ANLY512: Homework 3

Due Mar 10 11:59pm

Problem 1: Use polynomials and ridge regression to predict stock returns

This problem uses the built in `EuStockMarkets` dataset. The dataset contains time series of closing prices of major European stock indices from 1991 to 1998. We use only the `FTSE` column in this problem. The dataset is a time series object, but you will need to extract the `FTSE` column and make it into a data frame.

part a: Fit polynomial models of degrees 4, 8, 12 to the FTSE data and plot all three fitted curves together with a scatterplot of the data. Comment on the plots. Which features in the data are resolved by the polynomial models? Which features are not resolved? Do the polynomial curves show any artifacts such as oscillations?

part b: Use ridge regression to regularize the polynomial model of degree 12. Use $\lambda_1 SE$. Plot the resulting polynomial model onto the the data and comment on it.

Problem 2: Improve advertising budgets using GAMs

Use the Advertising dataset, which can either be found here. Split the data into a training and test set (70% / 30%).

part a: Fit generalized additive models to predict sales, using smoothing splines of degrees 2, 3, 4, 5, 6 for the three predictors. How do the rms prediction errors compare to the rms prediction error of a multiple regression model on the training set? On the test set?

part b: Is there evidence of overfitting?

part c: You now have six models (five GAM and one LM). Which model should be used? Explain your answer.

Problem 3: Use LASSO to predict housing prices in Boston

Consider the `Boston` data from the `MASS` package. We want to use LASSO to predict the median home value `medv` using all the other predictors.

part a: Set up the LASSO and plot the trajectories of all coefficients. What are the last five variables to remain in the model?

part b:Find the 1SE value of λ, using 10-fold cross-validation. What is the cross validation estimate for the residual standard error?

part c: Rescale all predictors so that their mean is zero and their standard deviation is 1. Then set up the LASSO and plot the trajectories of all coefficients. What are the last five variables to remain in the model? Compare your answer to part a.

part d: Find the 1SE value of λ using 10-fold cross-validation. What is the cross validation estimate for the residual standard error now? Does rescaling lead to a better performing model?

Problem 4:Predict bike share usage in Seoul using ridge and LASSO regressions

Access the dataset here. Filter the data to only include rows with "`functioning days`" == 'Yes'. Next drop the columns `Date`, `Hour`, `Seasons`, and `Holiday`, and `Functioning Day`. Then drop any rows that have any missing values in any columns. Hint: You will need to rename some of the variable names because they include non-ASCII characters. This will help you later on.

part a: Run a linear regression to predict rented bike count using the remaining 8 variables in the dataset. Report the MSE and the most influential variables.

part b: Fit a ridge regression model with the optimal λ chosen by cross validation. Report the CV MSE.

part c: Perform the same fit using a LASSO regression this time. Choose the optimal λ using cross validation. Report on the remaining variables in the model and the CV MSE. How does this performance compare to ridge and a plain linear model?

part e: Interpretion and communication. Write a short paragraph about your analysis and recommendations, explaining the most important factors for high bike share usage, why you came to that conclusion, and what actions can be taken by a bike rental company based on this information.

Problem 5: Compare the characteristics of two different smoothing splines

Consider two curves called $\hat{g}_1$ and $\hat{g}_2$ are as follows:

$$\hat{g}_1 = argmin_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int [g^{(3)}(x)]^2 \right)$$

$$\hat{g}_2 = argmin_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int [g^{(4)}(x)]^2 \right)$$

where $g^{(m)}$ represents the mth derivative of g.

part a: As $\lambda \to \infty$, which function ($\hat{g_1}$ or $\hat{g_2}$) will have the smaller training RSS?

part b: As $\lambda \to \infty$, which function ($\hat{g_1}$ or $\hat{g_2}$) will have the smaller test RSS?

part c: For $\lambda = 0$, which function ($\hat{g_1}$ or $\hat{g_2}$) will have the smaller training and test RSS?

Problem 6: Explain the behavior of the curve for a variety of λ and m values.

Suppose a curve $\hat{g}$ is fit smoothly to a set of n points as follows:

$$\hat{g} = argmin_g \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int [g^{(m)}(x)]^2 \right)$$

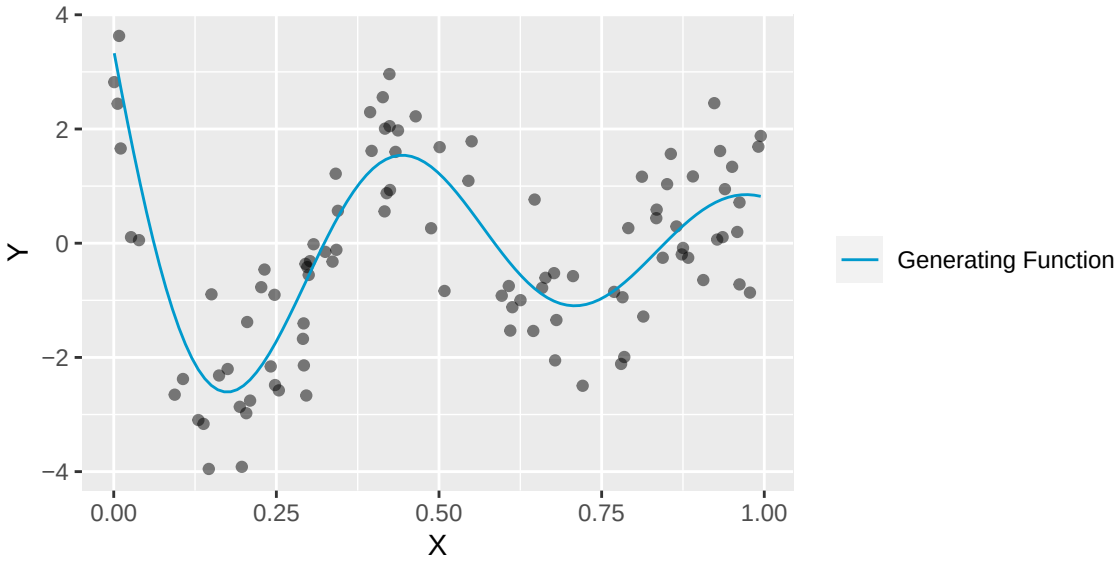
where $g^{(m)}$ is the mth derivative of $\hat{g}$ and $g^{(0)} = g$. Provide plots of $\hat{g}$ in each of the following scenarios along with the original points provided.

Use the following starter code to make your set of points and plot your various model predictions.

```
set.seed(325626)

X <- runif(100)
eps <- rnorm(100)
Y <- sin(12*(X + 0.2)) / (X + 0.2) + eps
generating_fn <- function(X) {sin(12*(X + 0.2)) / (X + 0.2)}
df <- data.frame(X, Y)

ggplot(df, aes(x = X, y = Y)) +
  geom_point(alpha = 0.5) +
  stat_function(fun = generating_fn, aes(col = "Generating Function")) +
  scale_color_manual(values = "deepskyblue3") +
  theme(legend.position = "right", legend.title = element_blank())
```



part a: $\lambda = \infty, m = 0$.

part b: $\lambda = \infty, m = 1$.

part c: $\lambda = \infty, m = 2$.

part d: $\lambda = \infty, m = 3$.

part e: $\lambda = 0, m = 3$.

part f: Fit a smoothing spline on the dataset and report the optimal lambda