Anand Kamat
ID # 260773313

Assignment 1

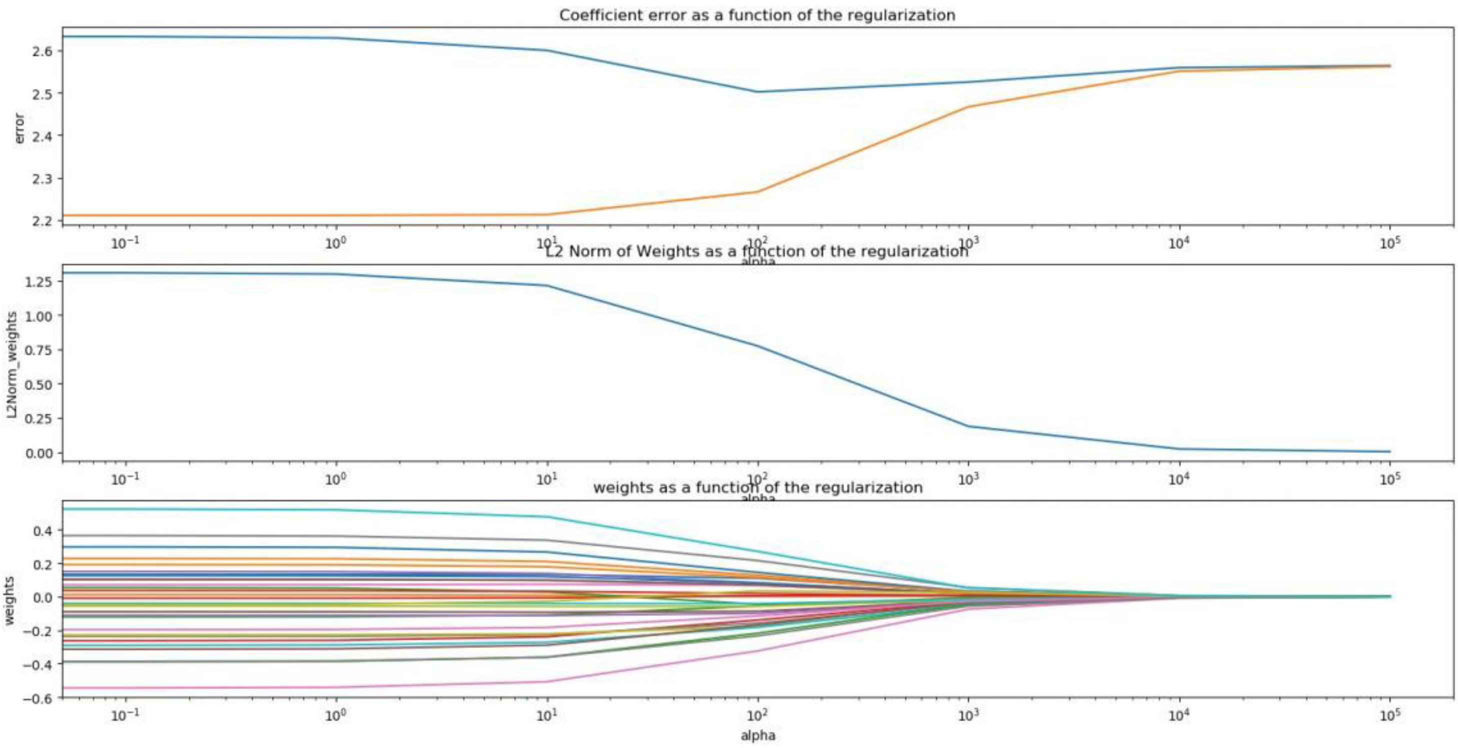
1. a) The only preprocessing done was scaling the data. The bias term was not ~~handlino~~ handled and the data was not centred as the sklearn functions handle all that.

b) The first graph depicts how the two graphs show an increase in error as the norm of weights increase. We also see the functions merging as error increases.
The second graph simply shows a downward sloping function which merges to zero gradually.

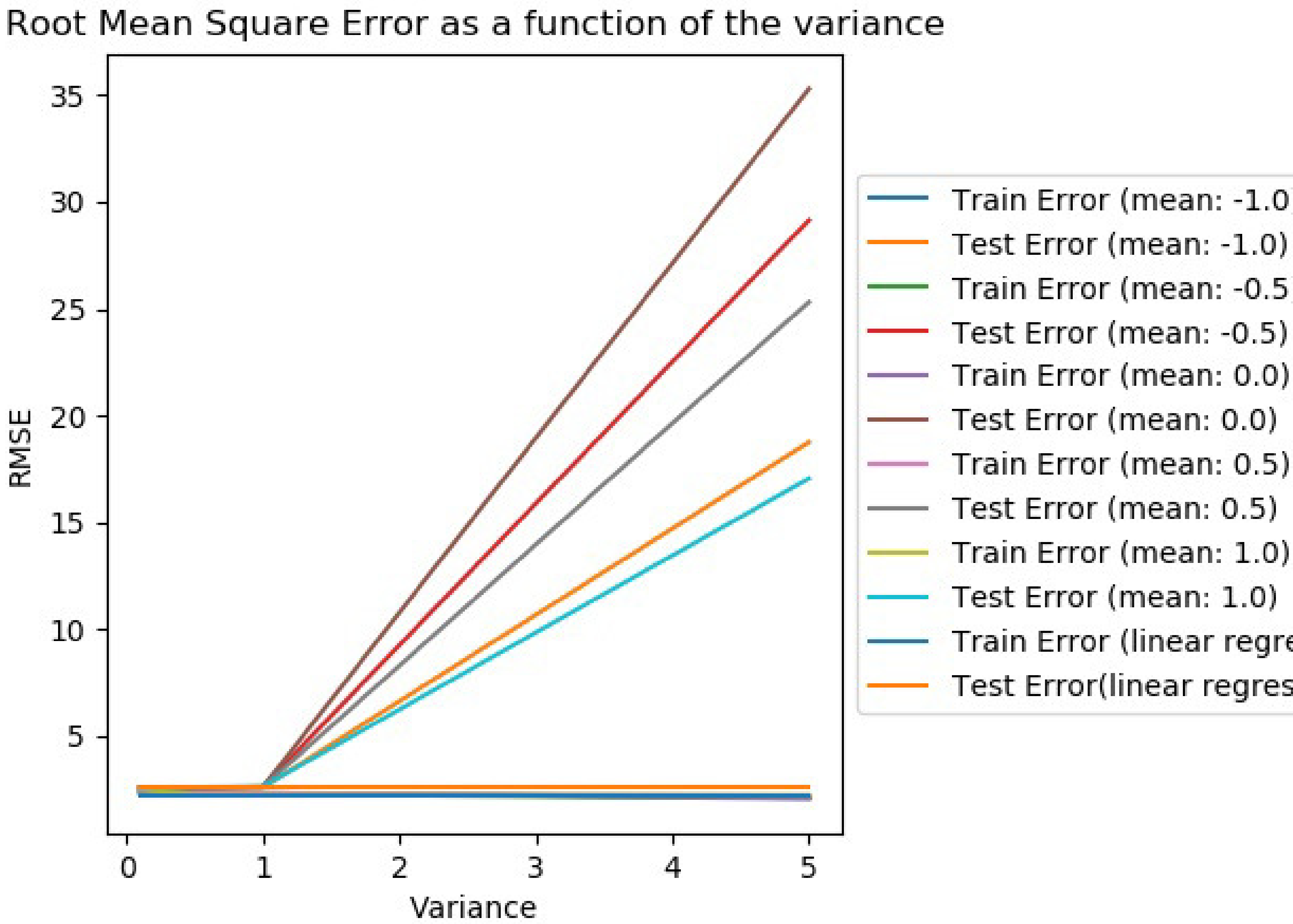
c) The third function depicts the functions converging as alpha increases and tending to zero for higher values of alpha.

d) In this scenario the error would drastically increase as the training data is provided as the model would be trained on a very small data and would fail to hold up to new data when its ~~as~~ exposed to it.

f) σ^2, according to our results increases overfitting, ~~and~~ increasing variance and decreasing bias.



Coefficient error as a function of the regularization
error
alpha
L2 Norm of Weights as a function of the regularization
L2Norm_weights
alpha
weights as a function of the regularization
weights
alpha



Root Mean Square Error as a function of the variance
RMSE
Variance
35
30
25
20
15
10
5
0
1
2
3
4
5
Train Error (mean: -1.0
Test Error (mean: -1.0)
Train Error (mean: -0.5
Test Error (mean: -0.5)
Train Error (mean: 0.0)
Test Error (mean: 0.0)
Train Error (mean: 0.5)
Test Error (mean: 0.5)
Train Error (mean: 1.0)
Test Error (mean: 1.0)
Train Error (linear regre
Test Error(linear regres

* Question Discussed with Amini, Deepanjan

g) There are five basis functions.

$$\phi_i(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu_i)^2/2\sigma^2}$$ where i is one of five functions $i = \{1, 2, 3, 4, 5\}$ which are parametric means.

Finding μ_i along $\vec{w}$ (weight vector) to minimize loss

$$\arg\min J(\omega, \mu) = \frac{1}{2}\sum_{j=1}^{m}\left(h_\omega(x_j) - y_j\right)^2 + \lambda \omega^T\omega \Rightarrow$$

$$\frac{\partial}{\partial \omega_k} J(\omega, \mu) = 0 . \quad \frac{\partial}{\partial \mu_k} J(\omega, \mu) = 0 \quad - (1)$$

$$\frac{\partial}{\partial \omega_k} J(\omega, \mu) = \frac{\partial}{\partial \omega_k} \frac{1}{2}\sum_{j=1}^{m}\left(\left(h_\omega(x_j) - y_j\right)^2 + \lambda \omega^T\omega\right)$$

$$\Rightarrow \phi(x^T)\phi(x)\omega - \phi(x^T)y + \lambda\omega = 0 \quad - (2)$$

$$\Rightarrow \frac{\partial}{\partial \mu_k} J(\omega, \mu) = \frac{\partial}{\partial \mu_k}\frac{1}{2}\sum_{j=1}^{m}\left(h(\omega)(x_j) - (y_j)\right)^2 \Rightarrow \phi(x)^T M \phi(x)\omega - \phi(x^T)My - 0$$

If we ~~kno~~ now use the gradient descent approach, equations (1) & (2) can converge to an optimal solution.

Here M is $R^{200} \times R^{5\times 200}$

b) In (1) the second derivative is positive ($\lambda > 0$). the problem is convex with a global solution. The second condition has no assurity of the derivative being positive. Solving using second derivation. We can hence ~~assume the~~ conclude that the solution only has local optimum.

2. $y_i = h_w(x_i) + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma_i)$

Maximum Likelihood Estimate $= \mathcal{L}(w)$

$$\mathcal{L}(w) = \prod_{i=1}^{m} P(y_i / x_i, w, \sigma_i) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma_i}} e^{-1/2\left(\frac{y_i - h_w(x_i)}{\sigma_i}\right)^2}$$

Using log:

$$\log(\mathcal{L}(w)) = \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi\sigma_i}}\right) - \sum_{i=1}^{m} \frac{1}{2} \frac{(y_i - h_w(x_i))^2}{\sigma_i}$$

$\Rightarrow$ Maximizing Right hand side = minimizing ~~left~~ Left hand side.

$$\Rightarrow \sum_{i=1}^{m} \frac{1}{2} \frac{(y_i - h_w(x_i))^2}{\sigma_i}$$

$$W^* = \arg\min_{w} \sum_{i=1}^{m} \left[\frac{(y_i - h_w(x_i))^2}{(\sigma_i)}\right]$$

3. Using maximum likelihood estimator method

$$P(X/\theta) = \prod_{i=1}^{n} \left(P(x_i/\theta)\right) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$\log P(x/\theta) = \sum_{i=1}^{n} \log\left\{\theta^{x_i}(1-\theta)^{1-x_i}\right\}$$

$$\arg\max_{\theta} P(X/\theta) = \arg\max_{\theta} \log P(X/\theta)$$

Taking partial derivative

$$\hat{\theta}_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\theta}_{ML} = \frac{1}{3}\sum_{i=1}^{3} x_i$$

$$= \frac{1+1+1}{3} = \underline{\underline{1}}$$

No this is not a good estimator as the biase drawn from the data indicates absolute certainity of heads. Hence this wont be the best estimator to make predictions.

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1} (1-\theta)^{\beta-1} \, dx$$

Deriving mean and mode:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$ {Using gamma functions}

$$B(\theta/\alpha, \beta) = \frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)} \, \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

$$\Rightarrow \int_0^1 B(\theta/\alpha, \beta) \, d\theta = 1$$

$$\text{mean} = \frac{\int_0^1 \theta^{\alpha} (1-\theta)^{\beta-1} \, d\theta}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+1)\,\Gamma(\beta)\,\Gamma(\alpha+\beta)}{\Gamma(\alpha+\beta+1)\,\Gamma(\alpha)\,\Gamma(\beta)}$$

$$= \frac{\alpha}{\alpha+\beta} = \frac{1}{2}$$

mode = most likely value of distribution

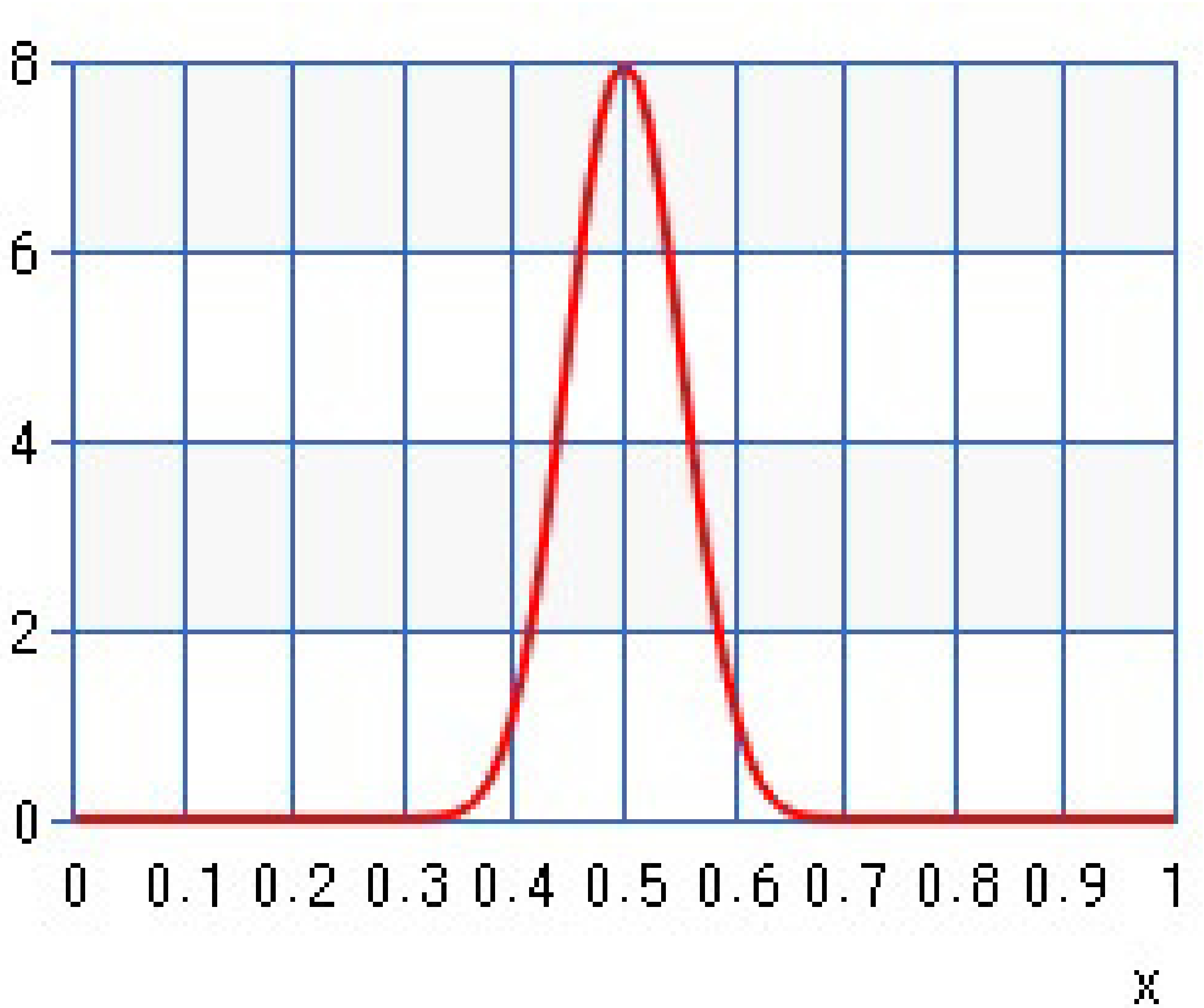
$$= \frac{\cancel{\alpha\pm}\,\alpha-1}{\alpha+\beta-2} = \frac{1}{2}$$

Probability density function:

$$f(\theta/\alpha, \beta) = \frac{1}{B(\alpha,\beta)} \, \theta^{\alpha-1} (1-\theta)^{\beta-1}$$

* Derivations were provided from Bishop

This is a much better estimate than the last example as the data is well distributed.


8
6
4
2
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
x

4.

a) $J(w) = \frac{1}{2}(Xw - Y)^2$ (Function in terms of X & Y)

$\nabla_w J = \nabla_w \frac{1}{2}(X_w - Y)^2 (X_w - Y)$

$= \frac{1}{2} \nabla_w (w^T X^T X_w - Y^T X_w - w^T X^T Y + Y^T Y)$

Rearranging and taking partial derivative

$\nabla_w J = X^T X_w - X^T Y$

* Derivations obtained from ~~ML 652~~ COMP 652 slides

Setting gradient $= 0$

$X^T X_w - X^T Y = 0$

$W = (X^T X)^{-1} X^T Y$

b) $y_t = x_t w_t$

$\sum_{t=1}^{p} y_t = \sum_{t=1}^{p} x_t w_t$

$Y = Xw$

Multiplying both sides by X^T

$X^T Y = X^T X w$

$w = (X^T X)^{-1} (X^T Y)$ which is what we got in (a).

4. (b) $$J(W) = \sum_{i=1}^{m} \| W x_i - y_i \|_2^2$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{p} \| w_j x_i - y_i \|_2^2$$

$$= \sum_{j=1}^{p} \left[\sum_{i=1}^{m} (w_j x_i - y_i)_2^2 \right] = \sum_{j=1}^{p} \left[(X w_j - y)^T (X w - y) \right]$$

We see that the classic regression problem, the one inside the bracket is also iterated p times. The p iterations of the classic regression problem.

Iterating over the regression function p times fails to account for correlated methods within functions. For example sunny weather and low humidity weather are correlated which dont get consided when ω is computed individually.

c) W is a matrix $\in R^{d \times p}$.

Since W has rank of R, its reduced row echelon form has R rows and other rows dont represent the matrix W.

$$\cancel{W \times x = R^{d \times p} \times x^{d} =}$$

$$x \times W = R^{d} \times R^{d \times p} = R^{p}$$

But since W has R significant rows in row echelon form. Hence $x \times W$ is represented in R rows also. Hence $(p - R)$ is something not significant to the representation i.e. noise.

d) [Method 1]

Source: 'Low Rank Sparce Subspace Representation for Robust Regression' by Zhang, Shi, Cheng & Gao

$$\min_{T} \| Y - T\hat{D} \|_F^2$$

$$X = D + E$$

$$\hat{D} = [D ; 1^T]$$

E = noise

D = data embedded in low rank subspaces.

$$\min_{T,D,E.} \frac{\eta}{2} \| W (Y - T\hat{D} \|_F^2 + \text{rank}(D) + \lambda \|E\|_0$$

The second term D is a low-dimensional subspace constraint. η & λ are scalars. $T \in R^{d_x \times (d_x+1)}$

Intuitive approach

d) This can also be solved using an intuitive method.

Method II

$W \in R^{d \times p}$ with rank R

$A \in R^{d \times R}$

$B \in R^{R \times p}$ such that

$W = A \cdot B$

$J(w) = \frac{1}{2} \| Xw - y \|_2^2$

$\nabla J(w) = \nabla_w \left(w^T X^T X w - y^T X w - w^T X^T y + y^T y \right)$

Substituting $W = A \cdot B$

The partial derivation of w now accounts ~~of~~ for A & B which are reduced rank matrices. The reduced rank matrices after partial derivation end up ~~bein~~ costing much less computation

5.

a) Multiplying a scalar to a kernel retains the symetric and positive semi definite properties of the kernel as long as the scalar is positive.

We also know

$aK_1(x, z) + bK_2(x, z)$ are still $R^n \times R^n \rightarrow R$ as they are combinations of kernels.

Hence using Mercers Rule & the definition of the Kernel $aK_1(x, z) + bK_2(x, z)$ is a positive semidefinite symetric matrix making it a Kernel.

b) $K(x, z) = K_1(x, z)\, K_2(x, z)$

We know multiplication of two positive semidefinite matrices gives a positive semidefinite matric. Also since K_1 & K_2 are symetrical, the product of them would also be symetric as they are bound by dot product

$K(x, z) = \phi_1(x)\, \phi_1(z)\, \phi_2(x)\, \phi_2(z)$

They are associative.

We can also see them forming the definition of the matrix.

$$K(x,z) = \phi_1(x)\phi_2(x)\ \phi_1(z)\phi_2(z)$$
$$= \phi'(x)\ \phi'(z)$$

The feature spaces can be defined again.

Hence the matrix K is a kernel.

c) $K(x,z) = f(x)\ f(z)$

We can see this to fit in the definition of the kernel.

$$f(x) :\Rightarrow R^n \rightarrow R$$
$$f(z) :\Rightarrow R^n \rightarrow R$$

$$f(x)\ f(z) :\Rightarrow R^n \times R^n \rightarrow R$$

The function of x & z can be defined as a feature space. This ~~how~~ is the categorical definition of a kernel.

Hence K is a kernel matrix.