FIN 510 Big Data Analytics in Finance

Lab 11: Cross-Validation

Due on 10/09/2021

Predicting Boston Housing Prices

The file BostonHousing.csv contains information collected by the US Bureau of the Census concerning housing in Boston, Massachusetts. The dataset includes information on 506 census housing tracts in the Boston area. The dataset contains the median housing price (MEDV) and other information that may affect house prices, such as weighted distances to five Boston employment centers (DIS), nitric oxides concentration (NOX), and the crime rate in the area (CRIM). Description of variables can be found below.

The goal is to validate three different models that predict median house value (MEDV), each one increasing in complexity.

Model 1: model MEDV as a cubic function of DIS.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2 + \beta_3 {x_1}^3 + \epsilon$$

where x_1 is DIS and Y is MEDV

Model 2: add a 4th degree polynomial in NOX as predictors, in addition to predictors in Model 1

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2 + \beta_3 {x_1}^3 + \beta_4 x_2 + \beta_5 {x_2}^2 + \beta_6 {x_2}^3 + \beta_7 {x_2}^4 + \epsilon$$

where x_1 is DIS, x_2 is NOX, and Y is MEDV

Model 3: add a 5th degree polynomial in CRIM as predictors, in addition to predictors in Model 2

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 {x_1}^2 + \beta_3 {x_1}^3 + \beta_4 x_2 + \beta_5 {x_2}^2 + \beta_6 {x_2}^3 + \beta_7 {x_2}^4 +$$
$$+\beta_8 x_3 + \beta_9 {x_3}^2 + \beta_{10} {x_3}^3 + \beta_{11} {x_3}^4 + \beta_{12} {x_3}^5 + \epsilon$$

where x_1 is DIS, x_2 is NOX, x_3 is CRIM, and Y is MEDV

DESCRIPTION OF VARIABLES FOR BOSTON HOUSING EXAMPLE	
CRIM	Per capita crime rate by town
ZN	Proportion of residential land zoned for lots over 25,000 ft^2
INDUS	Proportion of nonretail business acres per town
CHAS	Charles River dummy variable (=1 if tract bounds river; =0 otherwise)
NOX	Nitric oxide concentration (parts per 10 million)

RM	Average number of rooms per dwelling
AGE	Proportion of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centers
RAD	Index of accessibility to radial highways
TAX	Full-value property tax rate per $10,000
PTRATIO	Pupil/teacher ratio by town
LSTAT	Percentage lower status of the population
MEDEV	Median value of owner-occupied homes in $1000s

0) Load the package

Use library() to load boot.

1) Create a data frame

Load the data and save the result in a data frame named housing.df. Return the first six rows and column names using head() and names(), respectively.

2) Compute leave-one-out cross-validation prediction error

In this question, use the leave-one-out cross-validation method to compute the mean squared error (MSE) for three models that predict median house value (MEDV).

2.1) Model 1: model MEDV as a cubic function of DIS

According to Model 1, fit a regression model of MEDV as a cubic function of DIS on data frame housing.df using glm(). Specify poly(DIS, 3) as predictors to compute a third degree polynomial in DIS. Save the result as glm1 and return summary(glm1).

Based on glm1, compute the leave-one-out cross-validation prediction error using cv.glm() and save the result as loocv.err1. Calculate the cross-validated mean squared error (MSE) using loocv.err1$delta[1].

2.2) Model 2: add a 4th degree polynomial in NOX as predictors, in addition to predictors in Model 1

According to Model 2, fit a regression model of MEDV as a cubic function of DIS and a quartic function (fourth degree polynomial) of NOX on data frame housing.df using glm(). Specify poly(DIS, 3) and poly(NOX, 4) as predictors to compute a third degree polynomial in DIS and a fourth degree polynomial in NOX. Save the result as glm2 and return summary(glm2).

Based on glm2, compute the leave-one-out cross-validation prediction error using cv.glm() and save the result as loocv.err2. Calculate the cross-validated mean squared error (MSE) using loocv.err2$delta[1].

2.3) Model 3: add a 5th degree polynomial in CRIM as predictors, in addition to predictors in Model 2

According to Model 3, fit a regression model of MEDV as a cubic function of DIS, a quartic function of NOX, and a quintic function (fifth degree polynomial) of CRIM on data frame housing.df using glm(). Specify poly(DIS, 3), poly(NOX, 4), and poly(CRIM, 5) as predictors to compute a third degree polynomial in DIS, a fourth degree polynomial in NOX, and a fifth degree polynomial in CRIM. Save the result as glm3 and return summary(glm3).

Based on glm3, compute the leave-one-out cross-validation prediction error using cv.glm() and save the result as loocv.err3. Calculate the cross-validated mean squared error (MSE) using loocv.err3$delta[1].

2.4) Identify the best a model according to the leave-one-out cross-validation method

Based on the leave-one-out cross-validated MSE you calculated in questions 2.1 to 2.3, which model is best for predicting median home values?

3) Compute 5-fold cross-validation prediction error

In this question, use the 5-fold cross-validation method to compute the mean squared error (MSE) for three models that predict median house value (MEDV).

3.1) Model 1: model MEDV as a cubic function of DIS

Use set.seed(1) to set the random seed. Based on glm1, compute the 5-fold cross-validation prediction error using cv.glm() with K=5 and save the result as kfcv.err1. Calculate the cross-validated mean squared error (MSE) using kfcv.err1$delta[1].

3.2) Model 2: add a 4th degree polynomial in NOX as predictors, in addition to predictors in Model 1

Use set.seed(1) to set the random seed. Based on glm2, compute the 5-fold cross-validation prediction error using cv.glm() with K=5 and save the result as kfcv.err2. Calculate the cross-validated mean squared error (MSE) using kfcv.err2$delta[1].

3.3) Model 3: add a 5th degree polynomial in CRIM as predictors, in addition to predictors in Model 2

Use set.seed(1) to set the random seed. Based on glm3, compute the 5-fold cross-validation prediction error using cv.glm() with K=5 and save the result as kfcv.err3. Calculate the cross-validated mean squared error (MSE) using kfcv.err3$delta[1].

3.4) Identify the best a model according to the 5-fold cross-validation method

Based on the 5-fold cross-validated MSE you calculated in questions 3.1 to 3.3, which model is best for predicting median home values?