Written examination paper for PGR210 – Machine Learning and Natural Language Processing
SEIT, Kristiania University College
Autumn 2024

Examination paper released: 15.11.2024

Examination deadline: see Wiseflow for details

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

Technical contact during examination: eksamen@kristiania.no

Exam type: Written home examination in groups
Exam submission files: Both the report and the code (from the ML and NLP sections) must be submitted. The code should be fully executable, with each result presented in the report having corresponding code in the implementation. The final grade will primarily be based on the quality of the report, while the code will serve as supplementary material for assessment

Support materials: All support materials are allowed

Final report format: use LaTeX or Word and font 12 with 1.5 spacing. The limit is 30 pages max that includes report, list of bibliography in the end, figures and tables. Both parts (ML and NLP) should be in the same PDF report

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
- knows the text analytics methods and tools (such as text classification, semantic textual similarity, neural language models etc) 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, 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

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

Part I Machine Learning

Problem 1: Classification Task (10 Marks):

The "Customer Health Profile" dataset has 3,500 rows, 5 features (Age, BMI, Smoking_Habit, Exercise_Frequency, Blood_Pressure) and a target variable (Health_Risk). The target variable Health_Risk has three categories: Low, Medium, and High. The dataset has some missing values.

Perform the following tasks:

a.) Data Preprocessing (3 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; Split the data into training and test sets. (1 marks)

b.) Model Building (4 marks)
- Select a suitable classification algorithm for the given dataset. Justify your choice. (2 marks)
- Train the model using the training set; Evaluate the model's performance on the test set and report the accuracy. (2 marks)

c.) Analysis (3marks)
- Visualize the distribution of the Health_Risk categories. (1 marks)
- Discuss any patterns or insights you can derive from the dataset. (1 marks)
- Provide any recommendations or suggestions for improving the classification results. (1 mark)

Problem 2: Clustering Task (10 Marks):

The "Online Shopping Behavior" dataset contains 6,000 rows and 4 features (Session_Duration, Page_Views, Purchase_Amount, Bounce_Rate) for each user's session on an e-commerce website. The goal is to cluster user sessions to identify different types of shopping behavior. The dataset has some missing values

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)
- Determine an appropriate number of clusters for the data using a suitable method. Justify your choice. (2 mark)

b.) Model Building (3 marks)

- Select a suitable clustering algorithm to identify number of clusters for the given dataset. Justify your choice. (3 marks)

c.) Analysis (3 marks)

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

Problem 4: **Deep Learning Based Regression Task (30 Marks):**
The "Rental Price Prediction" dataset includes 10,000 rows and 6 features (Property_Area, Bedrooms, Bathrooms, Property_Age, Proximity_to_City_Center, Monthly_Rent). The target variable is Monthly_Rent. This dataset has some missing values.

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. (3 marks)
- Normalize the features of the dataset. (1 marks)
- Split the data into training, validation, and test sets. (1 marks)

b.) Model Building (20 marks)

- Design a deep learning model suitable for regression tasks. Describe the architecture, including the number of layers, types of layers, and activation functions. (10 marks)
- Train the model using the training set and validate it using the validation set. (5marks)
- Evaluate the model's performance on the test set and report the mean squared error. Plot this. (5 marks)

c.) Analysis (5 marks)

- Plot the distribution of actual vs. predicted Monthly_Rent values (1 marks)
- Discuss any patterns or insights you can derive from the model's predictions. (2 marks)
- Provide any recommendations for improving the model's performance. (2 marks)

Part II Natural Language Processing

2.1. Practical Task: Text processing, feature extraction and representation by using both TF **and TF-IDF schemes, topic modelling (20 Marks)**

Given a data file (PGR210_NLP_data1.csv), where spam emails and non-spam emails are included. Please perform **data preparation**, **text pre-processing**, generate **TF and TF-IDF representation** for each sample, select and compare **two topic modelling algorithms** from LDiA, Truncated SVD, Word2Vec or any other topic modelling algorithms, and finally **analyze the results**.

In the report, provide a detailed description of the pipeline, outlining each step involved. Present your findings on topic modeling, and then include a comparison of the results for clarity and analysis.

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

Given a data file (PGR210_NLP_data2.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'.

Assume you would like to find similar movies as 'Spider-Man' based on the given dataset, what would you do? Please introduce your solution step-by-step, where such information should be provided in your report:

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

Assignment criteria*

Grade	Learning Outcome 1: Knowledge	Learning Outcome 2: Skills	Learning Outcome 3: Competence
A Excellent	Excellent and comprehensive understanding of concepts	Demonstrates excellent analytical, technical and writing skills	Outstanding degree of judgment and independent critical thinking
B Very good	Very good understanding of concepts	Demonstrates very good analytical, technical and writing skills	Sound degree of judgment and independent critical thinking
C Good	Good understanding of theory in most important areas	Demonstrates good analytical, technical and writing skills	Reasonable degree of judgment and independent critical thinking
D Satisfactory	Satisfactory understanding of theory, but with significant shortcomings	Demonstrates limited analytical, technical and writing skills	Limited degree of judgment and independent critical thinking

E Sufficient	Meets the minimum understanding of concepts	Demonstrates sufficient analytical, technical and writing skills	Very limited degree of judgment and independent critical thinking
F Fail	Fail to meet the minimum academic criteria.	No demonstration of analytical, technical and writing skills	Absence of judgment and independent critical thinking

*Adapted from The Norwegian Association of Higher Education Institutions