Advanced Statistical Learning - Assignment 1

Dataset

The data for this assignment are from a large survey from the US. The survey data were, for example, used to evaluate long-term benefits of participation in the Head Start program, that provides early childhood education, health, nutrition, and parent involvement services to children from low-income and families. The data comprise a birth cohort of children, who were enrolled in the National Longitudinal Survey of Youth (NLSY) around 1984. The cohort was tracked until 2004. The dataset contains both characteristics of the children and their families (mother, father, household). From many families, multiple children participated in the study.

All variables in the dataset were assessed from 1984 to 1996, except for the response, which was assessed in 2004. Your aim is to predict the response variable `Test_Pct104`.

For the purposes of the current assignment, we have retained only those observations for which the response variable was observed, and imputed missing values in the predictor variables, so results may not perfectly generalize to the real world. A table with variable names and descriptions is provided as an .xlsx file on Brightspace.

Read in the data as follows:

```
data <- readRDS(file = "Data_Assignment1_ASL.Rda")
```

Separate the data using a random split into 75% train and 25% test observations. Before random sampling, set the seed to your student number.

General instructions

- Hand in your report and R code in separate files (or add the code as an appendix to your report).
- Your report should preferably be a .pdf file and should be understandable without consulting the code. Your code can be an .R or .Rmd file or other format.
- Write your answers in full sentences. Your report does not need to read like an essay. Provide your answers in a numbered fashion, in the order the questions are presented here.
- Numbers of words are maxima: It is certainly possible to obtain a perfect score with less words.
- Make sure tables and figures are well-formatted. You may omit figure/table captions and numbering.

Questions

XGBoost

Question 1

For XGBoost, define a grid of hyper-parameter values, and use cross-validation on the training dataset to find the optimal set of hyper parameters. Make sure to specify the grid so that random-forest, gradient-

boosting and combinations of these approaches have been tried. Make sure to also optimize other important hyper parameters, but you may fix the size of subsamples to 0.632. It may be necessary to adjust your tuning grid after you inspect the results of your first iteration. In your report, include a plot of the results and describe how each hyper parameter affects cross-validated predictive performance. (Max. 200 words, excluding figure.)

Question 2

Using the optimal set of parameters found in question 1), fit an XGBoost ensemble to the training dataset. Compute and report predictive accuracy on the test observations. (Max. 50 words.)

Question 3

Extract the importances computed by XGBoost. Pick the importance measure you think is most meaningful, and motivate your choice. (Max. 50 words.)

Question 4

Report the variable importance measure your chose in question 3) for the ensemble fitted in question 2) in a table or figure. Report which variables you deem obviously predictive of the response. (Max. 50 words.)

Gaussian Process Regression

Question 5

a) Use the training set to compare the following kernel families: `cf_lin`, `cf_sexp`, `cf_nn` (which corresponds to a neural network). The goal is to select the kernel family and its hyperparameters that optimize the model evidence. (Max. 50 words.)

b) Explain two advantages of selecting a model via the model evidence compared to cross-validation. Base your discussion on a comparison of the cross-validation you did for tuning the hyper-parameters of XGBoost and the model selection you just did for Gaussian process regression. (Max. 200 words.)

Hints:

- GPR can deal with categorical variables, but specifying a proper kernel for categorical variables is out of the scope of this lecture series. Thus, only use the numeric predictors: `Data_Assignment1_ASL[sapply(Data_Assignment1_ASL, is.numeric)]`
- For GPR, if data are on completely different scales as they are here, standardization is heavily advised. The reason is essentially the same as for SVMs or regularized regression. `gplite` can automate this for you; see the explanation about `normalize` in `?cf`

Question 6

a) For the best kernel, obtain predictions on the test set and calculate the MSE. (Max. 50 words.)
b) Additionally, calculate the proportion of test points that are within ± 1.96 standard deviations (of the predictive distribution) of the predictive mean. (Max. 50 words.) **Hint:** Remember from lecture 2, that there is a difference between predicting the latent $y(x)$ and the observations $t(x)$, `gplite` can only predict the latent. The formulas in lecture 2 explain how to transform the predictive distribution from the latent to the observation.

Question 7

Consider the following two linear Bayesian regression models representend in their GPR form. Which of them is more complex. Justify your answer. (Max. 200 words.)

1. $k(x, x') = 10x^\top x'$ and $\sigma_\epsilon^2 = \beta^{-1} = 1$
2. $k(x, x') = 2x^\top x'$ and $\sigma_\epsilon^2 = \beta^{-1} = 1$

Question 8

a) Does the XGBoost or the GPR model lead to more accurate predictions on your test set? (Max. 50 words.)
b) Outside of prediction accuracy, name two advantages of XGBoost as compared to GPR and vice versa. (Max. 200 words.)