MACHINE LEARNING HOMEWORK 2

ELİF HANGÜL

a) Why do we normalize f?
We want to see the effect of noise, N and target complexity(Qf) on overfit measure. Since y values depending on f and noise level, we normalize the f to get more compact data points, only varied by noise.

b) How can we obtain g2,g10?
We can think this problem like a linear regression. To generate a higher order equation we can add powers of the original features as new features. This will still considered as a linear model as the coefficients associated with the features are still linear. After that, applying linear regression on the new features will give us the expected curve.

c) How can we compute Eout for a given g10?
Basically we try polynomial regression. With the help of PolynomialFeatures method from sklearn library we can select the degree of polynomial as 10. Then using pipeline we can combine the degree with linear regression to create our pipeline frame. After that using r2 scoring we can calculate the Eout for a given g10.

d) Since trying all the Qf, N and noise level values would take too much time and effort, I pick 6 values from each set to try over. The selected values are:

```
n_samples_array=[20,30,50,80,95,120]
qf_array=[3,18,30,43,88,92]
var_array = [0,0.3,0.5,1.2,1.7,2]
```

In total 216 experiments are run.
During each experiment out of sample error for H2 and H10 are calculated, the difference between them is assigned to overfit_measure variable. Each iteration the minimum and maximum H2 error and H10 error are hold, also the N,Qf and variance values that leads to these minimum and maximum errors are kept. A sum of H2 and H10 errors are calculated. After each experiment is completed, the average H2 error, H10 error, the minimum overfit measure, the maximum overfit measure and the N,Qf,variance values that lead to them are printed.

From the observations we can come up to these conclusions:

Increase in the number of data points(N) decreases the overfitting.
Increase in the variance(noise level) increases the overfitting.
Increase in the target complexity(Qf) increases the overfitting.

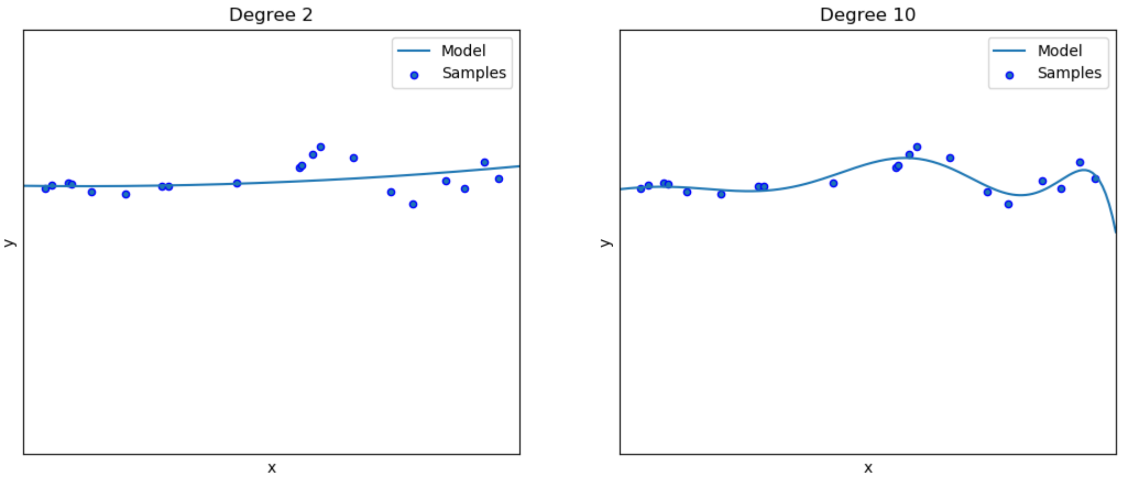
The output is given as:

```
Average out of error for H10  637101.6832350367
Average out of error for H2  1294.8303817816009
```

```
Minimum overfit measure  -121660.69520884981
Maximum overfit measure  66479136.46022499
N,QF,noise values that making overfit measure minimum  30   3   0.3
N,QF,noise values that making overfit measure maximum  30   30   0.3
```

e) Why do we take the average over many experiments?
As we select the data points and training and testing datasets randomly, it is always better to try different experiments to average over. Especially, in our case we want to see if H2 or H10 model fits the data, trying with different parameters such as N size, different noise level and different f(x), then calculate the out of sample errors for each different circumstances for both model and getting the average, will give us a very clear picture on which model to use. If we try not enough experiment then we could get wrong assumptions, such as H10 model fitting better, on the other hand after a satisfactory number of experiments, continuing to do so would just be a waste of time. The balance over number of experiments conducted needs to be found.
To select the acceptable number of experiments, we look into bias-variance trade-off. A high bias results in underfitting while a high variance results in overfitting. As the model complexity increases, the bias decreases and the variance increases.

Notes:
A sample model plot for N=30, Qf=50, noise=0.5 can be given as:


Degree 2
Model
Samples
y
x
Degree 10
Model
Samples
y
x