ELG 5225: Applied Machine Learning

Assignment 2

Due date posted in Brightspace

Max mark: 100

Submission Instructions

You must submit two separate documents for this assignment:

- **Report:** Submit a report in PDF, `.doc`, or `.docx` format, including solution explanations and important code snippets. Please ensure that:
  - The document is machine-readable, meaning it contains selectable and searchable text rather than scanned images or screenshots.
  - Code snippets are presented in their original, editable text format (not as images), allowing for proper indexing and plagiarism verification.
  - All figures have clear captions in text format (e.g., below each figure), providing a machine-readable description of each figure's content. The caption should go beyond a simple title, if necessary, to clarify the figure's purpose.
- **Code File:** Submit the entire code separately, preferably in a `.ipynb` (Jupyter Notebook) format, as it offers added transparency for any disputed points, such as intermediate calculations or code outputs. If a `.ipynb` file is not possible, a `.py` file is also acceptable. The file name must include your group number and assignment number, e.g., `Group1_HW3.ipynb` (or `Group1_HW3.py`) for the code, and `Group1_HW3.docx` (or `.pdf` if preferred) for the report.

These requirements (machine-readable report, readable code snippets, and textual captions for figures) are essential to facilitate thorough examination by plagiarism detection software. Additionally, the `.ipynb` format provides clear, inline output and context for each step, making it easier to verify specific points in your work if needed.

Assignments must be submitted online through Brightspace only. No email submissions will be accepted, and assignments not submitted on Brightspace cannot be graded. Ensure proper submission before the deadline, as late submissions are not accepted.

Part 1: Calculations

1. Use the k-means algorithm and Euclidean distance to cluster the following 5 data points into 2 clusters: $A_1 = (3, 6)$, $A_2 = (6, 3)$, $A_3 = (8, 6)$, $A_4 = (2, 1)$, $A_5 = (5, 9)$. Suppose that the initial centroids (centers of each cluster) are A_2 and A_4. Using k-means, cluster the 5 points and show the following for one iteration only:

(a) Show step-by-step calculations for clustering the 5 points. **(7 Marks)**

(b) Draw a 10x10 space, marking the clustered points and new centroid coordinates. (4 Marks) **(4 Marks)**

(c) Calculate the silhouette score and WSS score. **(5 Marks)**

Part 2: Programming

1. Use scikit-learn to implement Naive Bayes (NB) and K-Nearest Neighbor (KNN) classifiers on the provided Mobile Crowd Sensing (MCS) dataset. Use the following features: Latitude, Longitude, Day, Hour, Minute, Duration, RemainingTime, Resources, Coverage, OnPeakHours, GridNumber, and Legitimacy. The Legitimacy column is the target, and the Day column will be used to split the dataset into training and test sets. Use Day values 0, 1, and 2 for training, and value 3 for testing. ID indicates the index and cannot be used as a feature. Please use color code and different line style in your figures.

   (a) Create training and test datasets based on the Day feature. **(6 Marks)**

   (b) Provide confusion matrices and F1 scores for both NB and KNN classifiers as baseline performances. **(6 Marks)**

   (c) Generate 2D TSNE plots, one for the training set and one for the test set. **(4 Marks)**

Part 3: Dimensionality Reduction and Analysis

1. Apply PCA and Autoencoder (AE) for dimensionality reduction.

   (a) Plot F1 scores against the number of components (dimension) for PCA and AE using NB and KNN classifiers. The baseline performances should be included for comparison. Plot four graphs in total. **(16 Marks)**

   (b) Generate 2D TSNE plots for the best-performing dimensionality reduction technique, one for the training set and one for the test set. **(4 Marks)**

Part 4: Feature Selection Exploration

1. Use feature selection methods to identify optimal feature subsets for improved classification.

   (a) Apply Filter Methods (e.g., Information Gain or Variance Threshold) and plot the number of features vs F1 score with baseline performance. **(8 Marks)**

   (b) Apply Wrapper Methods (e.g., Forward Selection, Backward Elimination) and plot the number of features vs accuracy with baseline performance. **(8 Marks)**

   (c) Generate 2D TSNE plots for the best feature selection method, one for the training set and one for the test set. **(4 Marks)**

Part 5: Clustering of Geographic Features

1. Use latitude and longitude features for clustering.
   (a) Apply K-means clustering to plot the number of clusters vs legitimate-only members in clusters. **(8 Marks)**

   (b) Apply SOFM clustering and plot the number of clusters vs legitimate-only members. **(8 Marks)**

   (c) Apply DBSCAN clustering, adjusting midpoint and epsilon parameters to obtain approximately 5 cluster numbers, and report results. **(8 Marks)**

Part 6: Conclusion

1. List conclusions for each question, with at least one conclusion per question. **(4 Marks)**

Important Notes

- The report should include all answers briefly. All plots must have titles and proper axis labels. Missing items will result in a deduction of one point per item.
- Use `random_state=0` when necessary for reproducibility.