Assignment: K Nearest Neighbor and VDM

In this assignment you'll be applying KNN algorithm to a healthcare data set. The data files for training and test are named *healthcareTrain.csv* and *healthcareTest.csv*, respectively. The details of this data set will be discussed in the Healthcare Case Study module. You can also find the description of each feature in the *DataDictionary* file.

Here is a brief description of those features that you'll be using in this assignment:

1. pre-rx-cost: Total pharmacy costs per person
2. numofgen: Number of generic scripts
3. numofbrand: Number of brand scripts
4. generic-cost: Cost of generic scripts filled
5. adjust-total-30d: 30 day adjusted fill rate
6. num-er: Number of ER visits
7. region: US Census Region (1 Northeast, 2 Midwest, 3 South, 4 West)
8. pdc-80-flag: Adherent (A categorical variable that indicates if patients have adhered to taking their medications more than 80% of the time; =1 if pdc $\geq$ 0.80; =0 otherwise)

Problem 1 (20 points)

In this problem you apply KNN to the healthcare data to predict adherent class (pdc-80-flag).

1. (10 points) Predict the pdc-80-flag using the following features "pre-rx-cost","numofgen","numofbrand","generic-cost","adjust-total-30d", and "num-er". Determine the accuracy rate for test set for k = 75 to 105 with a step size of 2 and report it in a table. Use linear normalization method to normalize the input features and Euclidean distance for distance measure. Note that you must use the training parameters for normalization of test points. You can use built-in knn function in R for this problem.
2. (10 points) Plot the accuracy rate vs. K. Which value of K gives you the best accuracy rate?

Problem 2 (40 points)

In this problem you'll continue using the healthcare data from the previous problem. You'll use the Value Distance Metric (VDM) to find the distance between symbolic feature values Northeast, Midwest, South, and West, and further use this information in KNN algorithm to predict pdc-80-flag.

1. (10 points) Find all the relevant conditional probabilities for finding VDM for symbolic variable region and report your results in a table.

2. (10 points) Use results in part 1 to find the distance between symbolic feature values Northeast, Midwest, South, and West using VDM equation. Report the distances in a table.

3. (10 points) Use this variable (region) in conjunction with the variables of problem 1 and regenerate your model, for k = 75 to 105 with a step size of 2. Report the mean accuracy rate. Compare this mean with mean accuracy rate from previous problem. Has it increased for decreased?

4. (5 points) Plot the accuracy rate vs. K. Which value of K gives you the best accuracy rate?

5. (5 points) What did your model predict for the $100^{th}, 200^{th}$, and 300^{th} test points?