FINAL EXAM

AML 3104- Neural Networks and Deep Learning

Final Exam Guidelines

Questions:

Follow the following 5 stages of on the data set and express your results:

1. Preliminary Data Analysis (20%)
2. Data Processing & Feature Engineering (20%)
3. Modelling (35%)
    a. X1 Conventional ML Model (11%)
    b. X3 Deep Networks (24%)
4. Evaluation (15%)
5. Conclusion (10%)

Guiding Principles:

- **Preliminary Data Analysis:** Understanding behavior of features against one another (Multi collinearity) , ranges of the inputs, anomalies in your data and assumptions you made to simplify or improve model accuracy
- **Data Processing & Feature Engineering:** Potential normalization, development/removing features affecting the overall performance of the model, encoding if applicable
- **Modeling:** On the modeling section, you are required to deliver:
    1. Conventional Machine Learning model of your choice fine-tuned using grid-search techniques to yield best results. This model will be used as a baseline to compare results against Neural Network.
    2. X3 Deep Sequential Neural network with dimensions of your choice. Purpose of this section is to evaluate your competency in parameter tuning of neural networks to obtain better results.
- **Evaluation:** Thought process on your approach on improving individual models (How you improved individual models by fine-tuning relevant hyper parameters). Some key objectives for evaluating your models are training vs. validation set, training vs. test set, does your model have bias or variance? If so, how can you tell?

- **Conclusion of your work:**

    1. Step by step process of how you tackled the problem
    2. How were the models compared with one another and why?
    3. What do you think is the most important feature in your dataset

Wine Quality Data Set – Red Wine

Link to the dataset: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

Group 1:

Roshan Acharya	C0831342@mylambton.ca
Rishi Phaneendra Varma Bhupathi Raju	C0825285@mylambton.ca
Kamaldeep Kaur Dhami	C0826633@mylambton.ca
Latharani Radhakrishnan	C0833847@mylambton.ca

Dataset Description:

The red wine datasets are related to red variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

Wine Quality Data Set – White Wine

Link to the dataset: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/

Group 2:

Davinderjit Singh	C0833117@mylambton.ca
Onyinye Mbanefo	C0831578@mylambton.ca
Hukamdeep Singh	C0834020@mylambton.ca
Tania Tangri	C0828053@mylambton.ca

Dataset Description:

The white wine datasets are related to white variants of the Portuguese "Vinho Verde" wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are munch more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

US Honey Production 1995-2021

Link to the dataset: **https://www.kaggle.com/datasets/mohitpoudel/us-honey-production-19952021**

Group 3:

Ganesh Chaulagain	C0838561@mylambton.ca
Milanjeet Kaur	C0829899@mylambton.ca
Bibek Shah Shankhar	C0835648@mylambton.ca
Tika Thapa	C0832724@mylambton.ca

Dataset Description:

Honey is the natural product made by bees—one of our planet's most important species. Honey bees visit millions of blossoms in their lifetimes, making pollination of plants possible and collecting nectar to bring back to the hive to produce honey.

Honey bees use honey as their primary energy source and their instinct is to make more than their colony needs. Beekeepers harvest the excess and bottle it for consumption, just like they've been doing since the beginning of time. Harvesting honey is good for the bees and also part of what makes successful beekeeping businesses.

This dataset provides insights into honey production, demand, and supply across different states of America.

We would like to predict the average cost of honey (As our dependent variable) against other features listed in the dataset.

Music & Mental Health Survey

Link to the dataset: https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results

Group 4:

Priti Bhale	C0835691@mylambton.ca
Jyoti Shukla	C0817905@mylambton.ca
Vindhya Reddy Teegapuram	C0833791@mylambton.ca
Greeshma Gayathri Yarlagadda	C0834358@mylambton.ca

Dataset Description:

Music therapy, or MT, is the use of music to improve an individual's stress, mood, and overall mental health. MT is also recognized as an evidence-based practice, using music as a catalyst for "happy" hormones such as oxytocin.

However, MT employs a wide range of different genres, varying from one organization to the next.

The MxMH dataset aims to identify what, if any, correlations exist between an individual's music taste and their self-reported mental health. Ideally, these findings could contribute to a more informed application of MT or simply provide interesting sights about the mind.

In this problem we are trying to understand the effect of music on mental health. This effect is recorded in the corresponding column “Music effects”.

Students Performance in Exams

Link to the dataset: **https://www.kaggle.com/datasets/whenamancodes/students-performance-in-exams**

Group 5:

Piyush Bhatia	C0827347@mylambton.ca
Nikhilesh Shyamkumar Koshti	C0828136@mylambton.ca
Meet Anilkumar Patel	C0827470@mylambton.ca
Shreyas Nithyanand Shetty	C0834023@mylambton.ca

Dataset Description:

This data set consists of the marks secured by the students in various subjects. The objective of this problem is to determine the overall success of student in all areas of math, reading and writing based on other features indicated in the dataset. Since there are three dependent variables involved as the output, think about how can you perform feature engineering to fit this into your model.

E-Commerce Shipping Data

Link to the dataset https://www.kaggle.com/datasets/prachi13/customer-analytics

Group 6:

Sagar Dahiya	C0833880@mylambton.ca
Palwinder Kaur	C0827804@mylambton.ca
Ajay Pal Singh	C0828307@mylambton.ca
Nimmo Usman	C0836309@mylambton.ca

Dataset Description:

An international e-commerce company based wants to discover key insights from their cust omer database. They want to use some of the most advanced machine learning techniques to study their customers. The company sells electronic products.

Content

The dataset used for model building contained 10999 observations of 12 variables. Given these variables, we would like to predict if the shipment is going to reach to destination on time or not. There is a specific column in the data set that you can use as your output (Reached on time)

Personal Key Indicators of Heart Disease

Link to the dataset https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Group 7:

Kirandeep Kaur	C0833848@mylambton.ca
Sujit Khatiwada	C0835126@mylambton.ca
Anjana Kuriakose	C0829580@mylambton.ca
Venkata Sai Manikanta Ponakala	C0833772@mylambton.ca

Dataset Description:

According to the CDC, heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, and white people). About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator include diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare. Computational developments, in turn, allow the application of machine learning methods to detect "patterns" from the data that can predict a patient's condition.

What we'd like to do is to predict HeartDisease (Binary classification) based on the ~20 variables listed in this dataset.

Billion-Dollar Disasters Dataset

Link to the dataset https://www.kaggle.com/datasets/michaelbryantds/billiondollar-disasters

Group 8 (Team Neuron):

Aadarsha Chapagain	C0825975@mylambton.ca
Sreya Treesa Johny	C0829268@mylambton.ca
Rajasekhar Katta	C0833766@mylambton.ca
Anirudh Siddula	C0830486@mylambton.ca

Dataset Description:

As “the Nation’s Scorekeeper in terms of addressing severe weather and climate events in their historical perspective,” the US National Centers for Environmental Information maintains an inventory of the most costly such disasters in the US — those that have caused at least $1 billion in estimated direct losses. The quarterly-updated dataset contains more than 330 severe storms, floods, droughts, wildfires, freezes, and other extreme events since 1980.

What we’d like to do is to predict the cost of a disaster based on features available in the dataset. This dataset, while being simple, yields a decent opportunity for feature engineering. So It might be worthwhile to gain more insight from this dataset