Written examination paper for PGR210 – Machine Learning and Natural Language Processing
Department of Technology, Kristiania University College
Autumn 2023

Examination paper released: 09.11.2022 (See Wiseflow if there is any change)

Examination deadline: 07.12.2022 (see Wiseflow if there is any change)

Academic contact during examination: Huamin Ren, huamin.ren@kristiania.no; Arvind Keprate, arvindke@oslomet.no

Technical contact during examination: eksamen@kristiania.no

Exam type: Written home examination in groups (2-5 students)

Support materials: All support materials are allowed

Final report format: use LaTeX or Word and font 12 with 1.5 spacing. The limit is 20 pages max that includes abstract, report, list of bibliography in the end, figures and tables. Both parts (ML and NLP should be in the same PDF report)

An example of possible report structure. Abstract, Introduction, State of the Art literature, Your used approach/Methodology, Analysis of results, Conclusions

Grading scale: Norwegian grading system using the graded scale A - F where A is the best grade, E is the lowest pass grade and F is fail

Weighting: 100% or overall grade

Plagiarism control: We expect your own independent work. Please, use citations and quotations in case if there is a material you want to include in the report.

Learning outcomes as per course description:
Knowledge. The student

- can understand the basic data structures and algorithms for machine learning.
- knows the mathematical concepts underlying the design and analysis of machine learning techniques appropriate for a given data science problem
- has insights into the strengths and weaknesses of Dimensionality Reduction Algorithms: variance thresholds, correlation thresholds, principal component analysis (PCA), linear discriminant analysis (LDA)
- can explain the basic natural language processing concepts and techniques for text analytics
- understands the basic pipeline for natural language processing; for a simple topic modelling task, is able to carry out step-by-step processing, representation and come out with a solution
- knows text processing and analytics methods (such as tokenization, word representation, topic modelling and clustering) for various data science domains

Skills. The student

- can select appropriate machine learning methods (such as linear models, classification models, text classification, semantic textual similarity, word sense disambiguation and neural language models) and tools for a given data science problem
- can analyze mathematically the performance of machine learning methods and techniques
- can apply techniques from the course to new data science problems in terms of selection of appropriate machine learning methods, techniques and tools
- can use python or similar to implement machine learning methods and techniques
- can discuss concepts and applications of machine learning (including text)

Competence. The student

- can differentiate the suitability and efficiency of programs in terms of the machine learning methods and techniques (incl. text analytics) employed
- can apply the knowledge of and skills in machine learning in various data science domains
- can critically reflect on the tradeoffs in the design and implementation of machine learning methods and techniques.

Exam Task

1. Machine Learning

1.1 Clasification Task (15 Marks):

The given “Health Monitoring”dataset has 3000 rows ,4 features (Heart_Rate, Blood_Pressure, Cholesterol, Blood_Sugar), a target variable (Risk_Level) and some missing values. The target dataset has 3 classes ('Low', 'Medium', or 'High'). Perform the following tasks:

a.) Data Preprocessing (4 marks)

- Handle any missing values present in the dataset. Describe the method you chose and justify your choice. (2 marks)
- Normalize the features of the dataset. (1 mark)
- Split the data into training and test sets. (1 marks)

b.) Model Building (5 marks)

- Select a suitable classification algorithm for the given dataset. Justify your choice. (2 marks)
- Train the model using the training set. (2 marks)
- Evaluate the model's performance on the test set and report the accuracy. (1 mark)

c.) Analysis (6 marks)

- Visualize the distribution of the three classes in the dataset. (2 marks)
- Discuss any patterns or insights you can derive from the dataset. (2 marks)
- Provide any recommendations or suggestions for improving the classification results. (2 mark)

1.2 Clustering Task (15 Marks):

The "Urban Mobility Patterns" dataset provides insights into the daily commuting patterns of individuals in a metropolitan city. The dataset consists of 5,000 entries, 4 features (Average_Speed, Waiting_Time, Daily_Commute_Distance, Traffic_Congestion_Score) each representing an individual's daily mobility metrics. The objective is to cluster individuals based on these metrics to identify distinct mobility patterns and provide insights into urban transportation challenges.

Perform the following tasks:

a.) Data Preprocessing (5 marks)

- Handle any missing values present in the dataset. Describe the method you chose and justify your choice. (2 marks)
- Normalize the features of the dataset. (2 marks)

- Determine an appropriate number of clusters for the data using a suitable method. Justify your choice. (1 mark)

b.) Model Building (5 marks)

- Select a suitable clustering algorithm for the given dataset. Justify your choice. (2 marks)
- Apply the clustering algorithm to the dataset. (3 marks)

c.) Analysis (5 marks)

- Visualize the clusters formed. (2 marks)
- Discuss any patterns or insights you can derive from the clusters regarding urban mobility. (2 marks)
- Provide any recommendations or suggestions based on the clustering results. (1 mark)

1.3 **Deep Learning Based Regresion Task (20 Marks):**

The "House_Price_Prediction" dataset is designed to represent properties with features like size, number of bedrooms, number of bathrooms, age, and proximity to the city center. The target variable is the property price. Given the dataset, perform the following tasks to predict the price of properties based on various features.

a.) Data Preprocessing (4 marks)

- Handle any missing values present in the dataset. Describe the method you chose and justify your choice. (2 marks)
- Normalize the features of the dataset. (1 marks)
- Split the data into training, validation, and test sets. (1 marks)

b.) Model Building (10 marks)

- Design a deep learning model suitable for regression tasks. Describe the architecture, including the number of layers, types of layers, and activation functions. (6 marks)
- Train the model using the training set and validate it using the validation set. (2 marks)
- Evaluate the model's performance on the test set and report the mean squared error. (2 marks)

c.) Analysis (6 marks)

- Visualize the distribution of actual vs. predicted property prices. (2 marks)
- Discuss any patterns or insights you can derive from the model's predictions. (3 marks)
- Provide any recommendations for improving the model's performance. (1 mark)

2. Natural Language Processing

2.1. Practical Task: Text processing, feature extraction and representation by using both TF and TF-IDF schemes (10 Marks):

Given a data file (Exam_NLP.csv), where reviews on movies are provided.

1. Data preparation: load the file, access the columns, then through printing and visualization, understand the meaning in each column. Then create a new column, name it as ‘description’ by concatenating the strings from two columns: tagline and overview.

2. Text processing: convert words in ‘description’ to lower case, remove white space, remove words from stop_words (from nltk package), remove special characters (such as ‘/n’) and add other necessary processings.

3. TF and TF-IDF representation on ‘description’: for each sample in the dataset, generate TF and TF-IDF representation for each sample based on the column of ‘description’.

2.2. Practical Task: Topic modelling (10 Marks):

Use TF and TF-IDF representation generated in task 2.1 to perform topic modelling. Select and compare two topic modelling algorithms from LDiA, Truncated SVD, Word2Vec or any other topic modelling algorithms, and then analyze the results.

2.3. Analysis Task: Searching for similar movies (30 Marks):

Assume you would like to find similar movies as 'Harry Potter and the Half-Blood Prince' based on the given dataset, what would you do? Assume you already know that the user DOES NOT like 'Harry Potter and the Half-Blood Prince', then which movies would you suggest the user to watch? Then write a report using "**An example of possible report structure**" shown above. Please introduce your solution, where minimum information should be provided as follows:

1. Details on each step and expected inputs/outputs of each step
2. Major algorithm to be used to solve this problem
3. The results
4. Analysis on the results

Be noted that visualization should be used when exploring the data or illustrating the results.