Assignment: K Nearest Neighbor-VDM-Logistic Regression

In the first 2 problems, 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. You can 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 ≥ 0.80; =0 otherwise)

Problem 1 (20 points)

In this problem you apply built-in KNN package to the healthcare data to predict adherent class (pdc-80-flag). Use of pipeline is recommended.

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.
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. You can NOT use the built-in packages for this problem.

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 best model predict for the $100^{th}, 200^{th}$, and 300^{th} test points?

Problem 3 (15 points)

Given the function $f(x) = x^2 + 6x$:

1. Use derivative of $f(x)$ to find the value of x that minimizes this function. (2 points)

2. Use gradient descent to find the value of x that minimizes this function. Compare your answer with the previous part. (13 points)

Problem 4 (35 points)

The Space Shuttle Challenger disaster occurred on January 28, 1986, when it broke apart 73 seconds into its flight, leading to the deaths of its seven crew members. The spacecraft disintegrated over the Atlantic Ocean, off the coast of central Florida at 11:38 EST. Disintegration of the entire vehicle began after an O-ring seal in its right solid rocket booster failed at liftoff. Subsequently, a special commission was appointed to investigate the accident. The commission found that NASA disregarded warnings from engineers about the dangers of launching posed by the low temperatures of that morning, claiming that engineers could not provide a convincing argument against the launch (source: Wikipedia, Applied Probability for Engineers).

File Oring.csv provides data on launch temperature and O-ring failure for the 24-space shuttle launches prior to the Challenger disaster. There are six O-rings used to seal field joints on the rocket motor assembly. A +1 in the O-rings indicates that at least one O-ring failure had occurred on that launch and a 0 indicates that no failure had occurred.

1. Normalize the launch temperature using the expression $\frac{x-\mu}{\sigma}$. (3 points)
2. Create a logistic regression model using the gradient decent technique to predict the probability of O-ring failure based on the launch temperature. Provide the equation for your model. You can NOT use built-in packages for this problem. (20 points)
3. Provide a plot of the original data along with your logistic model. (5 points)
4. The actual temperature at the Challenger launch was 31 degrees Fahrenheit. According to your model what was the probability of O-ring failure on the Challenger launch? Could the engineers have used your model to provide a convincing argument to NASA? Elaborate. (7 points)