10 Academy Batch 4 - Weekly Challenge: Week 7

Industry - Casualty Challenge

Overview

Business Need

A common frustration in the industry, especially when it comes to getting business insights from tabular data, is that the most interesting questions (from their perspective) are often not answerable with observational data alone. These questions can be similar to:

- "What will happen if I halve the price of my product?"
- "Which clients will pay their debts only if I call them?"

Judea Pearl and his research group have developed in the last decades a solid theoretical framework to deal with that, but the first steps toward merging it with mainstream machine learning are just beginning.

The causal graph is a central object in the framework mentioned above, but it is often unknown, subject to personal knowledge and bias, or loosely connected to the available data. The main objective of the task is to highlight the importance of

the matter in a concrete way. In this spirit, trainees are expected to attempt the following tasks:

1. Perform a causal inference task using Pearl's framework;
2. Infer the causal graph from observational data and then validate the graph;
3. Merge machine learning with causal inference;

The first is straightforward, the second and third are still open questions in the research community, hence may need a bit more research, innovation, and thinking outside the box from trainees.

Data

You can extract the data from kaggle or from UCI Machine Learning Repository. In the latter you can find even more data that you may explore further. To understand more about the data, and how it is collected we highly recommend reading this paper: (PDF) Breast Cancer Diagnosis and Prognosis Via Linear Programming (researchgate.net).

Features in the data are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

Attribute Information:

1. ID number
2. Diagnosis (M = malignant, B = benign)
3. The remaining (3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) Perimeter
d) Area
e) smoothness (local variation in radius lengths)
f) compactness (perimeter^2 / area - 1.0)
g) concavity (severity of concave portions of the contour)
h) concave points (number of concave portions of the contour)
i) Symmetry

j) fractal dimension ("coastline approximation" - 1)

The mean, standard error and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recorded with four significant digits.

Missing attribute values: none

Class distribution: 357 benign (not cancer), 212 malignant (cancer)

Expected Outcomes

Skills:

- Modeling a given problem as a causal graph
- Statistical Modelling and Inference Extraction
- Building model pipelines and orchestration

Knowledge:

- Knowledge about casual graph and statistical learning
- Hypothesis Formulation and Testing
- Statistical Analysis

Competency Mapping

The tasks you will carry out in this week's challenge will contribute differently to the 17 competencies 10 Academy identified as essential for job preparedness in the field of data science, and Machine Learning engineering. The mapping below shows the change (lift) one can obtain through delivering the highest performance in these tasks.

MC0: Marginal contribution - causes no significant change
MC1: Minor contribution - recognized for routine performance gain
MC2: Measurable contribution - will contribute a value towards portfolio and job readiness metric

MC3: Major contribution - the best performance of these types of tasks at least three times within our training leads one to attain a job-ready level along that competency dimension.

Competency	Value	Potential contributions from this week
Business Understanding	MC3	Understanding and reasoning the business context. Thinking about suitable analysis that matches the business need. Thinking about clients and their interests.
Data Engineering	MC1	Thinking about how to store data for easy analysis, and what format to use to build responsive dashboards.
Data Understanding	MC3	Understanding the data provided and extract insight. Exploring different techniques, algorithms, statistical distributions, sampling, and visualization techniques to gain insight.
Dashboard & Visualization	MC2	Building a dashboard to explore data as well as to communicate insight. Advanced use of modules such as plotly, seaborn, matplotlib etc. to build descriptive visualizations. Reading through the modules documentation to expand your skill set.
Mathematics and Statistics	MC3	Thinking about statistical distributions, sampling, bias, overfitting, correlations.
MLOps & Continuous Delivery	MC1	Using Github for code development, thinking about feature store, planning analysis pipeline, using MLOps tools for code, data, model, artifact versioning,

		setting up docker containers for automated microservice deployment.
Modeling and evaluation	MC3	Comparing multiple Deep learning techniques; training and validating DL models; choosing appropriate architecture, loss function, and regularisers; hyperparameter tuning; choosing suitable evaluation metrics.
Python programming	MC3	Advanced object-oriented python programming. Python package building.
SQL programming	MC2	Building feature stores using SQL or NoSQL databases.
Fluency in the Scientific Method	MC3	Thinking about evidence. Generating hypothesis, testing hypothesis. Thinking about different types of errors.
Ethics	MC1	data privacy, data security, ethical use of data. The 8 principles of responsible machine learning
Statistical & Critical Thinking	MC3	Thinking about the difference between causal vs chance correlation. Giving reasonable recommendations. Thinking about uncertainties.
Software Engineering & Dev Environment	MC1	Reading articles on software project planning. Unit testing.
Impact & Lifelong learning	MC3	Learning new concepts, ideas, and skills fast, and applying them to the problem at hand.
Professional Culture & Communication	MC3	Writing a well-formatted presentation with no mistakes, formatted nicely.

Social Intelligence & Mentorship	MC3	Asking for help early, providing help for those who need it, avoiding being stuck.
Career Thinking	MC3	Working within groups in a successful way

Team

Instructors: Yabebal, Abubakar, Mahlet, Kevin

Group Work Policy

This submission is to be done individually or in groups of upto 4 people. Collaborative learning is highly encouraged. We believe that groups of 2-3 people are likely to be most effective. If you would like to join a group, you can ask the 10 Academy team to help match you into a group. We recommend forming groups right away, ideally by Monday 23 Aug 2021 1300UTC.

Key Dates

- **Discussion on the case** - 10:00 UTC time on Monday 23 August 2021. Use #all-week7 to ask questions.
- **Interim Submission** - 8:00PM UTC time on Wednesday 25 August 2021.
- **Final Submission** - 8:00PM UTC time on Saturday 28 August 2021

Leaderboard for the week

There are 100 points available for the week.

20 points - community growth and peer support.

13 points - technical public and group-based RC channels

- Total number of messages (5)
- Total number of Mentions (3)
- Total number of DM connections (5)

7 points - community activities

- Number of messages in non-technical channels (4)
- On-time presence in Gmeet sessions (3)

30 points - presentation and reporting.

15 points - interim submission. PDF slide or report format. Evaluated for:

- Overview of Causal inference and its distinction from statistical inference (5)
- Basic explanation of causal graph, and its elements (4)
- Discussion of your data exploration result (3)
- Report of what is completed. (3)

15 points for the final submission. Blog entry or PDF with 5-8 pages.

- Abstract, review, and data exploration and representation (5).
- Discussion on causal graph, and conclusion of your analysis showing clearly the insights you derived (10),

50 points - Technical content

20 points - Interim submission

1. Github link submission (20)
    - Object-oriented programming (5)
    - Git issues or project that shows your work plan (5)
    - Jupyter notebook that illustrate you data exploration and feature engineering with (Professional plot production code, readable axes labels, title, and legend; good choice of color) (3)
    - Unit testing (3)
    - Screenshot of DVC data versions (3)
    - Screenshot of MLFlow dashboard (3)

30 points - Final submission

- MLOps setup (10)
    - Working CML git workflow implementation in repo (5)
    - Screenshot of example CML report when git push a code (5)
- Github Link submission (25)
    - Working scripts and notebooks to build causal graphs (15 points)
    - Well explained readme (5)
    - SQL database integration (5)

Badges

Each week, one user will be awarded one of the badges below for the best performance in the category below.

In addition to being the badge holder for that badge, each badge winner will get +20 points to the overall score.

Visualization - the quality of visualizations, understandability, skimmability, choice of visualization
Quality of code - reliability, maintainability, efficiency, commenting - in the future this will be CICD/CML
An innovative approach to analysis -using latest algorithms, adding in research paper content and other innovative approaches
Writing and presentation - clarity of written outputs, clarity of slides, overall production value
Most supportive in the community - helping others, adding links, tutoring those struggling

The goal of this approach is to support and reward expertise in different parts of the Machine learning engineering toolbox.

Late Submission Policy

Our goal is to prepare successful learners for the work and submitting late when given enough notice, shouldn't be necessary.

For interim submissions, those submitted 1-6 hours late will receive a maximum of 50% of the total possible grade. Those submitted >6 hours late may receive feedback, but will not receive a grade.

For final submissions, those submitted 1-24 hours late, will receive a maximum of 50% of the total possible grade. Those submitted >24 hours late may receive feedback, but will not receive a grade.
When calculating the leaderboard score:

- From week 8 onwards, your two lowest weeks' scores will not be considered.

Instructions

The fundamental tasks in this week's challenge are the following;

- Read the data
- Perform exploratory analysis on it
- Extract features and scale the extracted feature
- Split the data into training and hold-out set
- Create casual graph using different technique
- Examine the model performance based on the graph

Task 1: Data Exploration

1. Conduct an exploratory data analysis on the data & communicate useful insights. Ensure that you identify and treat all missing values and outliers in the dataset by using appropriate methods.
2. Perform feature extraction and scaling

Task 2: Casual learning

1. Split data into training and hold-out set
2. Create a causal graph using all training data and get the insights (this will be considered the ground truth)
3. Create new causal graphs using increasing fractions of the data and compare with the ground truth graph
    a. The comparison can be done with a **Jaccard Similarity Index**, measuring the **intersection** and **union** of the graph edges
4. After reaching a stable causal graph, select only variables that point directly to the target variable
5. Train one model using all variables and another using only the variables selected by the graph
6. Measure how much each of the models overfit the hold-out set created in step 1.

Submission

Interim: Due Wednesday 25.08 8pm UTC

1. A pdf file that helps others understand the main concepts behind causal inference (5-8 pages). This should include:
    a. Literature review
    b. Overview of the data source and formats
    c. Techniques used to perform causal inference
    d. Insights derived from analysis carried out
2. Github link submission that demonstrates
    a. Object-oriented programming
    b. Data exploration
    c. Workflow plan through Github issues

Final: Due Saturday 28.08 8pm UTC

1. A pdf file that explains your understanding of causal inference (max 3 Pages). This should include:
    - Literature review
    - Overview of the data source and formats
    - Techniques used to perform causal inference
    - Insights derived from analysis carried out
2. Github link submission that demonstrates
    - Object-oriented programming
    - Notebook(s) that shows data exploration and inference derivation
    - Reusable scripts that can be used in other similar projects

References:

Notebooks & Github codes

- Quickstart Notebook for using Causalgraphicalmodels python module: used to describe and manipulate Causal Graphical Models and Structural Causal Models.
- Introduction to CasualgraphicModel

Key Papers & blogs

- If correlation doesn't imply causation, then what does?
- DoWhy: An End-to-End Library for Causal Inference (arxiv) - DoWhy package paper
- Mini course on Causality, Cambridge MIT
- Controlling Confounding Bias
- Slides on Causality
- Causality lecture note
- Confounding Bias
- Analysis of Breast Cancer Detection Using Different Machine Learning Techniques | SpringerLink
- ANALYSIS OF FEATURE SELECTION WITH CLASSFICATION: BREAST CANCER DATASETS

Talks & Videos

- Tutorial Session B – Causes and Counterfactuals: Concepts, Principles and Tools (Microft, 2014)
- Plenary 2: The Mathematics of Causal Inference: with Reflections on Machine Learning

General

- Causality by Judea Pearls 2nd edition
- Causal inference in statistics: An overview (2019)
- Main reference

Wikipedia

- Causal Graphical Models
- Structural Causal Models
- Rubin causal model - Wikipedia
- Instrumental variables estimation - Wikipedia

-