Dan Schumacher

1/17/24

Predictive Modeling

Assignment #1, (ISLR CH2 Questions)

1. For each part, indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.
   * The sample size n is extremely large, and the number of predictors p is small.
     + ***Inflexible*** would be better. Generally, more flexible models require more predictors.
   * The number of predictors p is extremely large, and the number of observations n is small.
     + ***Inflexible*** would be better. Using a small dataset with a flexible model is a great way to cause overfitting.
   * The relationship between the predictors and response is highly non-linear.
     + ***Flexible***. Flexible methods, because they don’t start with a given function, can generate shapes that are more unique the less flexible methods.
   * The variance of the error terms, i.e. σ2 = Var(ϵ), is extremely high.
     + ***Inflexible***. High variance means that changing the test data would yield wildly different results. Simpler models are less likely to change with different test data.
2. Explain whether each scenario is a classification or regression problem and indicate whether we are most interested in inference or prediction. Finally, provide n and p.
   * We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry, and the CEO salary. We are interested in understanding which factors affect CEO salary. 2.4 Exercises 53
     + Regression
     + Inference
     + n = 500, p = 3
   * We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.
     + Classification
     + Prediction
     + n = 20, p = 13
   * We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence, we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.
     + Regression
     + Prediction
     + n = ~52 weeks in a year, p = 3
3. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

|  |  |
| --- | --- |
| FLEXIBLE | |
| Pros | Cons |
| Good for fixing Bias.  Typically, better for prediction | Typically need more parameters  Easy to overfit  Likely increase variance |

|  |  |
| --- | --- |
| INFLEXIBLE | |
| Pros | Cons |
| Good for fixing variance.  Typically, better for inference bc easier to understand | Small range of shapes  Can cause Bias |

1. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?
   * Parametric assumes a shape/starting function then plugs in weights.
   * Nonparametric don’t assume a starting place but build a function from scratch.
   * Pro parametric
     1. Easy to understand.
     2. Structure makes them less likely to overfit.
   * Con parametric
     1. Can oversimplify. -> Bias
     2. Can make weaker predictions than non-parametric
2. dfdf