STATS 3DA3

Homework Assignment 5

Pratheepa Jeganathan

03/18/2025

Instruction

- **Due before 10:00 PM on Tuesday, April 1, 2025.**
- Upload a **PDF copy** of your solutions to Avenue to Learn. You do **not** need to rewrite the questions in your submission.
- Late Submission Penalty: A 15% deduction per day will be applied to assignments submitted after the deadline.
- Late Submission Limit: Assignments submitted more than 72 hours late will receive a grade of zero.
- **Grace Period for Accommodations**: A 72-hour extension beyond the due date is granted for students with approved accommodations through SAS.
- Your submission must follow the **Assignment Standards** listed below.

Assignment Standards

- Include a **title page** with your **name** and **student number**. Assignments without a title page will not be graded.
- Use Quarto Jupyter Notebook for your work (**strongly recommended**).
- Format your document with an **11-point font** (Times or similar), **1.5 line spacing**, and **1-inch margins** on all sides.

- Use a **new page** for the solution to **each question** (e.g., *Question 1*, *Question 2*, *Question 3*).
  - Clearly **number** all solutions and **sub-parts**.
- Do not include screenshots in your submission; they will not be accepted.
- Ensure your writing and referencing are appropriate for the undergraduate level.
- You may discuss homework problems with other students, but you must prepare and submit your own written work.
- The originality of submitted work will be checked using various tools, including publicly available internet tools.

Assignment Policy on the Use of Generative AI

- The use of **Generative AI** is **not permitted** in assignments, **except** for using **GitHub Copilot** as a coding assistant.
  - If GitHub Copilot is used, you must clearly indicate this in the code comments.
- In alignment with McMaster academic integrity policy, it "shall be an offence knowingly to submit academic work for assessment that was purchased or acquired from another source". This includes work created by generative AI tools. Also state in the policy is the following, "Contract Cheating is the act of"outsourcing of student work to third parties" with or without payment." Using Generative AI tools is a form of contract cheating. Charges of academic dishonesty will be brought forward to the Office of Academic Integrity.

Question

In this assignment, you will explore the logistic regression algorithm and apply it to a dataset from Kaggle. This will provide an opportunity to practice key data science skills, including data retrieval, preprocessing, model fitting, inference, and evaluation.

You will use the diabetes dataset available on Kaggle, which originates from a study conducted by the National Institute of Diabetes and Digestive and Kidney Diseases. The goal is to develop a classification model to predict whether a patient has diabetes based on their diagnostic measurements. Additionally, you will analyze and interpret the logistic regression model to understand its findings.

Link to the dataset: Diabetes dataset on Kaggle

1. Ensure you have a Kaggle account. If not, sign up at kaggle.com. Download the diabetes dataset from Kaggle. Properly cite the dataset in your work.
2. Load the dataset into a pandas DataFrame for analysis.
3. Perform an initial exploration of the dataset to understand its structure (the number of features, observations, and variable types). Write at least three findings from the exploratory data analysis.
4. Generate summary statistics for the dataset, including descriptive statistics for categorical variables if there is any. Provide at least two statements based on the results. Don't include the response variable in this summary statistics.
5. Visualize the distribution of the diabetes outcome variable. Provide at least one statement based on the plot.
6. Check for missing values in the dataset. If any are found, report the number of observations with missing values. Do not remove them from the analysis, as the reason for their absence or irrelevant values is unknown.

Hint: glucose, blood pressure, insulin, and BMI should not be zero.

7. Skip the outlier analysis. Standardize the numerical predictor variables to ensure they are on the same scale.

Hint:

a. Identify the numerical columns in the dataset.

b. Scale the selected columns in (a) and update the DataFrame with the transformed values.

8. Split the dataset into a training set(75%) and a testing set (25%).

Hint:

a. Store only the predictor variables in a DataFrame X before splitting.

b. Store the response variable in y before splitting.

c. Use stratified random sampling to maintain the class distribution.

9. Create an instance of the logistic regression using `scikit-learn`. Change the maximum number of iterations taken for the solvers to converge to 120.

10. Train the model on the training set and generate probability predictions for the test set.

    - Identify the order of categories in the target variable within the predicted probability array.

11. Calculate the test accuracy of the model using a probability cutoff of 0.5 to classify individuals as at risk for diabetes. Discuss the model's performance, such as comparing it to random guessing (e.g., flipping a coin).

12. Evaluate the sensitivity and specificity of the model on the test set using a probability cutoff of 0.5 for diabetes classification.

13. Discuss potential improvements for diabetes prediction based on the test set results, including accuracy and the confusion matrix. Provide at least one statement based on your analysis.

14. Perform ROC analysis on the test set and plot the ROC curve.

    - Additionally, calculate the area under the ROC curve (AUC).
    - Determine the optimal probability cutoff point for diabetes classification.
    - Discuss how changing the cutoff affects sensitivity, specificity, and overall classification performance (at least two statements).

15. The remaining questions will use the selected predictor variables: 'Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', and 'DiabetesPedigreeFunction'.

    - Identify which predictor variable is missing from the model.
    - Fit a logistic regression model using the selected predictor variables (use the maximum number of iterations taken for the solvers to converge to 120).

16. Based on the model in (15), what can you conclude about the missing predictor variable in predicting the risk of diabetes ('Outcome')?

Hint:

a. Calculate the test accuracy of the model without the missing predictor (use the optimal probability cutoff found in (14) for diabetes classification).

b. Compare the test accuracy found in (14) with the model with all predictors to assess the impact of the missing variable.

Grading scheme

1.	1.	Cite the dataset in the References section in the APA citation format [1]
	2.	Code [1]
	3.	Codes [1] and describe three findings [3]
	4.	Codes [1] and two statements [2]
	5.	Codes [1] and one statement [1]
	6.	Codes [1] and answer [1]
	7.	Codes [1]
	8.	Codes [1]
	9.	Codes with appropriate iterations [1]
	10.	Codes [1] and answer for the order of categories [1]
	11.	Codes [1] and statement [1]
	12.	Codes [1]
	13.	statement [1]
	14.	Codes (ROC, AUC, optimal cut-off) [3] and two statements [2]
	15.	answer [1] and codes [1]
	16.	Codes [1] and answer [1]

The maximum point for this assignment is 31. We will convert this to 100%.