CS 699
Assignment 2

Note: R is recommended since that is what is taught in class, but you may use any software you choose. As explained in submission guidelines on the next page, please provide your "final answers" in a document and provide a separate file containing your code or other detail that explains how you obtained your answers.

Problem 1 (10 points). Consider the following data about 10 people.

ID	job	marital	education	default	housing	loan	contact
P1	unemployed	married	primary	No	no	no	cellular
P2	services	married	secondary	No	yes	yes	cellular
P3	management	single	tertiary	No	no	no	cellular
P4	management	married	tertiary	No	yes	yes	unknown
P5	blue-collar	single	secondary	No	yes	no	unknown
P6	management	single	tertiary	No	no	yes	cellular
P7	self-employed	married	tertiary	No	yes	no	cellular
P8	technician	married	secondary	No	yes	no	cellular
P9	entrepreneur	married	tertiary	No	yes	no	unknown
P10	services	married	primary	No	yes	yes	cellular

Find the distance between P4 and P5 and the distance between P4 and P9. Is P4 closer to P5 or P9? The attributes are nominal, not ordinal. Distances should be between 0 and 1.

Problem 2 (10 points). Consider the following dataset with two objects.

Object	A1	A2	A3	A4
O1	88	47	32	6
O2	97	63	18	4

(1). Calculate the distance between O1 and O2 using the Manhattan distance.
(2). Calculate the distance between O1 and O2 using the Euclidean distance.

Problem 3 (10 points). Use the dataset *accidents1000.csv*. This data set has 999 observations and 11 features. The class feature is named MAX_SEV.

(1). Generate training and holdout partitions on the data set, holding out 1/3 of the data.
(2). Fit a k-Nearest Neighbor model. Be sure to center and scale the data. Select an optimal value of k. Use the optimal model to make predictions on the holdout data. Generate a confusion matrix and compute the accuracy.
(3). Does this seem to be a good model? Discuss why or why not.

(Assignment continues on the next page)

Problem 4 (35 points). Use the dataset *accidents1000.csv*. This data set has 999 observations and 11 features. The class feature is named MAX_SEV.

(1). Remove the observations with MAX_SEV = no-injury
(2). Generate training and holdout partitions on the remaining data. Use 1/3 of the data in the holdout.
(3). Fit a logistic regression model. Use the model to make predictions on the holdout data. Generate a confusion matrix and compute the accuracy and the F-score.
(4). Does this seem to be a good model? Discuss why or why not.
(5). This model has class imbalance. Think about what you've learned about over-sampling and under-sampling. One of these techniques is less likely to work when applied to this data set. Which one, and why?
(6). Apply the method (over-sampling or under-sampling) that is more likely to be helpful to the training data you already created in step (2).
(7). Fit a logistic regression model to the class-balanced data set you just created in step (6). Use the model to make predictions on the original holdout data you created in step (2). Generate a confusion matrix and compute the accuracy and the F-score.
(8.) Compare this model to the one you fit in step (3). Which, if any, seems to perform better? Discuss.
(9.) Calculate variable importance for the model you fit in step (7). What are the top 3 most important variables?

Problem 5 (20 points). Use the dataset *powdermetallurgy.csv*. This data set has 6253 observations and 8 features. The numeric value to be predicted is named Shrinkage.

(1). Generate training and holdout partitions on the data set. Use 1/3 of the data in the holdout.
(2). Fit a multiple linear regression model. Use the model to make predictions on the holdout data. Compute the MAE and the RMSE.
(3). Does this seem to be a good model? Discuss why or why not.
(4). Apply a regularization method of your choosing, such as LASSO or ridge regression. Use the regularized model to make predictions on the holdout data. Compute the MAE and the RMSE.
(5). Compare this model to the one you worked with in steps (2) and (3). Which, if any, seems to perform better? Discuss.

(Assignment continues on the next page)

Problem 6 (15 points). This problem is about the logistic regression we discussed in the class. Consider a dataset that has two independent variables A1 and A2 and a class attribute, which takes on either yes or no. Suppose you ran a logistic regression algorithm on the dataset and obtained the following coefficients for class yes:

Coefficient of A1 = 0.045
Coefficient of A2 = 0.003
Intercept = -3.485

Classify the following two unseen objects using the above model:

O1: A1 = 47, A2 = 213
O2: A1 = 65, A2 = 276

Assume that the classification threshold is 0.5.

Submission Guidelines:

- **Submit the solutions in a single Word or PDF document and upload it to Blackboard. You should separately upload the code or another document that explains how you arrived at the answers. Make sure each filename includes your name.**
    - **If you need more than 3-4 separate files, you might ZIP/RAR these files. Please still upload the Word/PDF separately. The separate Word/PDF greatly simplifies the facilitator's ability to comment on your assignment.**
- **Your submission should be organized well and easy to follow. Clearly list which question you are answering and provide the answer requested (one or more numbers, a screenshot, a plot, an explanation, whatever the question asks you to do). If a question contains multiple parts, you should clearly provide your answer for each part.**
- **If your facilitator tells you to submit the files differently than the above guidelines, you are expected to respect your facilitator's wishes starting on the next assignment.**
- **Facilitators can deduct up to 20% if you fail to follow these requirements (more if the questions are not actually answered).**
- **Facilitators can deduct 5% for each day the assignment is late.**
- **Unless your facilitator or the professor agrees, your assignment will not be graded if it is more than 3 days late (e.g., no credit will be given after Friday at 6 AM Boston time). The professor will usually ask the facilitator to make the decision but in rare cases (<1% of the time) has overridden a facilitator. Do not expect the professor to override in most cases.**