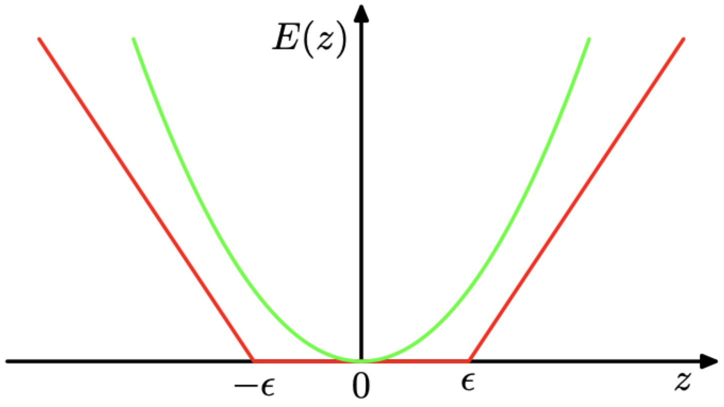
ε-SVR and RHSVR Report

-Abhinava Sikdar

ε – Support Vector Regression

In simple linear regression, we minimize the MSE and a regularization term. However to obtain a sparse solution, the MSE function is replaced by an ε-insensitive error function, which gives zero error if the absolute difference between the prediction y(x) and the target t is less than ε where ε>0. Such an error function can be realized as:

$$E_\epsilon(y(\mathbf{x})-t) = \begin{cases} 0, & \text{if } |y(\mathbf{x})-t| < \epsilon; \\ |y(\mathbf{x})-t|-\epsilon, & \text{otherwise} \end{cases}$$



We therefore minimize:

$$C\sum_{n=1}^{N} E_\epsilon(y(\mathbf{x}_n)-t_n) + \frac{1}{2}\|\mathbf{w}\|^2$$

Where y=w.T*ϕ(x) and C is the inverse regularization parameter.
We introduce two sets of slack variables allowing points to lie outside the tube provided the slack variables are nonzero and the corresponding conditions are:

$$\begin{aligned} t_n &\leqslant y(\mathbf{x}_n)+\epsilon+\xi_n \\ t_n &\geqslant y(\mathbf{x}_n)-\epsilon-\widehat{\xi}_n. \end{aligned}$$

The error function can then be written as:

$$C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|^2$$

Hence we can now write the Lagrangian from the objective function and the constraints as:

$$\begin{aligned} L = {} & C\sum_{n=1}^{N}(\xi_n+\widehat{\xi}_n) + \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N}(\mu_n\xi_n+\widehat{\mu}_n\widehat{\xi}_n) \\ & -\sum_{n=1}^{N} a_n(\epsilon+\xi_n+y_n-t_n) - \sum_{n=1}^{N}\widehat{a}_n(\epsilon+\widehat{\xi}_n-y_n+t_n). \end{aligned}$$

Substituting for y(x) and using the following 1st KKT condition:

$$\begin{aligned}
\frac{\partial L}{\partial \mathbf{w}} &= 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{n=1}^{N} (a_n - \widehat{a}_n) \boldsymbol{\phi}(\mathbf{x}_n) \\
\frac{\partial L}{\partial b} &= 0 \quad \Rightarrow \quad \sum_{n=1}^{N} (a_n - \widehat{a}_n) = 0 \\
\frac{\partial L}{\partial \xi_n} &= 0 \quad \Rightarrow \quad a_n + \mu_n = C \\
\frac{\partial L}{\partial \widehat{\xi}_n} &= 0 \quad \Rightarrow \quad \widehat{a}_n + \widehat{\mu}_n = C.
\end{aligned}$$

And substituting them back into the Lagrangian gives us the Wolfe dual as maximizing:

$$\begin{aligned}
\widetilde{L}(\mathbf{a}, \widehat{\mathbf{a}}) &= -\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} (a_n - \widehat{a}_n)(a_m - \widehat{a}_m) k(\mathbf{x}_n, \mathbf{x}_m) \\
&\quad - \epsilon \sum_{n=1}^{N} (a_n + \widehat{a}_n) + \sum_{n=1}^{N} (a_n - \widehat{a}_n) t_n
\end{aligned}$$

Subject To:

$$\sum_{n=1}^{N} (a_n - \widehat{a}_n) = 0$$

$$\begin{aligned}
0 \leqslant a_n \leqslant C \\
0 \leqslant \widehat{a}_n \leqslant C
\end{aligned}$$

This gives us:

$$y(\mathbf{x}) = \sum_{n=1}^{N} (a_n - \widehat{a}_n) k(\mathbf{x}, \mathbf{x}_n) + b$$

The parameter **b** can be found by considering a point for which 0<a(n)<C. Such a point must satisfy

$$\epsilon + y_n - t_n = 0.$$

Hence, we can obtain b as:

$$\begin{aligned}
b &= t_n - \epsilon - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \\
&= t_n - \epsilon - \sum_{m=1}^{N} (a_m - \widehat{a}_m) k(\mathbf{x}_n, \mathbf{x}_m)
\end{aligned}$$

Since b may not be unique, we can take an average over it.

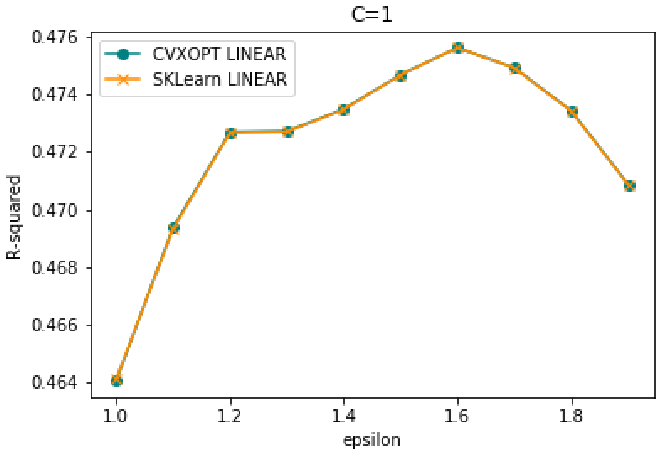
Data Set Used:

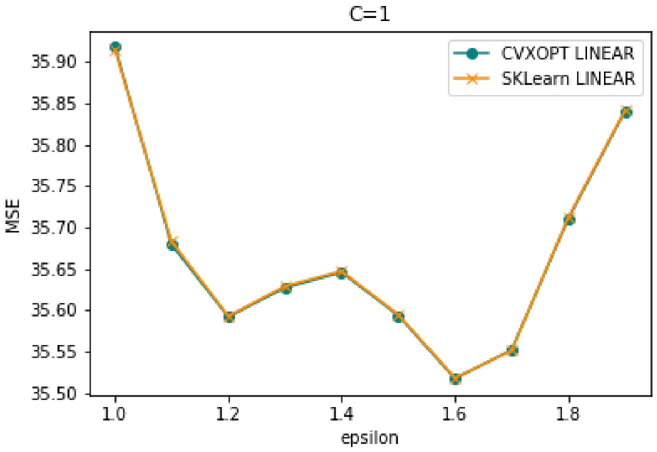
As given, I used the Boston Housing Price Dataset which consists of 506 datapoints, each consisting of 13 features.

Results:

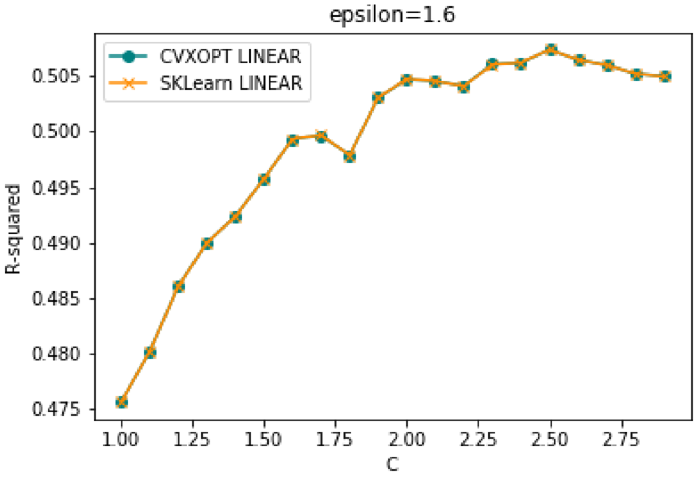
Note: All these graphs were plotted after taking the average of the k values of MSE and Rsquared obtained through k-fold cross validation of the dataset where k=5.

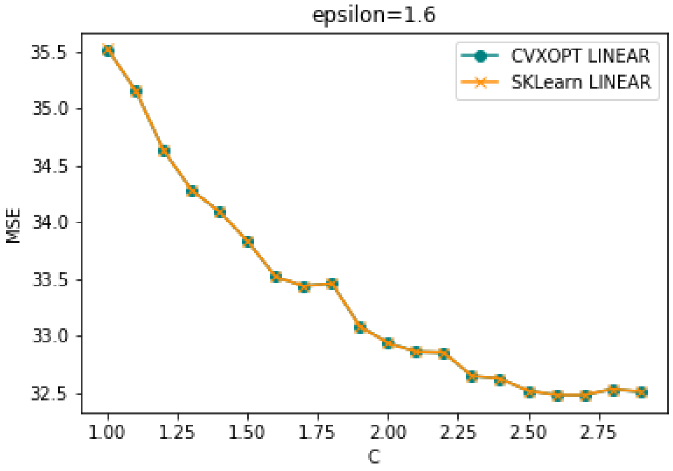
1a) LINEAR KERNEL- Varying epsilon:



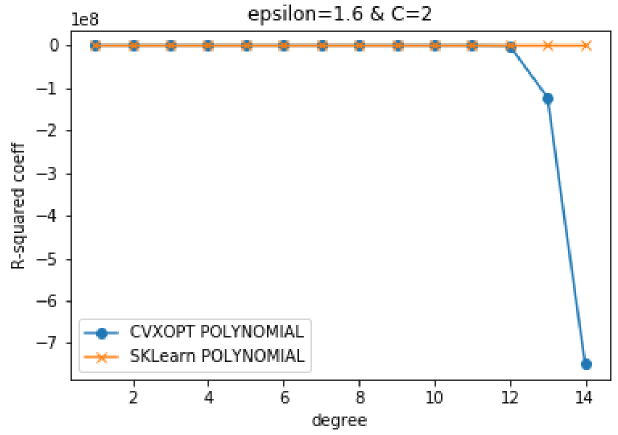


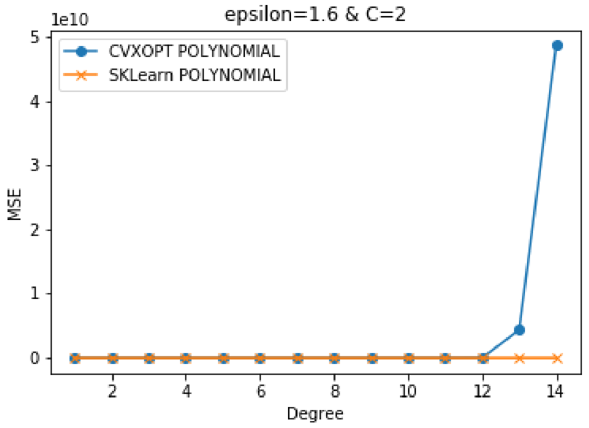
1b) LINEAR KERNEL- Varying C:





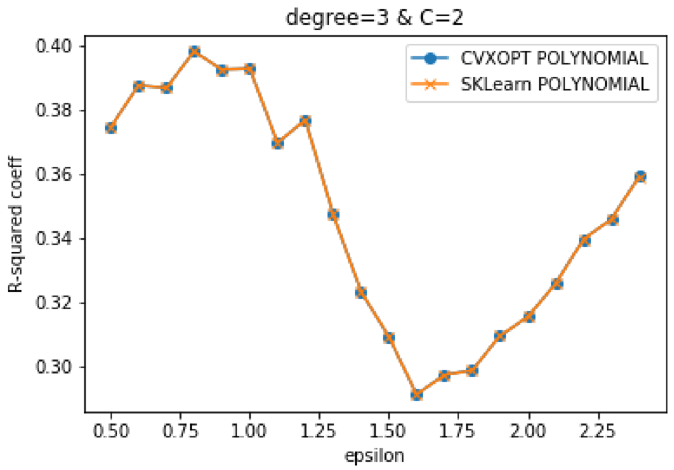
2a) POLYNOMIAL KERNEL- Varying Degree:

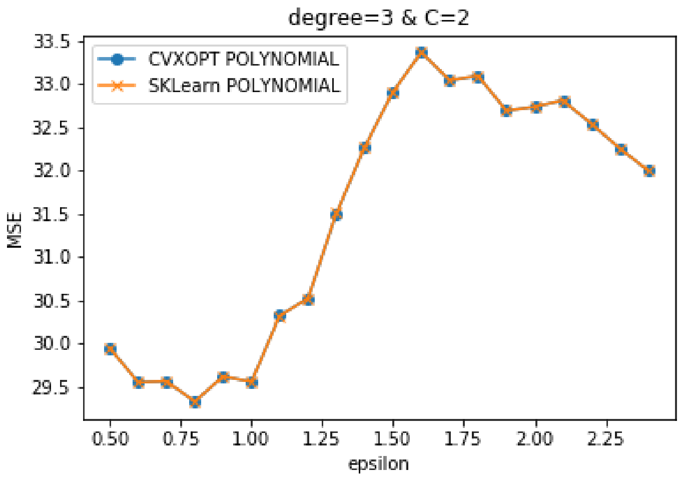




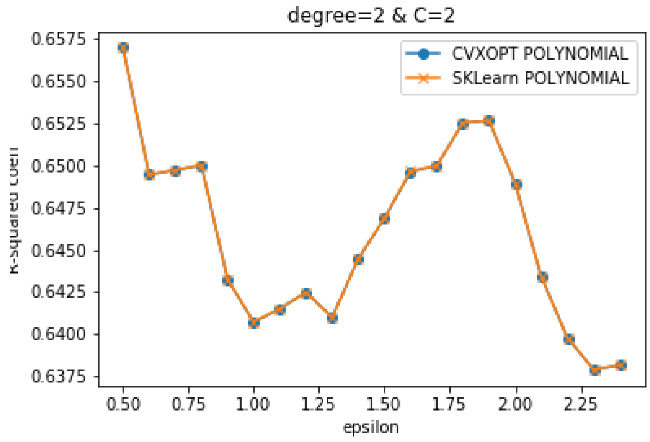
Note: CVXOPT is known to perform poorly for higher degree polynomial kernels

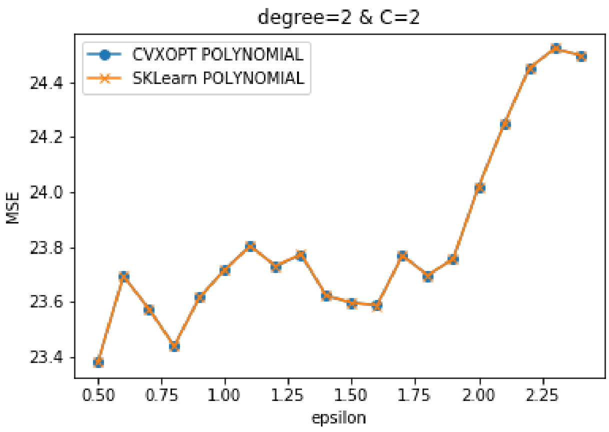
2b) POLYNOMIAL KERNEL- Varying epsilon with degree =3:



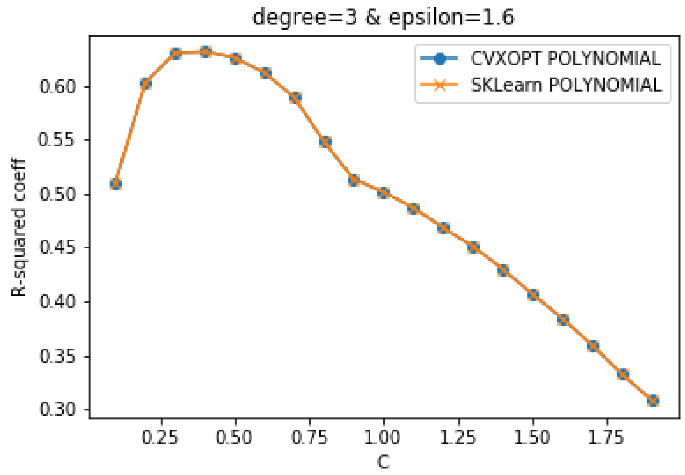


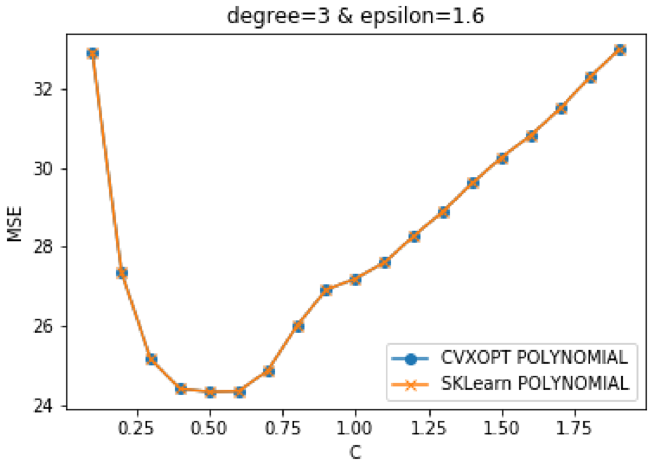
2c) POLYNOMIAL KERNEL- Varying epsilon with degree =2:



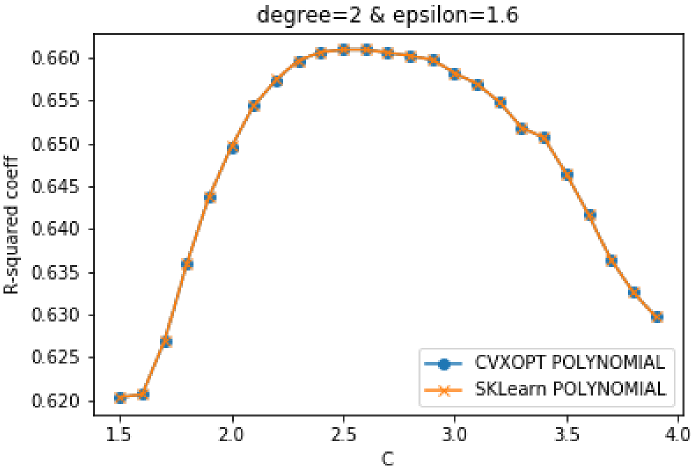


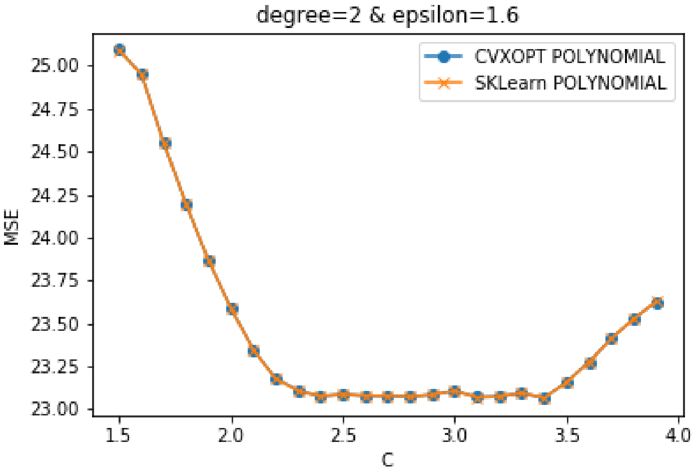
2d) POLYNOMIAL KERNEL – Varying C with degree=3:



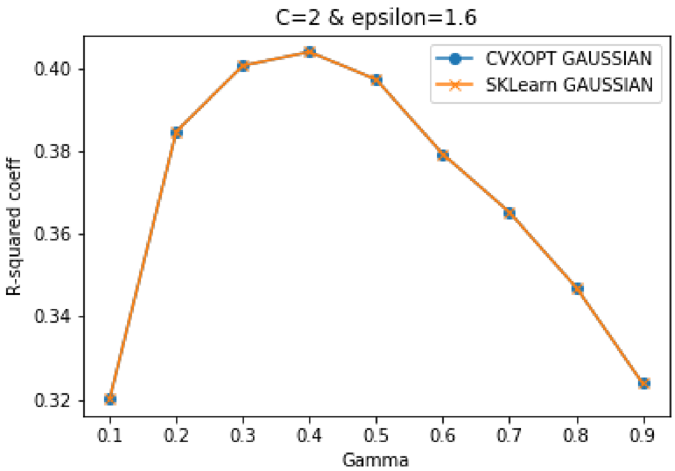


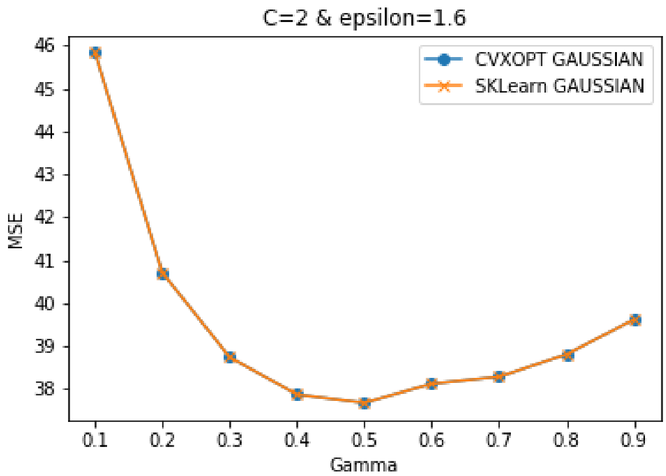
2e) POLYNOMIAL KERNEL – Varying C with degree = 2:



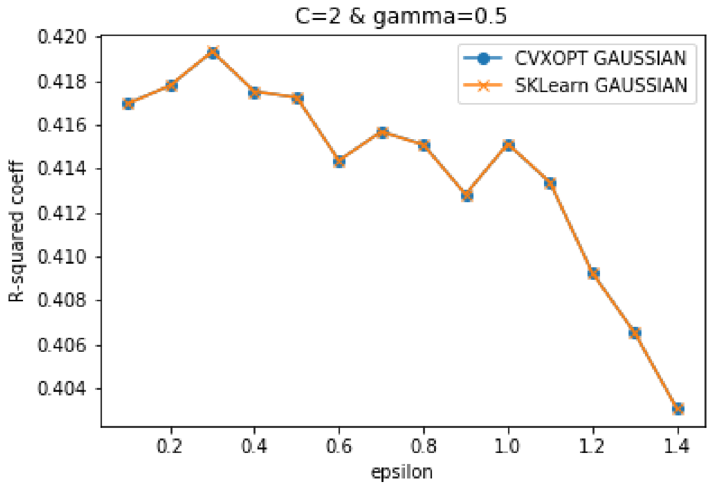


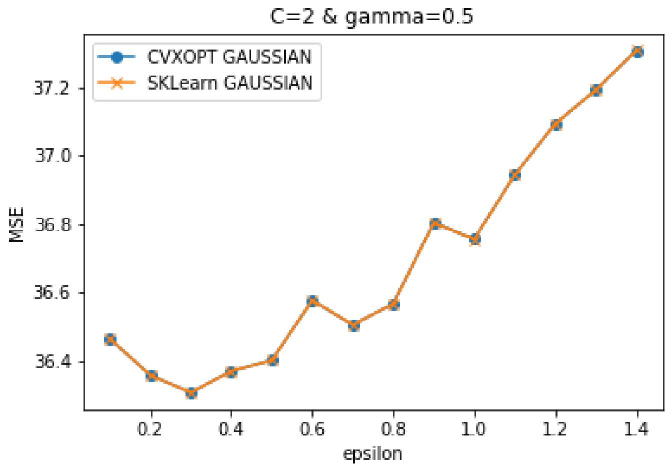
3a) GAUSSIAN KERNEL – Varying gamma:



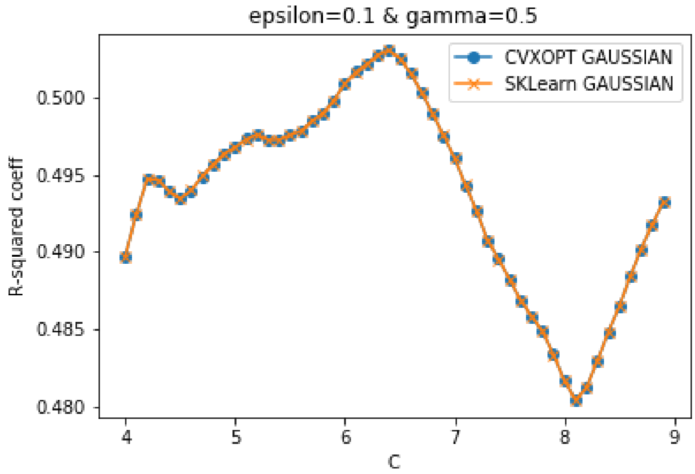


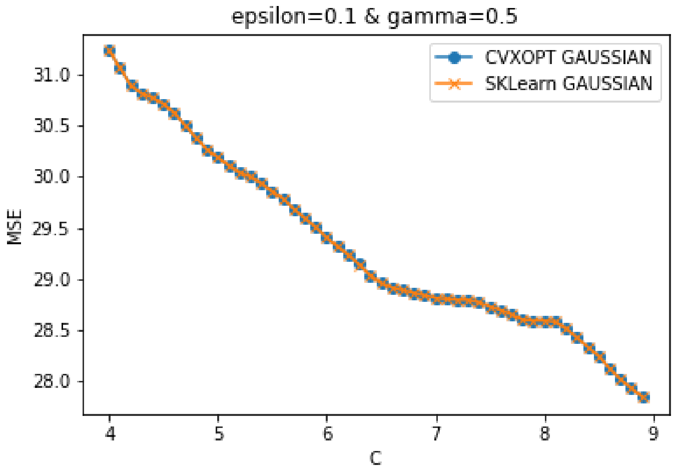
3b) GAUSSIAN KERNEL – Varying epsilon



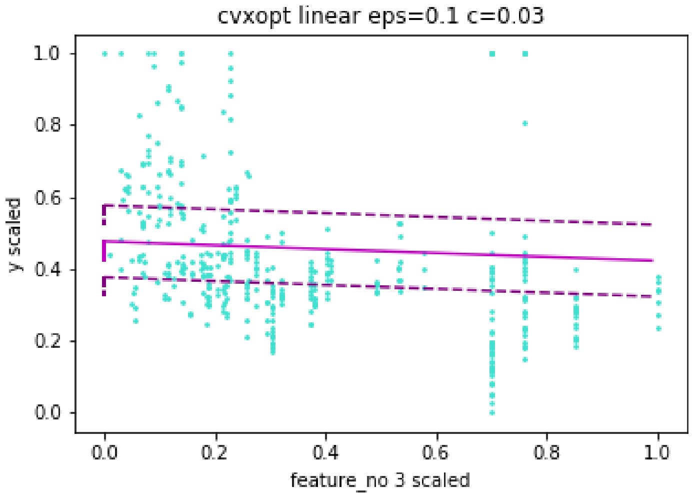


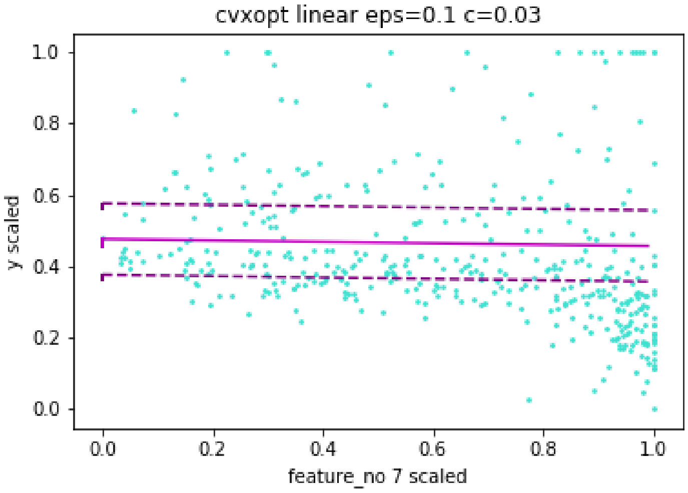
3c) GAUSSIAN KERNEL – Varying C

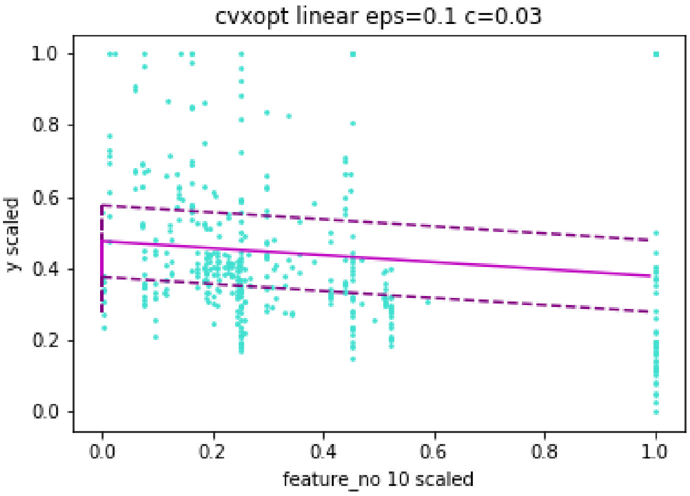


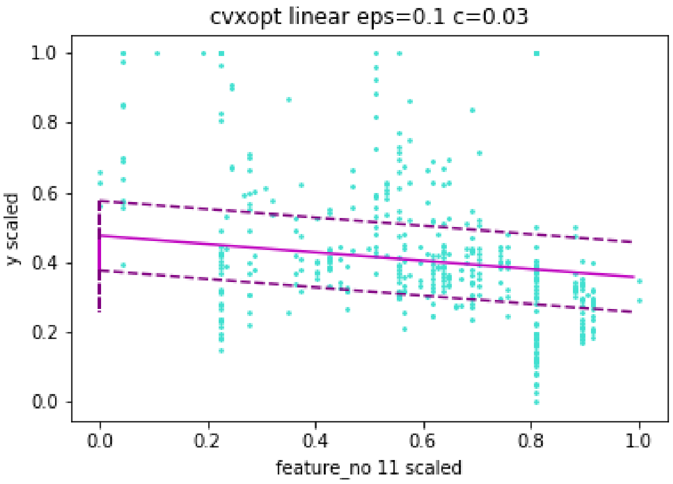


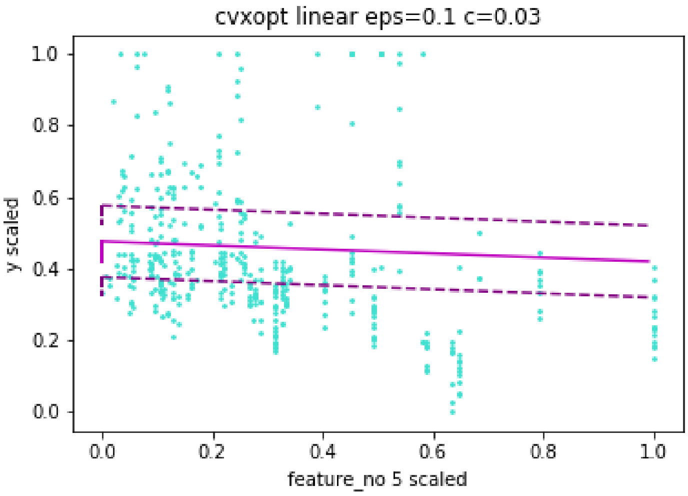
4) Feature-wise, projection of REGRESSION PLANE for LINEAR KERNEL in PRIMAL SPACE:

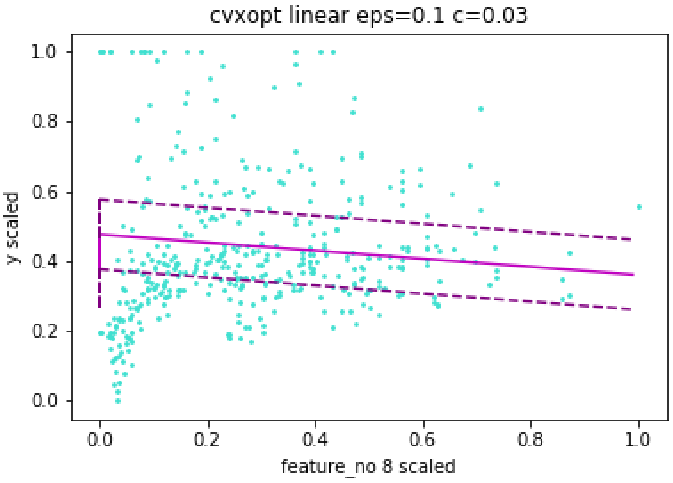












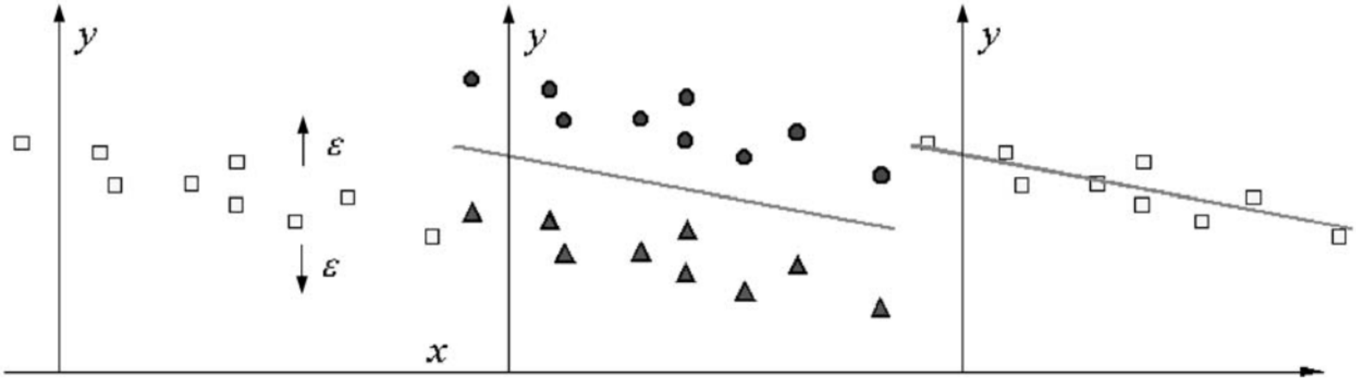
Conclusions:

- **Since SKLearn SVR is also an epsilon-SVR, as one would expect, their performance is identical in terms of MSE loss and Rsquare statistic.**
- **The only exception Is the case of polynomial kernel with high degrees. This is because CVXOPT is known to perform poorly on these.**
- **Overfitting was observed which caused the 2nd and 3rd folds to have higher MSE and lower Rsquare value as compared to the 1st and 5th fold. This can be rectified by shuffling the data before splitting for k fold.**
- **Overall, the lowest MSE and highest Rsquare value was observed for polynomial kernel of degree 2 and 3**
- **The feature set was scaled to 0-1. This is essential for SVMs/SVRs**

RH– Support Vector Regression

In RHSVR, proposed by Bi & Bennet, one converts the problem of regression into that of classification in a dimension which excees that of the original problem by 1.

In a classic SVM for a linearly inseparable classification problem, the primal problem finds the separating plane while maximizing the soft margin between the two classes while the equivalent dual problem computes the closest point in the reduced convex hulls of the two separate classes.



RH-SVM converts the problem of regression into that of classification by introducing another feature to all the data points. This data point is essentially the target variable 'y' of the data point perturbed in both the directions by a factor of epsilon. This leads to the creation of twice the number of original data points. Those having the new positively perturbed feature are said to be in one class while those negatively perturbed are said to be in the other class. Hence, we can perform a classification on these data points to get a maximum soft margin plane using our classic SVM and use this particular plane for regressing on test data.

It is important to note that we converted a regression problem in say d dimensional feature space to a classification problem in d+1 dimensional feature space.

Another salient feature is the removal of the parameter C used in epsilon and mu SVRs, for which the geometric role or interpretation is not known.

Outliers are handled in the dual space by using reduced convex hulls. They limit the influence of any point by reducing the upper bound on the multiplier for each point to D<1.

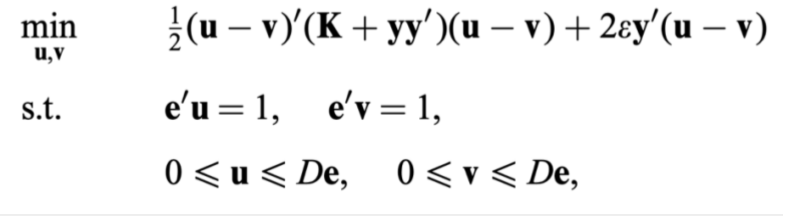
Since RH-SVR computes the closest point in the reduced convex hulls of the shifted data points z(i)=(x(i) y(i)) along the y dimension respectively up and down by epsilon, the problem is as:

$$\begin{aligned} \min_{\mathbf{u},\mathbf{v}} \quad & \frac{1}{2}\left\| \begin{pmatrix} \mathbf{X}' \\ (\mathbf{y}+\varepsilon\mathbf{e})' \end{pmatrix} \mathbf{u} - \begin{pmatrix} \mathbf{X}' \\ (\mathbf{y}-\varepsilon\mathbf{e})' \end{pmatrix} \mathbf{v} \right\|^2 \\ \text{s.t.} \quad & \mathbf{e}'\mathbf{u} = 1, \quad \mathbf{e}'\mathbf{v} = 1, \\ & 0 \leqslant \mathbf{u} \leqslant D\mathbf{e}, \quad 0 \leqslant \mathbf{v} \leqslant D\mathbf{e}. \end{aligned}$$

Where e is column vector of all 1's.

Note: If D < 1, then we define a the reduced convex hulls in the dual space for our problem. If D >= 1, it becomes equivalent to defining the complete convex hulls because if the summation of nonnegative coefficients equals 1, each of the coefficients cannot be larger than 1. The parameter D hence determines whether or not the convex hulls are reduced and the extent of the reduction.

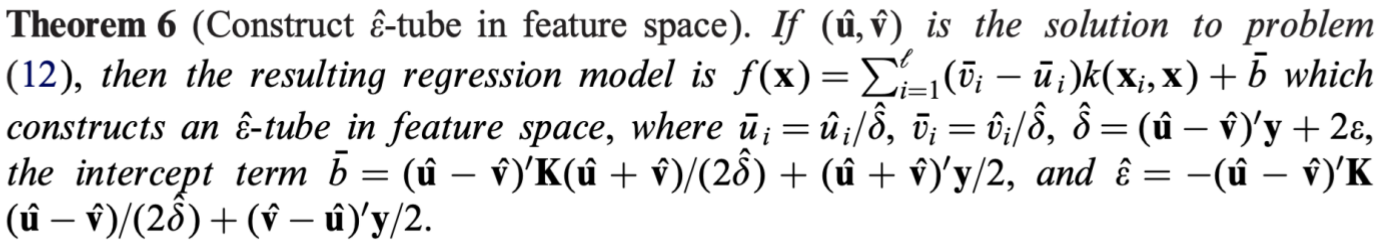
The primal of the above, after kernelizing can be written as:



$$\begin{aligned} \min_{\mathbf{u},\mathbf{v}} \quad & \tfrac{1}{2}(\mathbf{u}-\mathbf{v})'(\mathbf{K}+\mathbf{y}\mathbf{y}')(\mathbf{u}-\mathbf{v})+2\varepsilon\mathbf{y}'(\mathbf{u}-\mathbf{v}) \\ \text{s.t.} \quad & \mathbf{e}'\mathbf{u}=1, \quad \mathbf{e}'\mathbf{v}=1, \\ & 0 \leqslant \mathbf{u} \leqslant D\mathbf{e}, \quad 0 \leqslant \mathbf{v} \leqslant D\mathbf{e}, \end{aligned}$$

We use CVXOPT to find u & v.

Then, using Theorem 6 of the paper by Bi & Bennet:



Theorem 6 (Construct $\hat{\varepsilon}$-tube in feature space). *If* $(\hat{\mathbf{u}},\hat{\mathbf{v}})$ *is the solution to problem* (12), *then the resulting regression model is* $f(\mathbf{x})=\sum_{i=1}^{\ell}(\bar{v}_i-\bar{u}_i)k(\mathbf{x}_i,\mathbf{x})+\bar{b}$ *which constructs an* $\hat{\varepsilon}$*-tube in feature space, where* $\bar{u}_i=\hat{u}_i/\hat{\delta}$, $\bar{v}_i=\hat{v}_i/\hat{\delta}$, $\hat{\delta}=(\hat{\mathbf{u}}-\hat{\mathbf{v}})'\mathbf{y}+2\varepsilon$, *the intercept term* $\bar{b}=(\hat{\mathbf{u}}-\hat{\mathbf{v}})'\mathbf{K}(\hat{\mathbf{u}}+\hat{\mathbf{v}})/(2\hat{\delta})+(\hat{\mathbf{u}}+\hat{\mathbf{v}})'\mathbf{y}/2$, *and* $\hat{\varepsilon}=-(\hat{\mathbf{u}}-\hat{\mathbf{v}})'\mathbf{K}(\hat{\mathbf{u}}-\hat{\mathbf{v}})/(2\hat{\delta})+(\hat{\mathbf{v}}-\hat{\mathbf{u}})'\mathbf{y}/2$.

we regress on the test data.

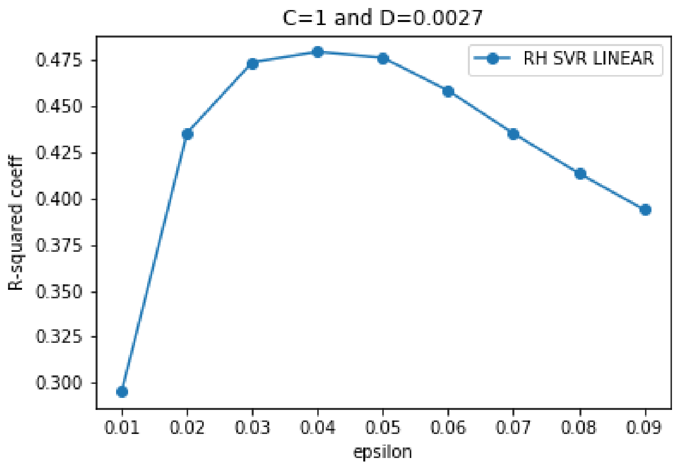
Data Set Used:

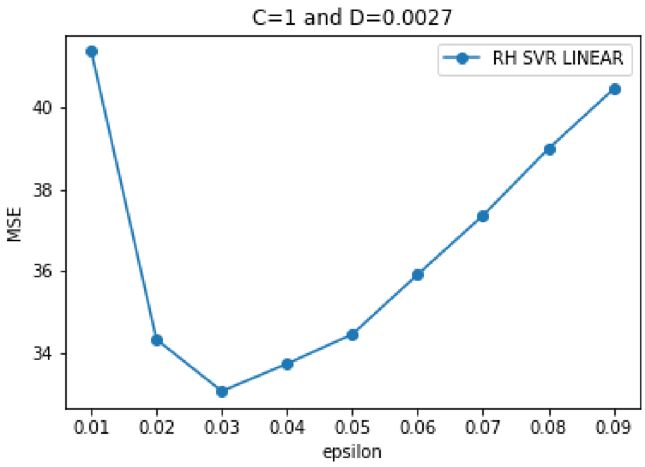
As given, I used the Boston Housing Price Dataset which consists of 506 datapoints, each consisting of 13 features.

Results:

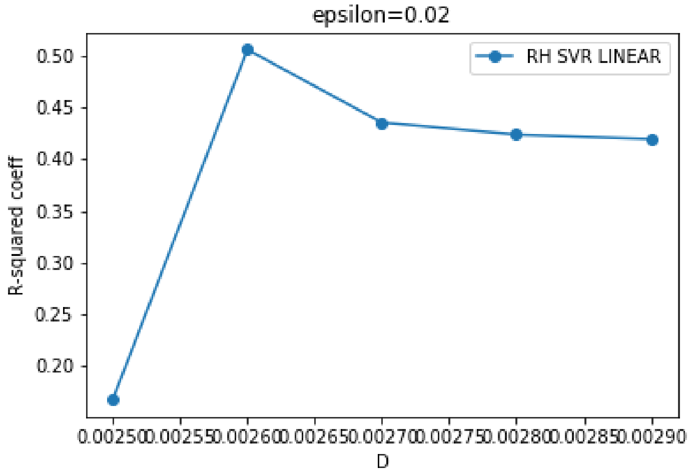
Note: All these graphs were plotted after taking the average of the k values of MSE and Rsquared obtained through k-fold cross validation of the dataset where k=5.

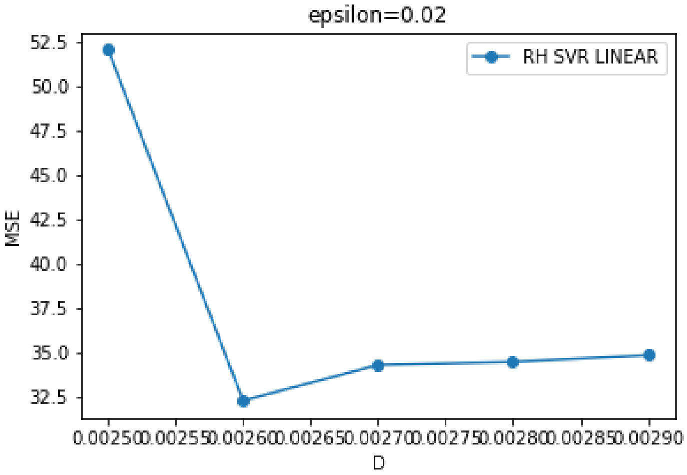
1a) LINEAR KERNEL- Varying epsilon:



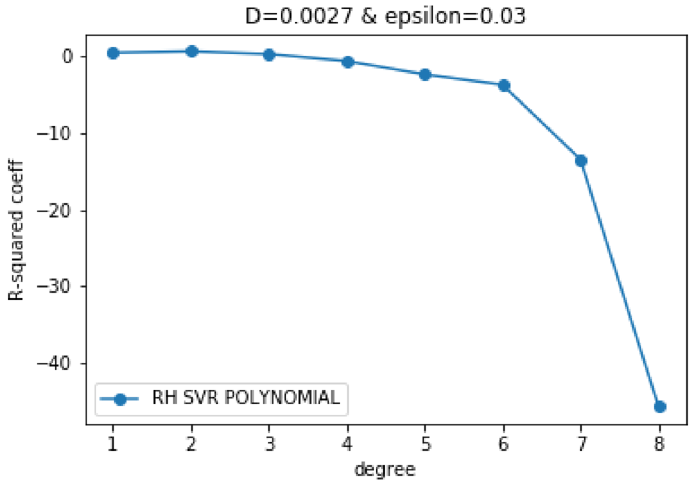


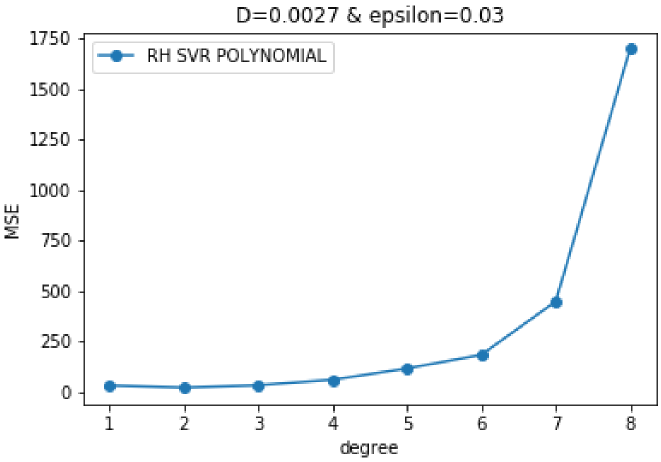
1b) LINEAR KERNEL- Varying D:



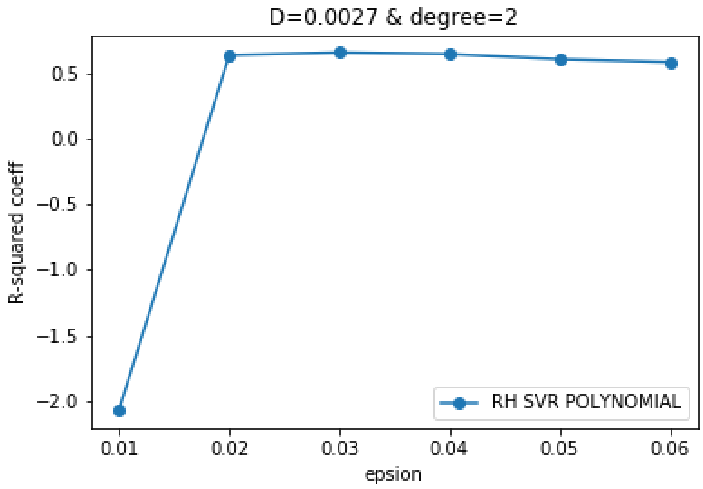


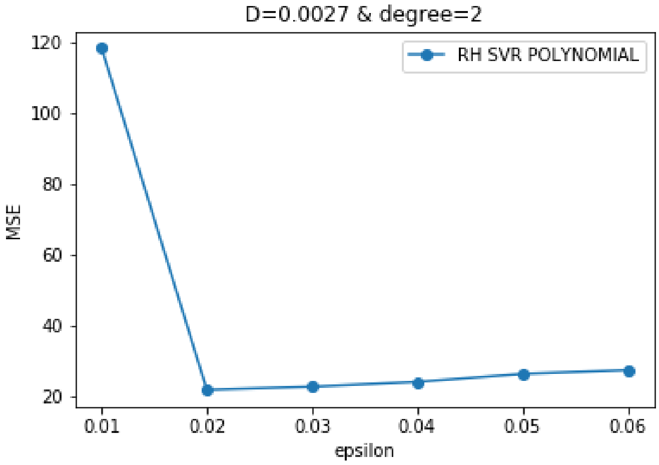
2a) POLYNOMIAL KERNEL- Varying degree:



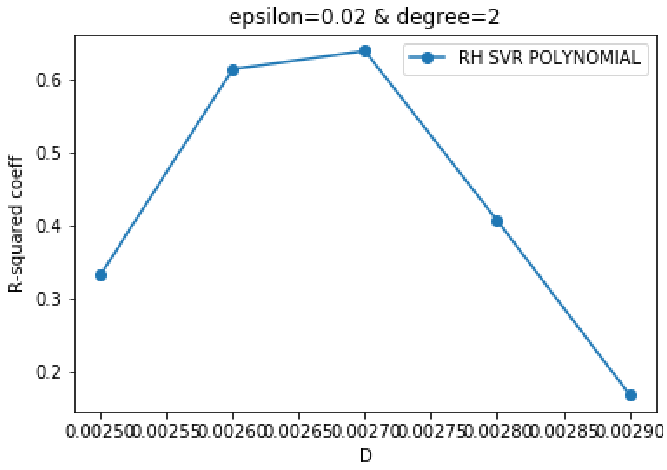


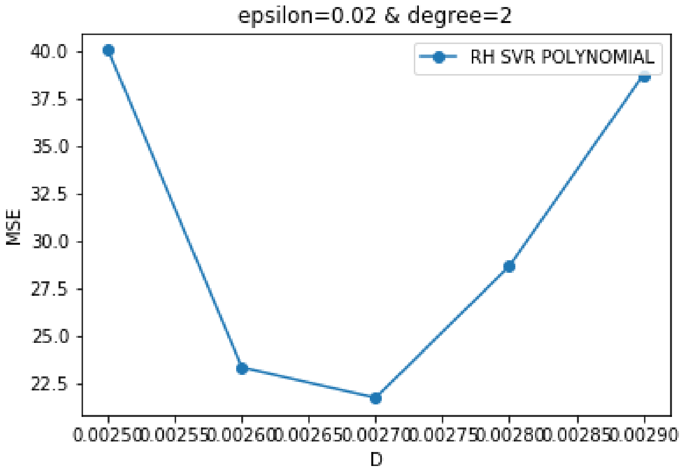
2b) POLYNOMIAL KERNEL- Varying epsilon:



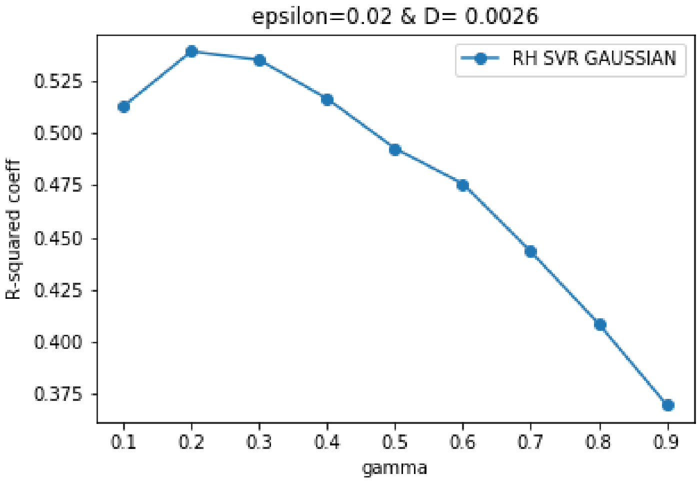


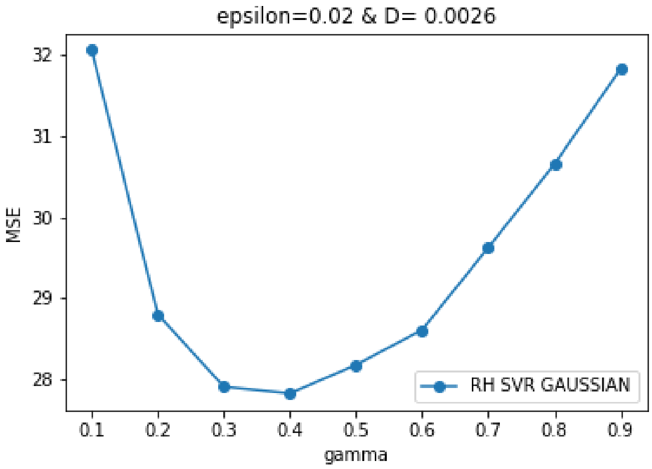
2c) POLYNOMIAL KERNEL- D:



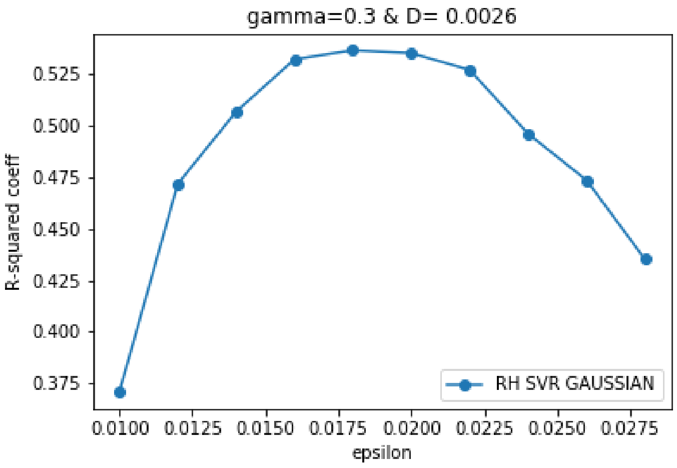


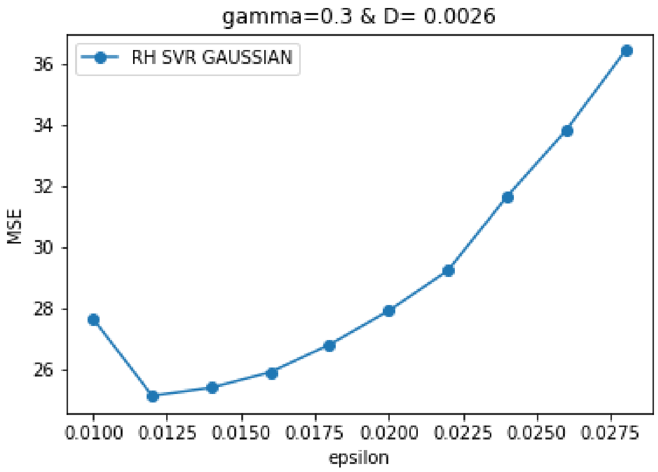
3a) GAUSSIAN KERNEL- Varying gamma:



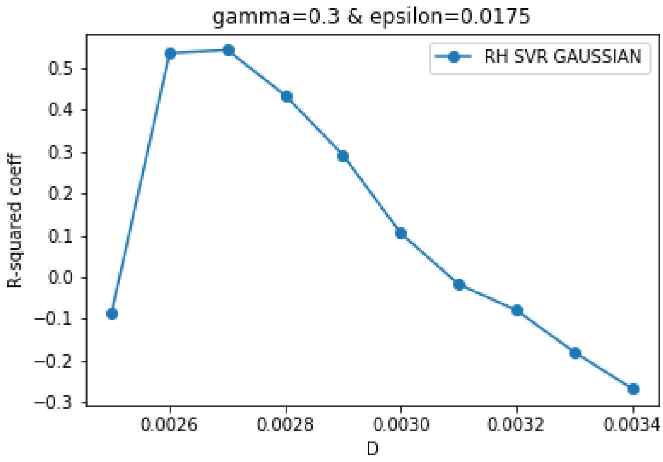


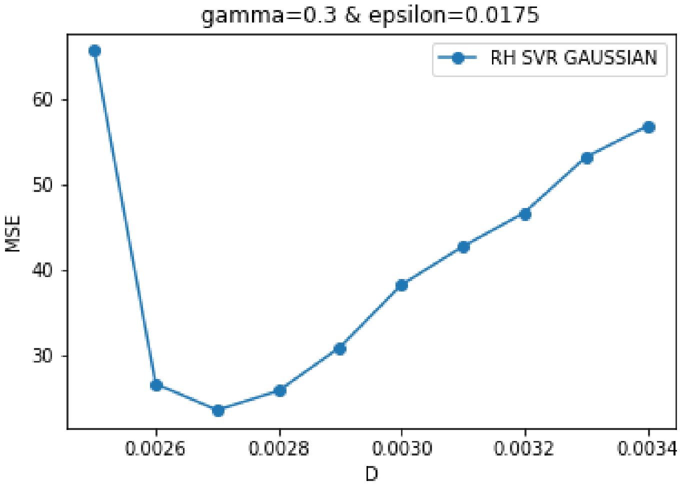
3b) GAUSSIAN KERNEL- Varying epsilon:



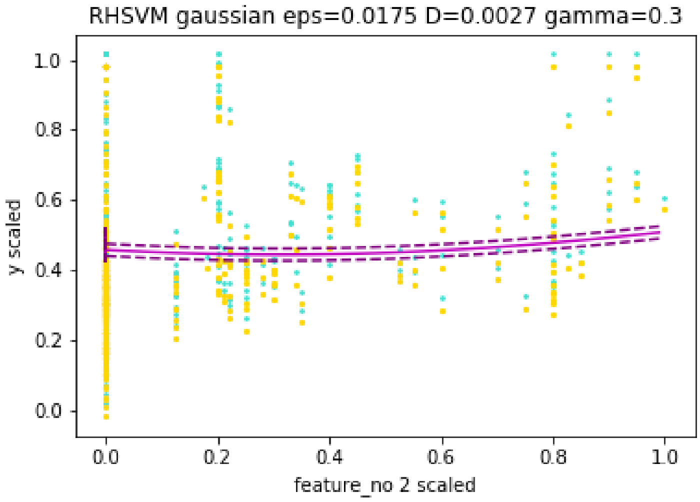


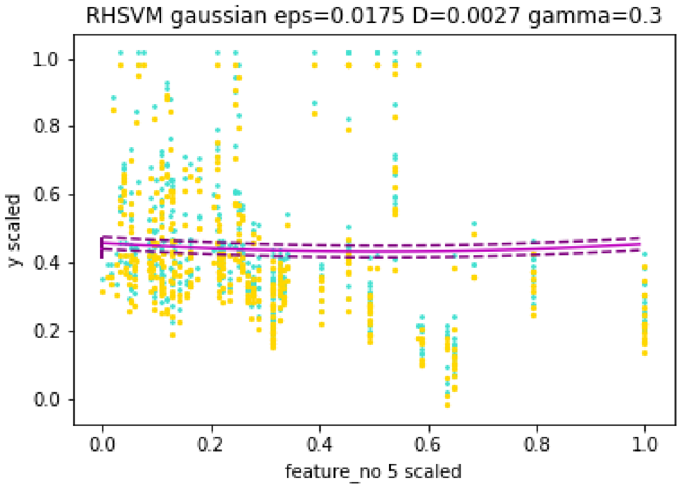
3c) GAUSSIAN KERNEL- Varying D:

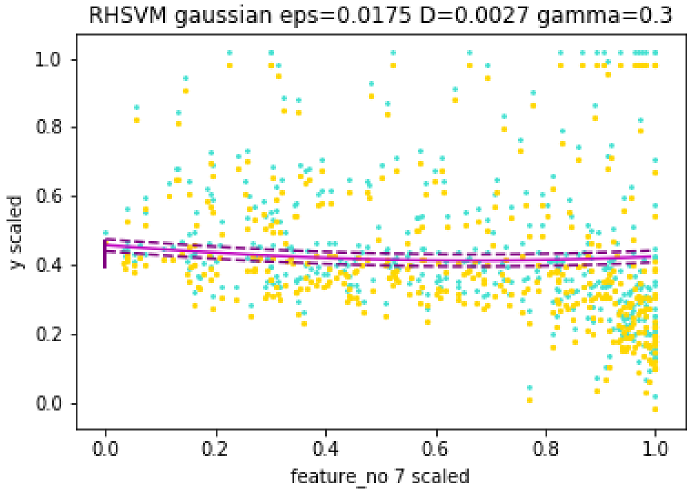


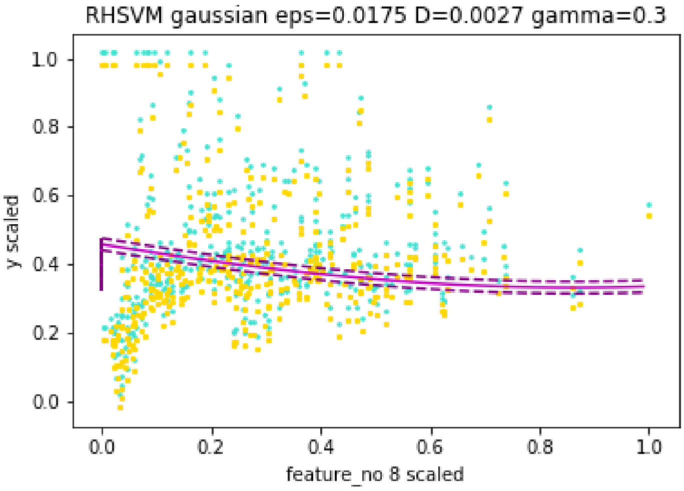


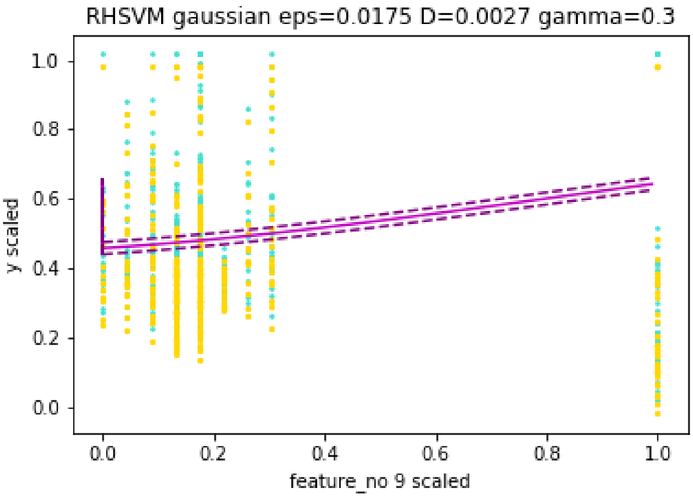
4) Feature-wise, projection of DECISION BOUNDARY for Gaussian Kernel in DUAL SPACE:

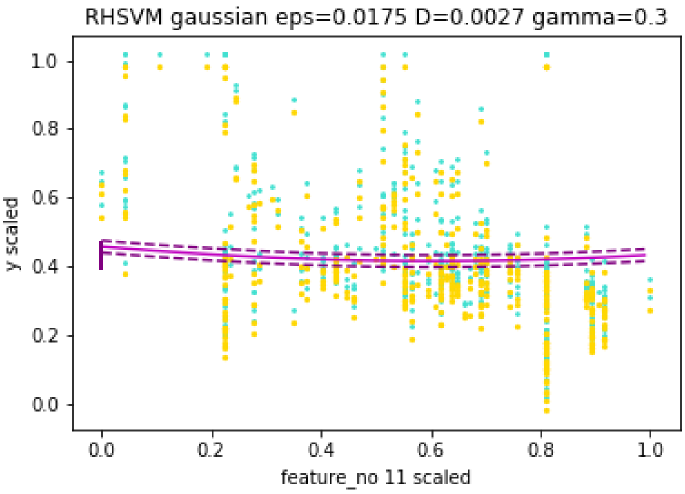












Conclusions:

- **Since the parameter C and D for an epsilon SVR and a RHSVR, respectively are completely different as to their role and geometric interpretation, a comparison based only on MSE and Rsquare can be made. Hence I did not plot the graphs of both together.**
- **Bad performance of polynomial kernel with high degrees was observed. This is probably because CVXOPT is known to perform poorly on these.**
- **Overfitting was observed which caused the 2nd and 3rd folds to have higher MSE and lower Rsquare value as compared to the 1st and 5th fold. This can be rectified by shuffling the data before splitting for k fold.**
- **Overall, the lowest MSE and highest Rsquare value was observed for polynomial kernel of degree 2 and 3**
- **The feature set was scaled to 0-1. This is essential for SVMs/SVRs**
- **MOST IMPORTANT**: While I scaled only the features for the epsilon SVR, I realized that it is extremely important to scale the 'y' values as well in the case of RHSVR. This is because the value of y+-epsilon is essentially a feature for the RHSVR(the d+1 feature!) since the classification is based on it too! This can be the difference between the program running and otherwise!!!

References:
[1] Pattern Recognition & Machine Learning by Christopher M. Bishop
[2] A geometric approach to support vector regression by Jinbo Bi. and Kristin P. Bennet