PROBLEMS

7.1 Calculating Distance with Categorical Predictors. This exercise with a tiny dataset illustrates the calculation of Euclidean distance, and the creation of binary dummies. The online education company Statistics.com segments its customers and prospects into three main categories: IT professionals (IT), statisticians (Stat), and other (Other). It also tracks, for each customer, the number of years since first contact (years). Consider the following customers; information about whether they have taken a course or not (the outcome to be predicted) is included:

Customer 1: Stat, 1 year, did not take course
Customer 2: Other, 1.1 year, took course

a. Consider now the following new prospect:

Prospect 1: IT, 1 year

Using the above information on the two customers and one prospect, create one dataset for all three with the categorical predictor variable transformed into 2 binaries, and a similar dataset with the categorical predictor variable transformed into 3 binaries.

b. For each derived dataset, calculate the Euclidean distance between the prospect and each of the other two customers. (*Note*: while it is typical to normalize data for k-NN, this is not an iron-clad rule and you may proceed here without normalization.)

c. Using k-NN with $k = 1$, classify the prospect as taking or not taking a course using each of the two derived datasets. Does it make a difference whether you use 2 or 3 dummies?

7.2 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.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.

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, 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?

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

c. Show the confusion 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 confusion matrix of the test set with that of the training and validation sets. Comment on the differences and their reason.

7.3 Predicting Housing Median Prices. The file *BostonHousing.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.

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, and choose function *knn()* from package `class` rather than package `FNN`. To make sure R is using the `class` package (when both packages are loaded), use *class::knn()*. 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	10

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?

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.