Homework 4

IE 7275 Data Mining in Engineering

Note: Read the materials "Tutorial on *k*-NN with R.pdf." and "Tutorial on Naïve Bayes with R.pdf". Practice example problems when possible.

Problem 1 (Personal Loan Acceptance, *k*-NN) [30 points]

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 of our analysis is to model the previous campaign's customer behavior to analyze what combination of factors make a customer more likely to accept a personal loan. This will serve as the basis for the design of a new campaign.

The file **`UniversalBank.xlsx`** 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.

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

a. Consider the following customer: Age=40, Experience=10, Income=84, Family=2, CCAvg=2, Education2=1, Education3=0, Mortgage=O, Securities Account=O, CD Account=O, 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?

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

c. Show the classification matrix for the validation data that results from using the best k.

d. 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. Classify the customer using the best *k*.

e. Repartition the data, this time into training, validation, and test sets (50% : 30% : 20%). Apply the *k*-NN method with the *k* chosen above. Compare the classification matrix of the test set with that of the training and validation sets. Comment on the differences and their reason.

Problem 2 (Predicting Housing Median Prices *k*-NN) [30 points]

The file **BostonHousing.xlsx** contains information on over 500 census tracts in Boston, where for each tract 14 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 13 columns. Partition the data into training (60%) and validation (40%) sets.

a. Perform a *k*-NN prediction with all 13 predictors (ignore the CAT.MEDV column), trying values of *k* from 1 to 5. Make sure to normalize the data (click "normalize input data"). What is the best *k* chosen? 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	10

c. Why is the error of the training data zero?

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

e. If the purpose is to predict MEDV for several thousands of new tracts, what would be the disadvantage of using *k*-NN prediction? List the operations that the algorithm goes through in order to produce each prediction.

Problem 3 (Automobile Accidents, Naïve Bayes) [40 points]

The file **Accidents.xlsx** 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."

a. Using the information in this dataset, if an accident has just been reported and no further information is available, what should the prediction be? (INJURY = Yes or No?) Why?

b. Select the first 12 records in the dataset and look only at the response (INJURY) and the two predictors WEATHER_R and TRAF_CON_R.

    i. Using Excel tools create a pivot table that examines INJURY as a function of the 2 predictors for these 12 records. Use all 3 variables in the pivot table as rows/columns, and use counts for the cells.

    ii. Compute the exact Bayes conditional probabilities of an injury (INJURY =Yes) given the six possible combinations of the predictors.

    iii. Classify the 12 accidents using these probabilities and a cutoff of 0.5.

    iv. Compute manually the naive Bayes conditional probability of an injury given WEATHER_R = 1 and TRAF_CON_R = 1.

    v. Run a naive Bayes classifier on the 12 records and 2 predictors. Obtain probabilities and classifications for all 12 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (= ordering) of observations equivalent?

c. Let us now return to the entire dataset. Partition the data into training/validation sets.

    i. 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? (Use the

Data_Codes sheet.)

ii. 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 classification matrix.

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

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

v. Examine the conditional probabilities output. Why do we get a probability of zero for P(INJURY =No | SPD_LIM = 5)?

Files Included in the Folder:

1. **Homework 4.pdf**
2. **Tutorial on *k*-NN with R.pdf**
3. **Tutorial on Naïve Bayes with R.pdf**
4. **UniversalBank.xlsx**
5. **BostonHousing.xlsx**
6. **Accidents.xlsx**