Semester 1 - 2024

CSC 3831 Final Assignment

Brief

This assignment consists of three tasks: data engineering (30%), machine learning (30%), and computer vision (40%). The components will be released sequentially as the course progresses.

1. Modify the provided iPython files or create your own iPython files which perform the required tasks
2. Clearly outline in each section where you accomplish the tasks required by the assignment
3. Include written analysis of the results/exploration either in a separate document or in a text block in the iPython document
4. Benefits/detriments of the algorithms used
5. Citations of rationale (either from course material, in canvas supplemental material, or found online)

Citations can be taken from online material (i.e., articles, papers, etc.), the supplemental material included at the bottom of each lecture, or directly from the lecture itself.

Deliverables

We require that you submit:

1. A single zipped folder containing all of the iPython notebooks, pdfs, and any other files requrested for submission.

It does not matter if you include your written explanation/descriptions within the text or in a separate document (whichever you feel more comfortable with), but if it is included in the iPython file please clearly demarcate your code and explanations.

Part I – Data Engineering **[30]**

1. **Data Understanding** [**7**]
   - Perform ad hoc EDA to understand and describe what you see in the raw dataset
     - Include graphs, statistics, and written descriptions as appropriate
   - Identify features with missing records, outlier records
2. **Outlier Identification** [**10**]
   - Utilise a statistical outlier detection approach to (i.e., **no** KNN, LOF, 1Class SVM)
   - Utilise an algorithmic outlier detection method of your choice
   - Compare results and decide what to do with identified outliers
     - Include graphs, statistics, and written descriptions as appropriate
   - Explain what you are doing, and why your analysis is appropriate
   - Comment on benefits/detriments of statistical and algorithmic outlier detection approaches
3. **Imputation** [**10**]
   - Identify which features should be imputed and which should be removed
     - Provide a written rationale for this decision
   - Impute the missing records using KNN imputation
   - Impute the missing records using MICE imputation
   - Compare both imputed datasets feature distributions against each other and the non-imputed data
   - Build a regressor on all three datasets
     - Use regression models to predict house median price
     - Compare regressors of non-imputed data against imputed datas
   -
4. **Conclusions & Thoughts** [**3**]
   - Discuss methods used for anomaly detection, pros/cons of each method
   - Discuss challenges/difficulties in anomaly detection implementation
   - Discuss methods used for imputation, pros/cons of each method
   - Discuss challenges/difficulties in imputation implementation

Part II – Machine Learning [**30**]

1. TBD

Part III – Computer Vision [**40**]

1. TBD

Submission

This assignment is worth 100% of the overall grade for CSC3831. Submissions must include both the code for implementation and rationale/explanation as described in Deliverables.

Referencing

You must follow the IEEE referencing style details which can be found here: https://ieeeauthorcenter.ieee.org/wp-content/uploads/IEEE-Reference-Guide.pdf

Plagiarism

Work which is submitted for assignment must be your own work. Plagiarism means presenting the work of others as though it were your own. Further details about the university policy on plagiarism can be found here:

https://www.ncl.ac.uk/academic-skills-kit/good-academic-practice/plagiarism/

Marking and Feedback

Marks will be given for all tasks and sub-tasks. Please note that the marking will not be linear, which means achieving higher grades will be increasingly challenging and should meet the expectations (and beyond) of the marking scheme.

Written feedback will be provided along with the marks within 20 working days of the submission deadline. If you are not satisfied with the marks and feedback, you are welcome and encouraged to discuss individually with Mr. Dixon, Dr. Gonzalez, and Dr. Ojha.