BIOST 2079 Homework 3

Distributed: 11/29/2023
Deadline: 12/15/2023

Theory

1. In decision trees for classification, we need to select an impurity function to determine the best split to construct the tree. Gini index and entropy are two most common choices. In order for them to be a valid impurity function, show that the function Gini index: $\phi(p) = \sum_{k \neq l} p_k \cdot p_l = 1 - \sum_k p_k^2 = \sum_k p_k(1 - p_k)$ and Entropy: $\phi(p) = -\sum_k p_k \cdot \log(p_k)$ takes the maximum value when $p_1 = \cdots = p_K = 1/K$ (the most impure) and minimum value when the probability concentrate on one of the K classes (the most pure). For simplicity, you may investigate with $K = 2$ ($p_1 = p,\ p_2 = 1 - p$) and express the indices in terms of just p. (5 points)

Computing

* You are encouraged to use R Markdown (template provided) to generate pdf reports with embedded R codes and outputs.

2. This question is modified from Question 10 in Chapter 4.7 in ISLR. The question should be answered using the Weekly data set, which is part of the ISLR package. This data contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010. Write a data analysis report addressing the following problems. (15 points)
   (a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to any patterns?
   (b) Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?
   (c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.
   (d) Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).
   (e) Repeat (d) using LDA.
   (f) Repeat (d) using QDA.
   (g) Which of these methods appears to provide the best results on this data?
   (h) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data.

[15% of final grade] *Kaggle competition* (submission closes on Friday December 15)

Work on the Kaggle prediction competition: Predicting hospital readmission for inpatient diabetes encounters. More details can be found on Canvas and the Kaggle page. Make sure to preprocess both training and testing data together using a consistent method, so that the trained model can be used on the testing data. I have provided a code template to get you started (Files>Labs>Kaggle Code). The code implements some preprocessing steps that helps handle missing data, categorical and string data. This is not necessarily the right/only way to go about it, but I want you to have something that you can build upon. You can think of ways to improve your model based on this as a starting point. Feel free to use your own training pipeline if you prefer, and you are more encouraged to be creative. You will be scored based on completion (making at least 1 submission) and be rewarded for extra points if you rank in the top 5.