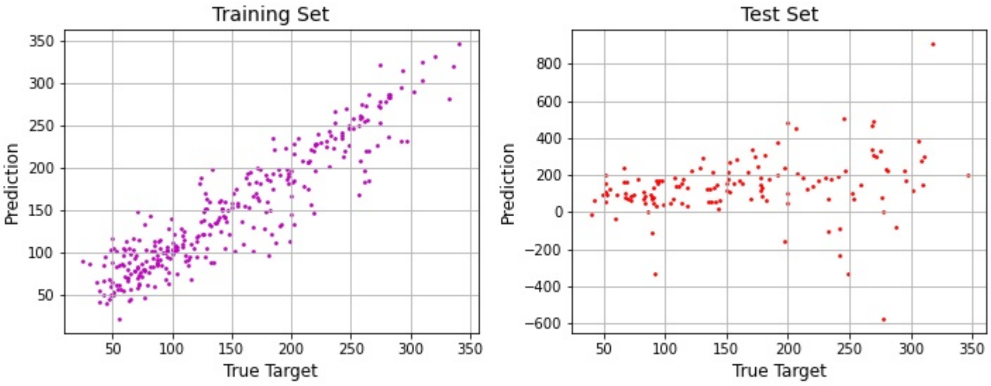
# Foundations of Machine Learning Lab 5 Report

Abhilash Pulickal Scaria
aps1n21@soton.ac.uk
Student Id:33124639

## Observation

1)Diabetes dataset is loaded using the sklearn package. Using the code snippet we can implement a Gaussian RBF model with $\sigma$ taken as distance between two randomly chosen points and basis function location as random points in input space.
2) The width parameter based on the current code can run into situations where the distance becomes zero, this can cause zero division error. We can change the width parameter to average of several pairwise distance to avoid this error. After setting the M basis function location to K-means cluster centres (K=M) and splitiing the data into training(70%) and test data(30%) we can train our model. From figure 1 we can observe the performance of the model on test and train data.

Figure 1: Prediction on test and train data.



3) When we initially introduce basis function and project data from 10 feature space to a 20 feature space we observe an improvement in training data prediction and a slight improvement or almost similar performance in test data prediction .

Figure 2: Number of basis functions = 20.

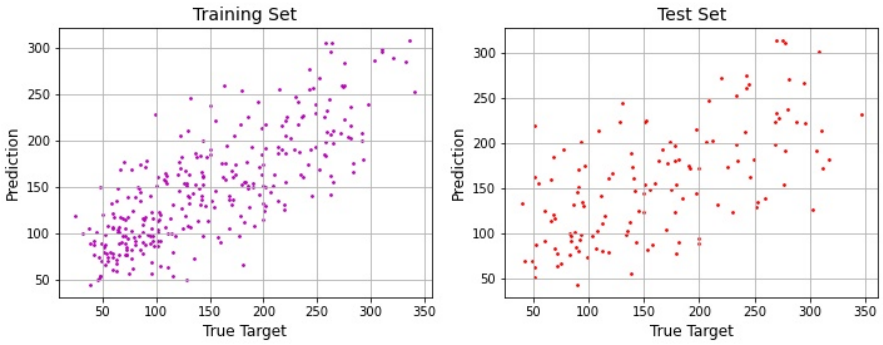
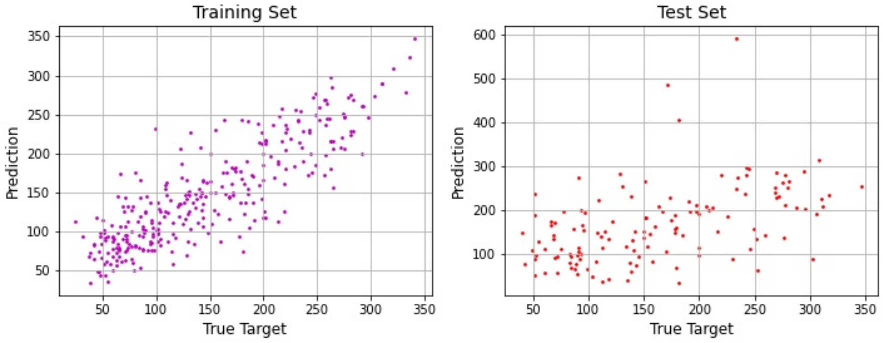


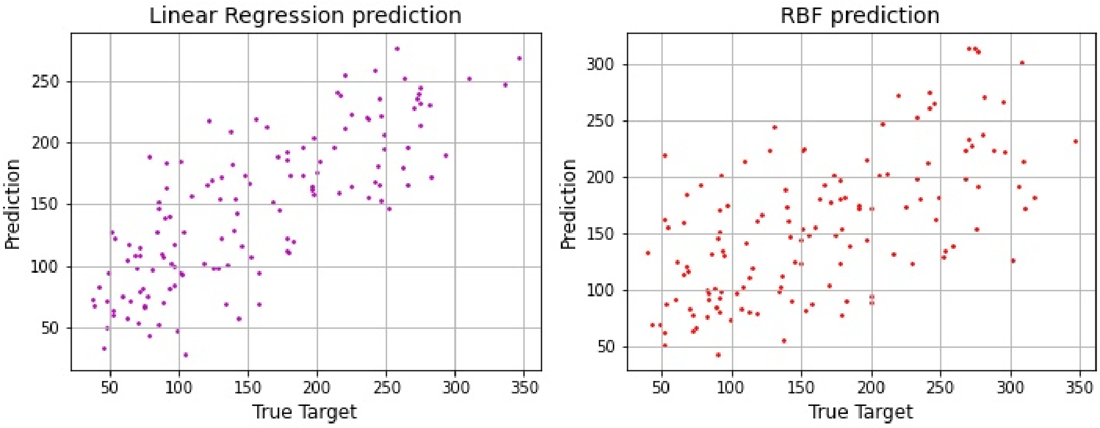
Figure 3: Number of basis functions = 100.



But as we continue increasing the basis function to 100 and then to 200, we can observe the model becoming overtrained. The increase in number of basis functions after a certain point causes the model to overfit and become less generalized. This can be observed in figure 2 and 3.

4) The distributions of prediction of test data for RBF and Linear Regression models is given below.

Figure 4: Prediction on test sets .


Linear Regression prediction
Prediction
True Target
RBF prediction
Prediction
True Target