## Instruction

1. **You need to upload your final exam by 16th of December Thursday at 11:59 pm in Brightspace.**

2. **Use one of R, Stata, SPSS, SAS, or Matlab.**

3. **On the first page of your exam document to submit, please include the following statement with your signature.**
   "I understand that this is an open exam with a statistical software package, and that I can use the course materials and web resources. However, I will not communicate with any other persons and I will not seek or accept help from any other persons. In addition. I will not share any information about the exam with any other persons after submission."

4. **Start each main problem (not subset problems) on a new sheet of paper. Show your work and provide brief justification for your answers. Make sure to include exact key output results from your statistical software package to show only necessary details. For this part, no hand writing is allowed. Remember to mark your final answer clearly.**

5. **Make sure to submit only one PDF for your final exam.**

6. **DO NOT share your answers with any other students in our class. Any similar approaches and answers will be scrutinized.** Any violations of Purdue University Student Conduct Code will be reported to the Office of the Dean of Students. https://cm.maxient.com/reportingform.php?PurdueUniv&layout_id=10

7. **To make sure fairness, you are NOT allowed to ask questions directly related to this final exam.**

**Problem 1 (34 points)**

Recurrent breast cancer is breast cancer that comes back after initial treatment. Although the initial treatment is aimed at eliminating all cancer cells, a few may have evaded treatment and survived. In particular, you are interested in understanding which risk factors are causal determinants of breast cancer recurrence. The data were collected from a total of 277 patients with the following variables:

| Variable | Description |
|---|---|
| class_num (dependent) | Cancer recurrence<br>0 = no recurrence and 1: recurrence |
| id | ID number of the patient |
| age (independent) | Age of the patient (quantitative) |
| tumor_size (independent) | Tumor size [mm] (quantitative) |
| inv_nodes (independent) | The number of axillary lymph nodes having breast cancer (quantitative) |
| deg_malig (independent) | Degree of malignancy (categorical)<br>1, 2, and 3 |
| menopause (independent) | Menopause status (categorical)<br>1 = ge40 and 2 = premeno |
| node_caps (independent) | Lymph node metastasis of cancer (categorical)<br>0 = no and 1 = yes |
| breast (independent) | Cancer in the left or right breast (categorical)<br>1 = left and 2 = right |
| breast_quad (independent) | Breast quadrant (categorical)<br>1 = left-up, 2 = left-low, 3 = right-up, 4 = right-low, and 5 = central |
| irradiat (independent) | Radiation therapy (categorical)<br>0 = no and 1 = yes |

1. Using multiple logistic regression, build a logistic regression model to predict a risk of breast cancer recurrence, including all of the risk factors in the dataset. Define dummy variables for the categorical variables as needed. Show both commands & statistical results in R or Stata.

2. For the logistic regression model in Question 1, test the performance characteristics of the model by showing a receiver operating characteristic (ROC) curve and calculating the area under the ROC curve.

3. Using multiple logistic regression, determine an optimal model to predict a risk of breast cancer recurrence by including only statistically significant risk factors. Do not consider any interactions (i.e. effect modifiers). Use likelihood ratio tests to justify your analyses. Translate statistical results into your own scientific explanation. Show both your best model & significant risk factors.

4. For the best model in Question 3, test the performance characteristics of the model by showing a receiver operating characteristic (ROC) curve and calculating the area under the ROC curve.

5. Using the best model in Question 3, estimate the probability of breast cancer recurrence for the patient (“id” is 100). What is a predicted probability with a 95% confidence interval? Clearly indicate the calculation process for both the probability & the confidence intervals.

**Problem 2 (33 points)**

Problem 2 uses the same dataset as Problem 1.

| Variable | Description |
|---|---|
| class_num (dependent) | Cancer recurrence<br>0 = no recurrence and 1: recurrence |
| id | ID number of the patient |
| age (independent) | Age of the patient (quantitative) |
| tumor_size (independent) | Tumor size [mm] (quantitative) |
| inv_nodes (independent) | The number of axillary lymph nodes having breast cancer (quantitative) |
| deg_malig (independent) | Degree of malignancy (categorical)<br>1, 2, and 3 |
| menopause (independent) | Menopause status (categorical)<br>1 = ge40 and 2 = premeno |
| node_caps (independent) | Lymph node metastasis of cancer (categorical)<br>0 = no and 1 = yes |
| breast (independent) | Cancer in the left or right breast (categorical)<br>1 = left and 2 = right |
| breast_quad (independent) | Breast quadrant (categorical)<br>1 = left-up, 2 = left-low, 3 = right-up, 4 = right-low, and 5 = central |
| irradiat (independent) | Radiation therapy (categorical)<br>0 = no and 1 = yes |

To predict a risk of breast cancer recurrence, you are interested in constructing a shallow neural network (NN), consisting of an input layer, a hidden layer and an output layer. For training, you will use 222 patients ("id" from 1 to 222). For testing, you will use 55 patients ("id" from 223 to 277).

1. Based on Question 3 of Problem 1, list the statistically significant risk factors you identified. In the case of categorical variables, use "one-hot encoding" to convert them. Once again, you will need to consider the statistically significant risk factors from Question 3 of Problem 1. Show one-hot encoded data for each categorical variable. After one-hot encoding of the categorical variables, you will need to use all of the risk factors you selected as input variables. How many neurons (or nodes) do you need in the input layer?

2. Construct a regression NN to estimate the probability of breast cancer recurrence using only 222 patients ("id" from 1 to 222) for training. Make sure that your NN has an overall performance, comparable to the logistic regression model that you constructed in Question 3 from Problem 1. In addition, make sure to incorporate the information as follows:

   - **Hidden layer**
     - ↘ The number of neurons (or nodes) is 4.
     - ↘ Batch normalization is applied.
     - ↘ The rectified linear unit (ReLU) is used as an activation function.

   - **Output layer**
     - ↘ Determine if you need 1 or 2 neurons (or nodes).
     - ↘ The sigmoid function is used as an activation function.

   - **Training**

- ↘ For an epoch, have a maximum of 500.
- ↘ For a learning rate, have an initial learning rate of 0.01.

*** For a full credit of Question 2, you should highlight the important hyperparameters, such as layer configurations and training options. You do not have to show the entire code.**

3. Based on the regression NN of Question 2, estimate the probabilities of breast cancer recurrence for 55 patients ("id" from 223 to 277). Plot the probability (y-axis) over the patient id (from 223 to 277). For only the testing patients, test the performance characteristics of the network by showing a receiver operating characteristic (ROC) curve and calculating the area under the ROC curve.

4. Construct a classification NN for a screening test of breast cancer recurrence. Make sure to use only 222 patients ("id" from 1 to 222) for training.

   - **Hidden layer**
     - ↘ The number of neurons (or nodes) is 4.
     - ↘ Batch normalization is applied.
     - ↘ The rectified linear unit (ReLU) is used as an activation function.

   - **Output layer**
     - ↘ Determine if you need 1 or 2 neurons (or nodes).
     - ↘ The softmax function is used as an activation function.

   - **Training**
     - ↘ For an epoch, have a maximum of 500.
     - ↘ For a learning rate, have an initial learning rate of 0.01.

   *** For a full credit of Question 4, you should highlight the important hyperparameters, such as layer configurations and training options. You do not have to show the entire code.**

5. Based on the classification NN of Question 4, estimate the status of breast cancer recurrence for 55 patients ("id" from 223 to 277). A patient can be classified as "0 (no recurrence)" or "1 (recurrence)". Plot the predicted binary classification over the patient id (from 223 to 277). For only the testing patients, calculate a classification accuracy (ratio of correct cases to the total number of testing patients). Hint: you can compare the classified results with the actual breast cancer recurrence "class_num".

**Problem 3 (33 points)**

Suppose that we are interested in investigating if the use of air emergency medical services (EMS) actually saves lives. In this study, accident victims were transported to hospitals by a medical helicopter or more typically by a road ambulance. This study was conducted by the Department of Emergency Medicine. The tables below present the dataset comparing survival rates in victims by EMS methods.

**All accidents**

| | Died | Survived |
|---|---|---|
| Helicopter | 64 | 249 |
| Road ambulance | 260 | 840 |

**Serious accidents**

| | Died | Survived |
|---|---|---|
| Helicopter | 48 | 129 |
| Road ambulance | 60 | 40 |

**Less serious accidents**

| | Died | Survived |
|---|---|---|
| Helicopter | 16 | 120 |
| Road ambulance | 200 | 800 |

1. Using Mantel-Haenszel tests with Woolf method,

   (a) Estimate a Mantel-Haenszel adjusted odds ratio for air EMS and death after adjusting for the seriousness of accidents.

   (b) Test if the survival rate is improved by air EMS after adjusting for the seriousness of accidents. Provide your test statistic along with the p-value. Translate the result into your own scientific explanation & conclusion.

   (c) Determine whether the seriousness of accidents is a confounding factor or an effect modifier in the association between survival rates and EMS types. Provide your test statistic along with the p-value. Translate the result into your own scientific explanation & conclusion.

2. Using multiple logistic regression,

   (a) Determine whether the seriousness of accidents is a confounding factor or an effect modifier in the association between survival rates and EMS types. Please make sure to use likelihood ratio tests. Does this result agree with that of Mantel-Haenszel tests?

   (b) Test goodness-of-fit of your final logistic regression model.

***No formatted dataset is available. Reformatting is a part of analysis.**