Assignment 4: Chapters 7 and 8

Gary Mudd

Solving Problems with Machine Learning

Professor Dr. Abdulaziz Ahmed

April 7, 2021

Q1 (20 points): Personal Loan Acceptance. Universal Bank is a relatively young bank growing rapidly in terms of overall customer acquisition. The majority of these customers are liability customers (depositors) with varying sizes of relationship with the bank. The customer base of asset customers (borrowers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business. In particular, it wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise smarter campaigns with better target marketing. The goal is to use k-NN to predict whether a new customer will accept a loan offer. This will serve as the basis for the design of a new campaign.

The file UniversalBank\_\_Assignment\_4\_Q1.csv contains data on 5000 customers. The data include customer demographic information (age, income, etc.), the customer’s relationship with the bank (mortgage, securities account, etc.), and the customer response to the last personal loan campaign (Personal Loan). Among these 5000 customers, only 480 (=9.6%) accepted the personal loan that was offered to them in the earlier campaign.

Assuming the data is partitioned into training (60%) and validation (40%) sets.Prepare the data file build a prediction model. Then, solve the following questions:

a. Consider the following customer: Age = 40, Experience = 10, Income = 84, Family = 2, CCAvg = 2, Education\_1 = 0, Education\_2 = 1, Education\_3 = 0, Mortgage = 0, Securities Account = 0, CD Account = 0, Online = 1, and Credit Card = 1. Perform a k-NN classification with all predictors except ID and ZIP code using k = 1. Remember to transform categorical predictors with more than two categories into dummy variables first. Specify the success class as 1 (loan acceptance), and use the default cutoff value of 0.5. How would this customer be classified? (10 points)

b. What is a choice of k that balances between overfitting and ignoring the predictor information? (5 points)

c. Show the confusion matrix for the validation data that results from using the best k. (5 points)

Q2 (20): Predicting Housing Median Prices. The file BostonHousing\_Assignment\_4\_Q2.csv contains information on over 500 census tracts in Boston, where for each tract multiple variables are recorded. The last column (CAT.MEDV) was derived from MEDV, such that it obtains the value 1 if MEDV > 30 and 0 otherwise. Consider the goal of predicting the median value (MEDV) of a tract, given the information in the first 12 columns. Partition the data into training (60%) and validation (40%) sets. Build a prediction model and perform the following tasks:

a. Perform a k-NN prediction with all 12 predictors (ignore the CAT.MEDV column), trying values of k from 1 to 5. Make sure to normalize the data. What is the best k? What does it mean?

b. Predict the MEDV for a tract with the following information, using the best k: CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO LSTAT 0.2 0 7 0 0.538 6 62 4.7 4 307 21 10MGT-665 2

c. If we used the above k-NN algorithm to score the training data, what would be the error of the training set?

d. Why is the validation data error overly optimistic compared to the error rate when applying this k-NN predictor to new data?

Q3 (20 points) Automobile Accidents. The file accidentsFull\_Assignment\_4\_Q3.csv contains information on 42,183 actual automobile accidents in 2001 in the United States that involved one of three levels of injury: NO INJURY, INJURY, or FATALITY. For each accident, additional information is recorded, such as day of week, weather conditions, and road type. A firm might be interested in developing a system for quickly classifying the severity of an accident based on initial reports and associated data in the system (some of which rely on GPS-assisted reporting). Our goal here is to predict whether an accident just reported will involve an injury (MAX\_SEV\_IR = 1 or 2) or will not (MAX\_SEV\_IR = 0). For this purpose, create a dummy variable called INJURY that takes the value “yes” if MAX\_SEV\_IR = 1 or 2, and otherwise “no.”

Partition the data into training (60%) and validation (40%).

a. Assuming that no information or initial reports about the accident itself are available at the time of prediction (only location characteristics, weather conditions, etc.), which predictors can we include in the analysis?

b. Run a naive Bayes classifier on the complete training set with the relevant predictors (and INJURY as the response). Note that all predictors are categorical. Show the confusion matrix.

c. What is the overall error for the validation set?

d. What is the percent improvement relative to the naive rule (using the validation set)?