Emily Robinson
Fall 2018

Final Project Proposal

Improved Classification with Ensemble Learning

Dataset:

The dataset I have selected for my project is the *Pima Indians Diabetes Database* found on Kaggle through the UCI data repository, originally from the National Institute of Diabetes and Digestive and idney Diseases. The patients in this study are all females at least age 21 years old and of Pima Indian heritage. The target variable, Outcome, is binary and indicates whether or not a patient has diabetes. The objective is to correctly classify individuals to have diabetes or not based on several medical predictor variables such as the number of pregnancies the patient has had, their BMI, insulin level, age, etc.

Software:

I will use R statistical software to obtain classifications and present my modeling process. One particular package I will explore is SuperLearner, used for ensemble methods by assigning weights to the predictions obtained by various classification models. Another package I may use to compare is caretEnsemble containing three primary functions to develop ensemble models.

Summary:

For my presentation, I will begin with a motivating example for ensemble modeling and transition into how it relates to prediction and classification methods we have been learning in class this semester. I will briefly describe a few different types of ensemble learning methods including bagging, boosting, stacking. I will initially split my dataset into three cross validation sets (training, validation, and testing). For modeling, I will be using the stacking method to combe heterogeneous classification methods that complement each other in order to improve the accuracy of classification on the testing set. I have selected a dataset used for classifying the incidence of heart disease. Using R, I will begin by using three to five classification methods learned in class and compare the misclassification rate and AUC values of the testing sets. With further exploration of ensemble learning, I will combine the previous models using one of the methods described as well as compare it to models developed using the SuperLearner and caretStack packages in R for ensemble learning. To conclude, I will provide resources for further exploration and application to prediction of continuous responses. One topic that relates to ensemble learning is hybrid classification. I plan to explore this method further to compare and contrast it to ensemble methods.

Potential Resources:

Eric Polley, Erin LeDell, Chris Kennedy and Mark van der Laan (2018). SuperLearner: Super Learner Prediction. R package version 2.0-24. https://CRAN.R-project.org/package=SuperLearner

Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. Chapter 16. New York, NY: Springer.

Kennedy, C., & University of California, Berkeley. (2017, March 16). Guide to SuperLearner. Retrieved from https://cran.r-project.org/web/packages/SuperLearner/vignettes/Guide-to-SuperLearner.html

Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. http://rexa.info/paper/04587c10a7c92baa01948f71f2513d5928fe8e81. *In Proceedings of the Symposium on Computer Applications and Medical Care* (pp. 261--265). IEEE Computer Society Press.

Steinki, Oliver and Mohammad, Ziad, Introduction to Ensemble Learning (August 15, 2015). Available at SSRN: https://ssrn.com/abstract=2634092 or http://dx.doi.org/10.2139/ssrn.2634092

Vardeman, S. B. (2015, April 17). Lecture Notes on Modern Multivariate Statistical Learning. Chapter 10. Retrieved from http://www.analyticsiowa.com/wp-content/uploads/2015/04/StatLearningNotesTOC.pdf

Vardeman, S. B. (2018, July 16). Lecture Notes on Modern Multivariate Statistical Learning-Version II. Chapter 11. Retrieved from http://www.analyticsiowa.com/wp-content/uploads/2018/07/StatLearningNotesII.pdf

Zachary A. Deane-Mayer and Jared E. Knowles (2016). caretEnsemble: Ensembles of Caret Models. R package version 2.0.0. https://CRAN.R-project.org/package=caretEnsemble