Data Analytics Course: 18-899

Recitation 2

Felix & Adolphe

Spring 2019

Carnegie Mellon University

Question 7

- In previous questions, we tried to predict the energy consumption using a quadratic model.
- We have multiple features, we would like to investigate if using many variables wouldn’t improve our accuracy
- However, not all the variables can linearly explain the dependent variables
- Need of selecting features using the stepwise approach

Question 7 -- continued

- We select significant feature and use them to predict the energy consumption
- Question: what are the variables were selected
- More significant question: is the multivariate linear regression more accurate than the quadratic model? → R-squared

Question 8

- Reminder of why quadratic

- Simple domain knowledge analysis:
  - Low temperature → high consumption
  - High temperature → high consumption
  - Medium temperature → low temperature

Question 8 -- continued

- Hence assumptions of quadratic relationship
- We increase the number of feature by adding all terms squared
- We have a large number of features, we repeat 7 to select significant features
- Same question: is this multivariate linear model with previous feature variables and squared terms improve on accuracy? → r-squared

Question 9

- Next question
- Is there any relationship between weekdays(Monday, Tuesday, ….) and the energy consumption?
- In other words, being Monday or Tuesday or … does it have any effect on energy consumption
- To our feature variables, we add the weekdays dummy variables
- Is there any improve in accuracy

Question 9 -- continued

- Procedure:
- My dataset have dates:
- Can we obtain weekdays from dates
  - Matlab: weekday → look up it return
  - Python: weekday()
- Now that I have weekdays, I need to create dummy variables

Question 9 -- continued

Hint:

- pandas: pd.get_dummies()
- Matlab:

  - dummyvar function or

  - you can deal with it logically

Question 10

- Question to ask yourself:

  -What is overfitting?

  -which techniques do we use to avoid overfitting?

  - Did we use any of those techniques?

QUESTIONS

GOOD LUCK