

## Stat 6601 In-Class Final (50 minute)

Open book & open note. Choose **only five questions** from below and answer them.  
(A plain English version is preferred; you can use math formula but it's not a must)

1. What is overfitting?
2. What is sample-splitting? What is cross-validation? Why do we need to use them?
3. What's the benefit of tree based regression method compared to multiple linear regressions?
4. Suppose you transform the response ('Y') variable and apply the linear regression. What's the potential danger of that? How do you know if the transformation is the correct thing to do?
5. Is there such a thing as the robust nonlinear regression? Justify your answer. (If it is possible, what R function can you use to fit it?)
6. In curve fitting, which is more important? Method (LOESS, ksmooth, locpoly, spline, etc) or bandwidth (or span)? Justify your answer.
7. What's the benefit of robust regression? What's the disadvantage?

## **Stat 6601 In-Class Final Solution**

1. Overfitting is a phenomenon of a classification/regression algorithm achieving a very small error in a test dataset (dataset on which the parameters of the algorithm is fitted), mostly from using excessively large number of parameters, and performing very poor on the new dataset (training dataset).
2. Both are methods for evaluating performance of an prediction/classification algorithm. Sample-splitting is dividing the data into test and training samples. The algorithm is fitted on the training sample and it is run on the test sample to compute the error rate. Cross-validation is sample-splitting repeatedly applied to the data after partitioning it into  $k$  equal parts. At each time,  $k-1$  parts are used as the training set and the other one as the test set. The performance is estimated as the average of the  $k$  individual values. These methods are used to correctly evaluate an algorithm's performance, avoid overfitting, and compare different algorithms.
3. It is a nonparametric method and doesn't assume the relationship between the predictor and response variables is linear, so it can work well even when the true relationship is not linear. [also, ease of interpretation]
4. Though you may achieve the linearity, the error structure (gaussian; homogeneous variance) can be destroyed by transforming the response variable. That can make various inferential conclusions from typical linear regression invalid. To check if it is OK to do the transformation, look at the plot of the residual against independent variable.
5. Yes, there are robust nonlinear regressions. Nonlinearity is about the shape of the regression function and robustness is about relaxing the Gaussian error assumption, and those two aspects are independent of each other. In R, one can use `nlm` (nonlinear minimization) to find such curve after specifying the correct function to minimize. ('`nls`' does nonlinear least squares regression)
6. Bandwidth is more important in determining the shape and smoothness of the fitted curve. Even within a single method, different bandwidths lead to tremendously different curve estimates. On the other hand, different methods can be made to yield quite similar curve estimate by tuning their bandwidth.
7. Advantage: It doesn't suffer from presence of outliers and influential points. Disadvantage: it doesn't provide as much detailed inference results as the typical linear model [it's also computer intensive, which may matter if you have to do run it many times]