-y\|^2}{2\sigma^2}) \\
&\quad = exp(-\frac{x^T x}{2\sigma^2}) * exp(\frac{x^T y}{\sigma^2}) * exp(-\frac{y^T y}{2\sigma^2})
\end{aligned}
$$

Taylor expansion of e^x:

$$exp(x) = \sum_0^\infty \frac{x^i}{i!}$$

So,

$$exp(\frac{x^T y}{\sigma^2}) = \sum_0^\infty \frac{(\frac{x^T y}{\sigma^2})^i}{i!} = \sum_0^\infty \frac{(x^T y)^i}{i!\sigma^{2i}}$$

Class results:

- $(x^T y)^d$ is a valid kernel
- if K is a valid kernel so is C.K where C is a constant
- if K_i is valid, so is $\sum \alpha_i K_i$ if $\sum \alpha_i^2$ is finite

Since $\sum \frac{1}{i!}^2 < \sum \frac{1}{i!} = e$ is finite, thus, $exp(\frac{x^T y}{\sigma^2})$ is a valid kernel as each term is a valid kernel. Now, K(x,y) is valid if for any square integrable function g(.) the following holds:

$$\int K(x, y) g(x) g(y) dx dy > 0$$

If g(x) is square integrable, so is $exp(-\frac{x^T x}{2\sigma^2}) g(x)$ because $g^2(x) >= (exp(-\frac{x^T x}{2\sigma^2}))^2 g^2(x)$. Thus, for all square integrable g and valid K, the following holds:

$$\int K(x, y).(exp(-\frac{x^T x}{2\sigma^2}) g(x)).(exp(-\frac{y^T y}{2\sigma^2}) g(y)) dx dy > 0$$

$$\int (exp(-\frac{x^T x}{2\sigma^2}) K(x, y) exp(-\frac{y^T y}{2\sigma^2})) g(x) g(y) dx dy > 0$$

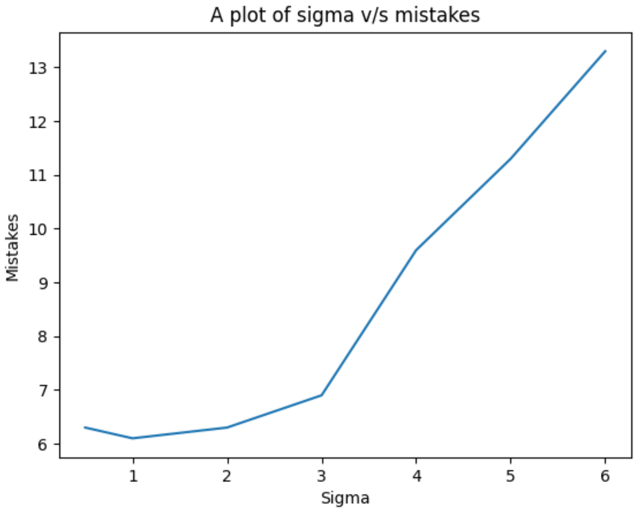
Substituting K(x,y) as $exp(\frac{x^T y}{\sigma^2})$ we get

$$\int K_\sigma(x,y)g(x)g(y)dxdy > 0$$

Hence Proved

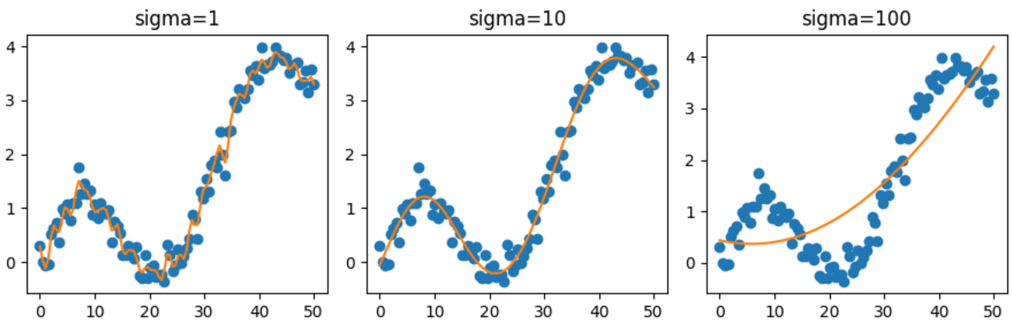
1.2

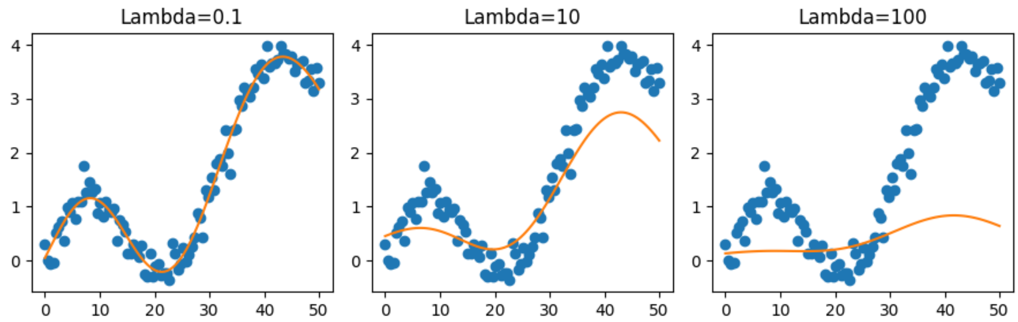
(b) (ii) The ideal value of σ is 1 as it shows the least number of incorrect labels for 10-fold cross validation.



(iii) RBF kernel can be observed as an infinite polynomial with decreasing weights. By increasing σ, we decrease the weight given to higher powers of this polynomial and hence the effective complexity of the model decreases. Thus as σ increases, the model under-fits the data. But on reducing it too much the model over-fits and reduces test accuracy. So there is an optimal value (= 1 in this case)

(c) Both λ and σ are parameters that can be used to control amount of overfitting or underfitting. However they are based on different mathematical ideas. λ reduces overfitting by reducing the weights and hence their variance. σ reduces overfitting by decreasing the variance between various training examples by controlling the complexity of ϕ. So we obtain a smoother curve for both the graphs as σ or λ increase. However, since λ reduces the weights, predictions reduce in value as λ increases whereas σ just makes it smoother.





2 Kernel Design

2.1

(i) Let ϕ be the basis function corresponding to the kernel K, that is $K(x, x') = \phi(x)^T\phi(x')$. Then this is an identity in x and x'. So we can substitute x with g(x) and x' with g(x') (as g preserves dimensions) to get $K(g(x), g(x')) = \phi(g(x))^T\phi(g(x'))$. Let $\phi'(.)$ be $\phi(g(.))$. Then $K'(x, x') = K(g(x), g(x')) = \phi'(x)^T\phi'(x')$. Thus there exists a basis function corresponding to our new kernel K' for all x and x'. So it is a valid kernel.

(ii) We have proven in class that $(\phi(x)^T\phi(x'))^d$ is a valid kernel. Let K(x,x') be $\phi(x)^T\phi(x')$. Then $K(x, x')^d = (\phi(x)^T\phi(x'))^d$. Now $q(K(x, x')) = \sum_{d=0}^{N} a_d K(x, x')^d$. Using the fact that sum of valid kernels is also a valid kernel and constant times kernel is also valid, we get that $a_d K(x, x')^d$ is valid and hence $\sum_{d=0}^{N} a_d K(x, x')^d = q(K(x, x'))$ is also valid.

2.2

- Kernel chosen :

$$K(x, y) = 0.01 * exp(x_1^T y_1) + (x_2^T y_2)^2$$

- Error obtained : 6925
- Validity : based on the fact that individually the two terms are valid kernels so their sum is also valid.
- Plot obtained

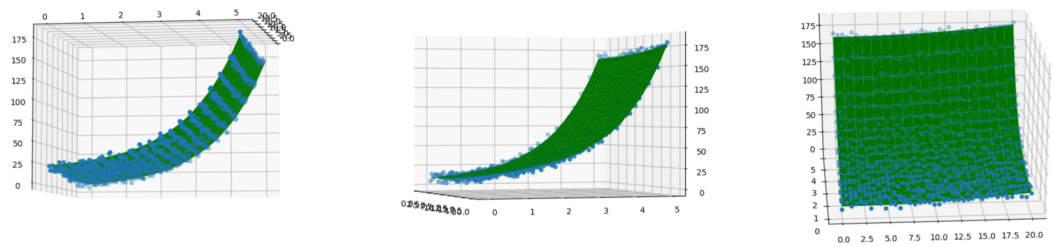
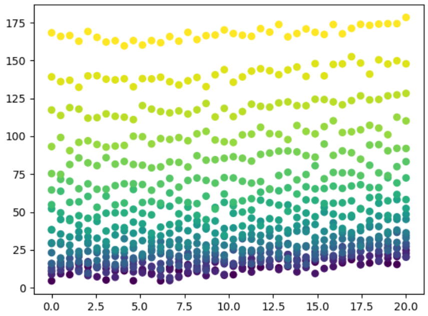
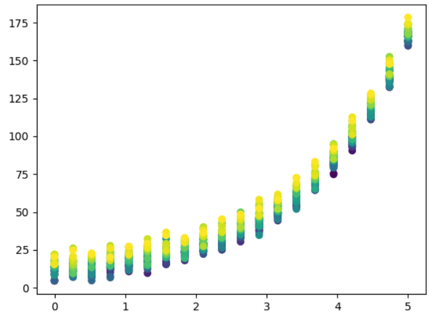


Figure 1: Plot obtained at various angles

- Explaination : The plots below represent my observations for variation of target value with x_1 and x_2 respectively. Same color represents closer target value. Clearly x_1 has very minor effect on the result. However removing it from the kernel gives poor results. So I have added a low degree polynomial (quadratic) to make its contribution. x_2 has an exponential effect on the target value. So first term of the kernel represents that. 0.01 is just a balancing value as exponential term rises way faster than polynomial one.





3 K-means clustering for Image compression

3.1

Let the center (mean) of first cluster be c^1 and that of second of cluster be c^2. Let P be the hyperplane of points equidistant to c^1 and c^2. So, as given, no point lies on P. Consider a point x^i for $i <= m$. It belongs to cluster 1. This means that $dist(x^i, c^1) < dist(x^i, c^2)$. Thus it lies on the same side of P as c^1. Similarly for $i > m$, x^i lies on the same side of P as c^2. Thus P separates the two clusters. Now we can find the values of **a** and b:

$$dist(x - c_1) = dist(x - c_2)$$

$$\sum_i (x_i - c_{1,i})^2 = \sum_i (x_i - c_{2,i})^2$$

$$\sum 2(c_{2,i} - c_{1,i})x_i = \|c_2\|^2 - \|c_1\|^2$$

$$\therefore a = 2\left(\frac{\sum\limits_{i=m+1}^{n} x_i}{n-m} - \frac{\sum\limits_{i=1}^{m} x_i}{m}\right)$$

$$b = \left|\sum_{i=1}^{m} x^i/m\right|^2 - \left|\sum_{i=m+1}^{n} x^i/(n-m)\right|^2$$

3.2

(i)
- Here k = 2 could just separate 'light-facing' faces of the cubes and the faces away from light. k = 5 reproduces most of the colors, but could not differentiate between various faces of a cube (which vary due to different shades (due to different amount of light exposure). k = 10 reproduces most of the image correctly.

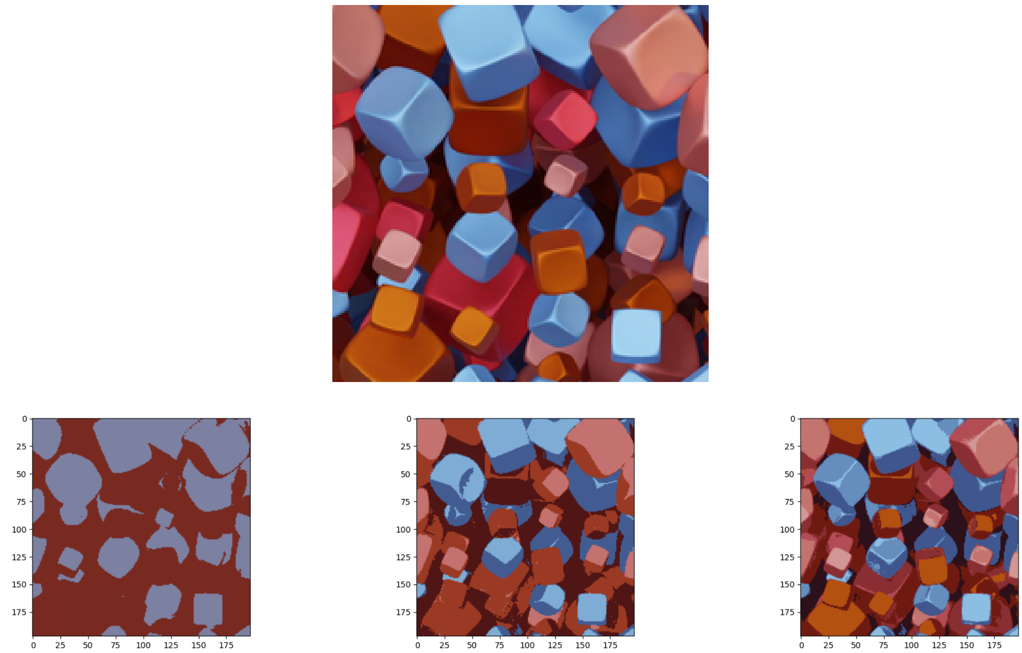


Figure 2: top : original image, bottom : k = 2, k = 5, k = 10

- Image 2 : Here k = 2 recognises dark parts of the image and boundaries. K = 5 correctly recognizes the colors but could not take care of the huge amount of shading in the original image because of just 5 colors. For k = 10, we have successfully extracted multiple shades from the sky and ground.

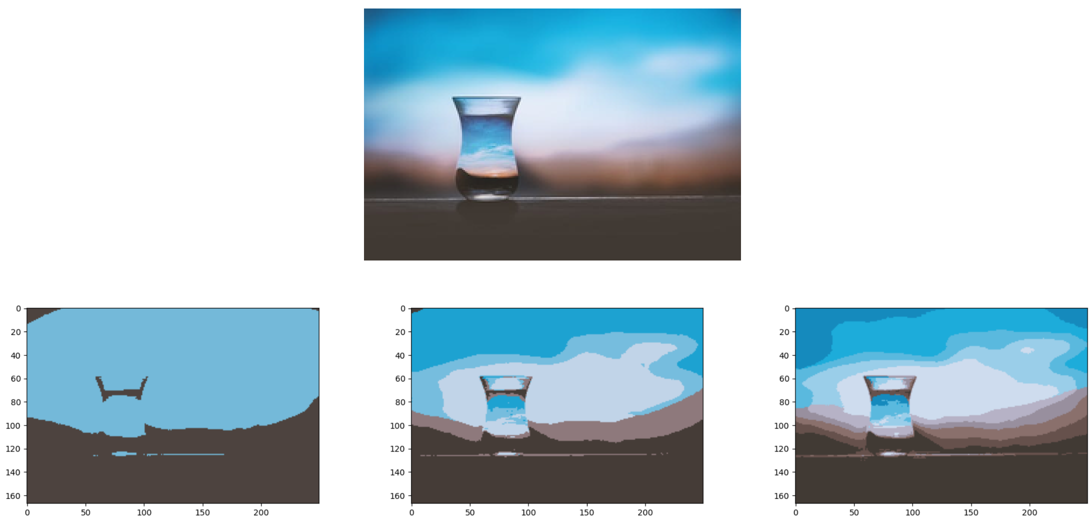


Figure 3: top : original image, bottom : k = 2, k = 5, k = 10

- Image 3 : This image does not have much shade variation (due to uniform light exposure). So k = 2 simple separated the already darker parts of the image from the lighter parts. k = 5 and 10 performed nearly the same as there are very few colors and almost no shade variation.

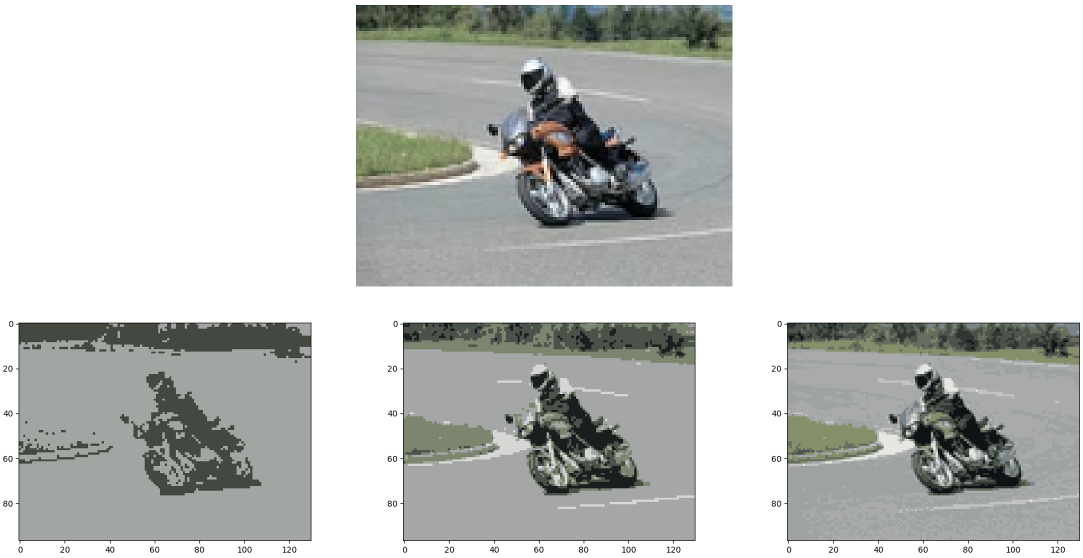


Figure 4: top : original image, bottom : k = 2, k = 5, k = 10

(ii) As we increase k, we add more colors to the image. However after certain k, the new information added by a new color(as judged by the naked eye) becomes nearly 0. Generally such a k is deemed ideal. The ideal number of k depends on the amount of colors and shades in the image.Image 1, has various faces with various intensity of light and hence various shades making k = 10 better. Image 3, being an open daylight photo,has low number of colors and shades and hence k = 5 seems ideal. In Image 2, there is extensive shading in the colors so as to promote higher k. Hence k = 10 seems ideal.