CS-335 / 337 Assignment-4

Niraj Mahajan - 180050069

Q-1.1). → Property-1 (Proved in class)

If K_1, K_2 are Kernels, then
$\alpha_1 K_1 + \alpha_2 K_2$ is also a kernel $\forall \alpha_1, \alpha_2 \geq 0$

→ Property-2 (Proved in class)

If K_1, K_2 are kernels, then $K_1 \cdot K_2$ is also a kernel.

→ Property-3 If K is a kernel, K^d is also a kernel, $\forall d \in N$

Proof by Induction :-
Base Case, $d=1$ $K^1 = K$ (pressumption) is a kernel.

Induction Hypothesis :- K^d is a kernel.
Induction Step : To prove K^{d+1} is a kernel,
$K^{d+1} = K \cdot K^d$

Using Property 2, since K, K^d are valid kernels, then K^{d+1} is a valid kernel
Hence proved.

⇒ Now, we need to prove

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

$$.\exp\left(\frac{x^Ty}{\sigma^2}\right)$$

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

In the RHS, just consider the middle term.

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

$$= \sum_{n=0}^{\infty} \frac{(x^Ty)^n}{n!\cdot\sigma^{2n}}$$

Now, since $K(x,y) = x^Ty$ is a valid kernel, using property 1 and property 3, we get $K'(x,y)$ is also a valid kernel.

Hence, there exist a $\phi(x) : R^n \to H$.

s.t $K'(x,y) = \phi(x)^T\phi(y)$.

Let us define $\phi_{new} = \phi(x) . \exp\left(\frac{-1}{2\sigma^2}x^Tx\right)$.

↳ scalar.

∴ using this ϕ_{new}, we can create a kernel

$$\therefore \phi_{new}(x)^T \phi_{new}(y) = \phi(x)^T \phi(y) \cdot \exp\left(\frac{-1}{2\sigma^2} x^T x\right) \cdot \exp\left(\frac{-1}{2\sigma^2} y^T y\right).$$

$$= \exp\left(\frac{x^T y}{\sigma^2}\right) \cdot \exp\left(\frac{-1}{2\sigma^2} x^T x\right) \cdot \exp\left(\frac{-1}{2\sigma^2} y^T y\right).$$

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

$$= K(x, y)$$

Hence Proved, the Gaussian (rbf) kernel can be expressed in terms of $\phi^T(x) \cdot \phi(y)$, Hence, rbf kernel is a valid kernel.

Q-1.2b).(i). ~~Best~~ $\sigma = 1$ (minimum errors at $\sigma = 1$).

(ii) As we ~~increase~~ decrease the σ, we basically are shrinking the neighborhood of points for performing our prediction.
Upto $\sigma = 1$, the errors decrease as we are shrinking the neighborhood to consider just a small meaningful set of neighbours.
But below $\sigma = 1$, our errors increase as ~~we begin~~ our interval shrinks down tremendously and we begin to overfit on our data.

[Plot on the next page].