Machine Learning – ES 654

Spring 2019

IITGN

1Ans.

a) Here is the secret gist link to the code of randomforest: Link

b) Yes, we could still learn the decision tree. By splitting the samples based on the probability values for each class at that depth of the decision tree. Here is the code for the same

c) Using increased parallelised run version with number of trees= 100. The train and test time together adds up to 3.9082seconds, where as parallel run takes around 2.0137seconds.

d) When ran with the given parameters and train size, the accuracy is as follows

Decision tree: 93.333%

Random forest: 97.777%

Hence using random forest does increase the accuracy

e) Using nested cross-validation I figured out that the no.of ideal estimators are 100. Here is the code for the same.

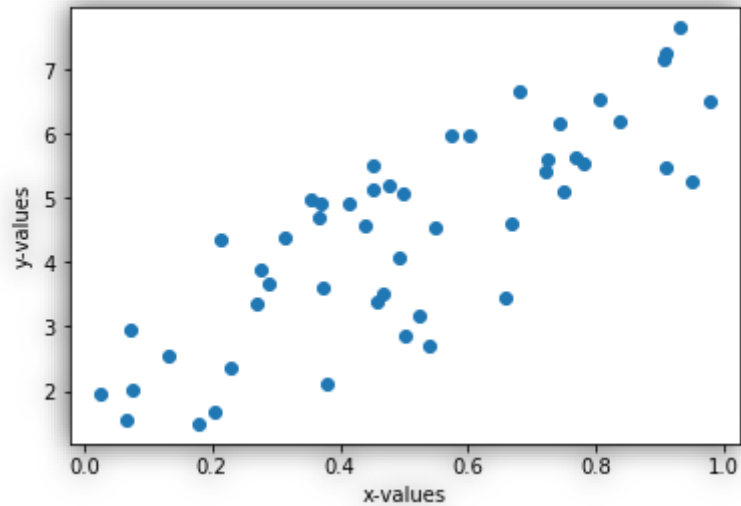**2Ans.** I have submitted my kernel to the kaggle with the Team Name : RA_GM1

   a) Kaggle score using decisiontree regressor is 5.07304
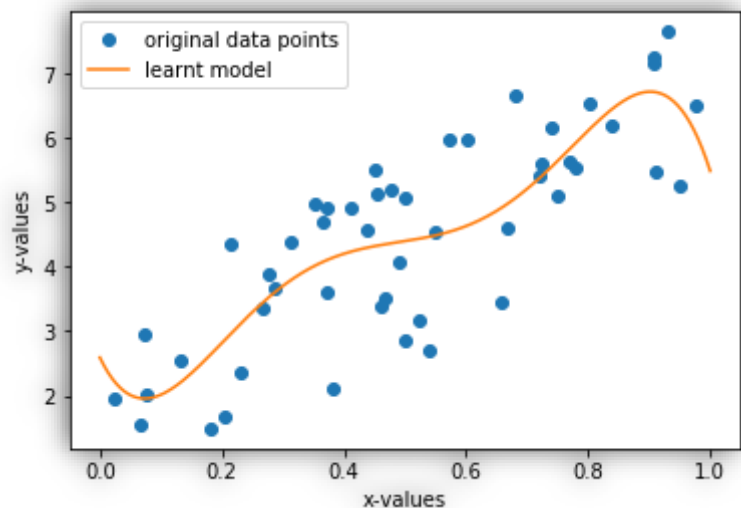   b) Kaggle score using randomforest model is 4.81226

4Ans.

 Part-(a):

Here is the code for generating random dataset of size 50. The plot of the generated dataset is shown below.

Part-(b):

Here is the code for fitting a 5-degree polynomial for the above dataset. The plot of the model is shown below along with the dataset.


Part-(c):

Performing bagging on the above dataset using degree-5 polynomials doesn't decrease the variance. If we 5 more points to the dataset then the model is changing is rapidly in comparison to the first learnt model (using 50 data samples). Plot of the model using bagging. Code for the 5-degree polynomial fit using bagging technique.
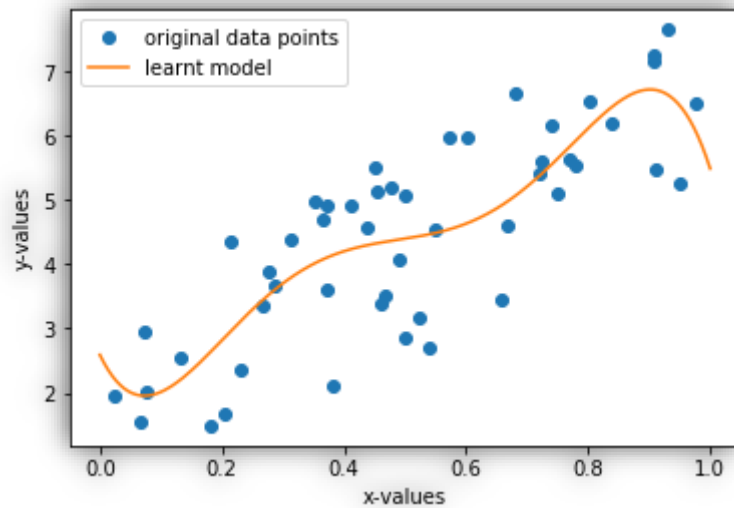
Bagging using 20-degree polynomials doesn't reduce the variance. Plot for the model learnt using degree 20 polynomials. Code for the 20-degree polynomial fit.
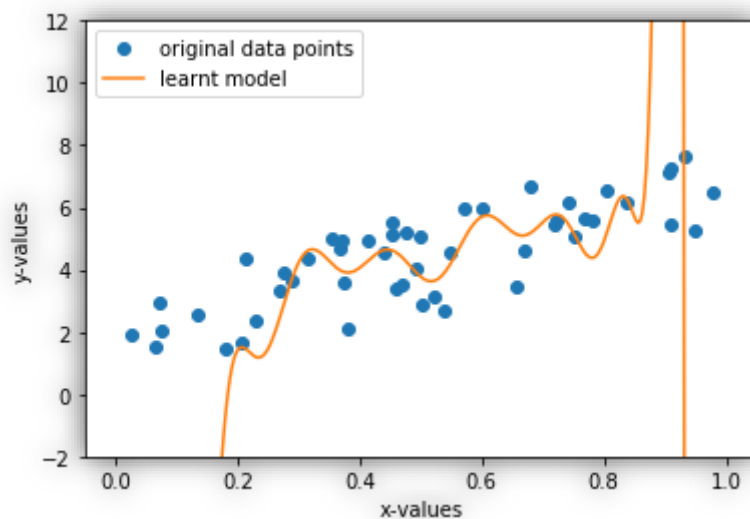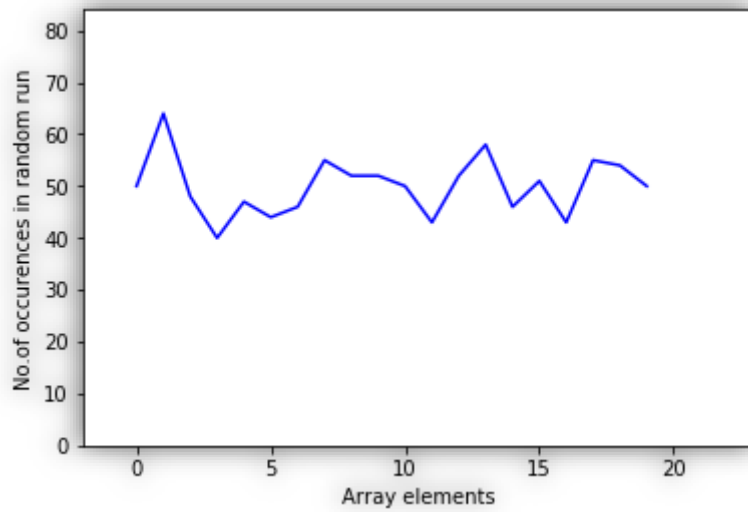

5Ans.

Part-(a):

 Here is the code of the random generator.

Part-(b):

 Here is the code of the plot where I ran the above code 1000 times and plotted the number of occurrences of element.

References:

- Lecture notes
- https://scikit-learn.org/stable/modules/ensemble.html