STATS/CSE 780 - Homework Assignment 2

Instructor: Pratheepa Jeganathan

2025-09-29

Instruction

- Due before 10:00 PM on Friday, October 10, 2025.
  - Submissions will not be accepted after this due date.

Assignment Standards

Your assignment must comply with the following standards:

- Writing Tools: Use Quarto-Jupyter Notebook for your report and supplemental materials.
- Title Page: Include your name and student number on the title page. Assignments without a title page will not be graded.
- Formatting:
  - Use an 11-point font (Times New Roman or a similar typeface).
  - Apply 1.5 line spacing and ensure 1-inch margins on all sides.
  - This document used these formats.
- Submission Components:
  - Submit a PDF copy of the report (2–3 pages) and supplemental material (less than 10 pages) to Avenue to Learn.
- Collaboration: While you may discuss problems with peers, the final assignment must be your own work.

- Ensure the writing and referencing are appropriate for graduate-level work.
- Various tools may be used to verify the originality of submitted work, including publicly available internet tools.

Report:

- The report must not exceed 3 pages, including tables and figures.
- Choose figures and tables judiciously for the report; additional figures and tables may be included in the supplemental material and referenced in the report.
- You may include one page for the `bibliography` and one page for the `title page`.
- `Python code are not` allowed in the report.

Supplemental Material:

- Include only Python code that generated the results discussed in the report.
- Ensure that the PDF does not display chunk messages, warnings, or extended data frames.
- `Screenshots of Python code are not` allowed in the supplemental material.

Generative AI Policy

- Generative AI is not permitted in this assignment, except for the use of `GitHub Copilot as an assistant for coding`.
- Clearly indicate in the code comments where GitHub Copilot was used as a coding assistant.
- 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.

Objective:

The goal of this assignment is to apply classification techniques to a real-world dataset, evaluate their performance, and draw meaningful conclusions. You will explore multiple logistic regression, k-nearest neighbors (kNN), and locally adaptive nearest neighbor classification methods.

In addition, you are also required to reflect on the use of generative AI for answering one of the questions in this assignment.

Dataset Selection Requirements:

- Find a dataset that is suitable for classification. The response variable may have more than two categories; in such cases, you may need to apply a suitable transformation for some classifiers, or adapt the method to handle multiclass classification.
- Some sites for dataset search are
  - Kaggle Datasets,
  - UCI Machine Learning Repository.
- **Do not** use any of the following:
  - Datasets used in class or in the assigned textbook(s),
  - Datasets from R or Python package libraries
  - Private datasets from your own research or work experience (i.e., datasets not publicly accessible).
- The dataset must have at least four predictor variables, including a nominal or ordinal predictor variable and qualitative response variable.

Assignment Questions:

1. Dataset Sourcing:
   - Briefly describe how the dataset was sourced and describe the chosen dataset (what are variables and observations).
2. Application Description:

- Clearly describe the application you are investigating based on the selected dataset.

3. Exploratory data analysis:

   - Produce numerical and graphical summaries of the data. Examine patterns and provide details such as:
     - Data types of the variables,
     - Data summaries,
     - Correlation and association analysis,
     - Outliers and missing values analysis.

4. Split the data into a training, validation, and test sets.

   - Describe the rationale behind your choice of data splitting.

5. Perform the following methods using the training and validation sets: logistic regression and k-nearest neighbor. For each classifier:

   (i) Describe if you used any transformations or selected a subset of predictors or categories.
   (ii) Describe the choice of tuning parameters (if any).
   (iii) Identify the most important predictor variable(s) in classifying the response, or explain why this cannot be determined.

6. Model evaluation:

   (i) Use the test set to evaluate the performance of the classifiers using the miss-classification error rate. - If the classifier needs a cutoff to classify the labels, use sensitivity and specificity analysis to determine the cutoff.
   (ii) Compare and contrast the performance of the classifiers. Provide at least two statements.

7. Perform kNN with an alternative nearest neighbor search strategy or distance metric learning method (different from what you used in (5)), using the training and validation sets. Example Python libraries for implementing alternative strategies include: 1. https://pypi.org/project/pyDML/ 2. https://github.com/scikit-learn-contrib/metric-learn

   (i) Describe the search strategy or distance metric learning method you used, including any tuning parameters (if applicable).

(ii) Compare and contrast the kNN performance with and without the new search strategy or distance metric learning method, using the test set.

8. State at least two conclusions based on your analysis in (4)–(7). These conclusions should be clearly connected to your selected dataset and application.

9. Can generative AI be used to answer one of the questions in (1)–(8)??

   (i) Provide the prompt(s) used to answer the question.

   (ii) Discuss whether generative AI tools could assist in completing any parts of this assignment.

Grading scheme

1.		source of dataset [1] describe variables and observations [1]
2.		describe the application [1]
3.		statistical summaries - no points if the graphs and tables are not readable [1] number of observations versus variables [1] data types [1] correlation, association analysis - no points if the graphs and tables are not readable [1] outliers detection and handling - no points if the graphs and tables are not readable [1] missing value detection and handling - no points if the graphs and tables are not readable [1]
4.		describe the rationale for data splitting [1]
5.	(i)	data and statistical transformation or subset of predictors for each classifier [2]
	(ii)	Choice of tuning parameters for each classifier & appropriate use of training and validation sets [2]
	(iii)	the most important predictor for each classifier [2]
6.		(i) Compute the appropriate error for each classifier [2] (ii) Compare and contrast the performance of the classifiers (at least two statements) - no points if the graphs and tables are not readable [2]
7.	(i)	Describe the nearest neighbor search startegy or distance metric learning [2]
	(ii)	Compare and contrast (at least one statement) [1]
8.		At least two conclusions drawn from your analysis [2]

9.	(i) Provide prompts and (ii) discussion [2]
References	Reference list starts on a new page, references are appropriate and list out in the report (at least two references) [2]
Supplementary material	Supplementary material starts on a new page [1], code readability [1], the Python codes and the outputs for the questions are presented [1], all codes are within the margins (we don't grade the supplementary material if too many codes are outside of the margins)

The maximum point for this assignment is 32. We will convert this to 100%.

Some datasets that may be used for this assignment

1. Hotel reservation in Kaggle: https://www.kaggle.com/datasets/ahsan81/hotel-reservations-classification-dataset/data
2. Loan approval in Kaggle: https://www.kaggle.com/datasets/rizqi01/ps4e9-original-data-loan-approval-prediction/data
3. Subscription status of Spotify in Kaggle: https://www.kaggle.com/datasets/nabihazahid/spotify-dataset-for-churn-analysis/data
4. Census income in UCI: https://archive.ics.uci.edu/dataset/2/adult