Bachelor of Informational Technology – Intelligent Systems

Module
TEK305 Machine Learning

Due date for submission
(see Wiseflow)

Module leader and e-mail
Noha El-Ganainy |Noha.El-Ganainy@kristiania.no

Teacher and e-mail
Arvind Keprate| arvindke@oslomet.no

Learning outcomes

After successfully completing the course the student:

Knowledge

- is able to explain the concept of machine learning and how this relates to the field of artificial intelligence.
- is able to explain the three main categories of machine learning: supervised learning, unsupervised learning and reinforcement learning.
- is able to be familiar with the concepts of overfitting and underfitting in connection with machine learning.
- is able to understand how machine learning can be used for tasks within classification, regression and clustering.
- is able to explain how common machine learning algorithms, such as support vector machines (SVMs), work.
- is able to explain how artificial neural networks work.
- is able to explain what is meant by deep learning.

Skills

- is able to use Python to solve machine learning tasks.
- is able to apply linear regression.
- is able to apply the most relevant machine learning algorithms.
- is able to map data using machine learning.
- is able to evaluate the performance of different machine learning algorithms.

General competence

The student ...

- is able to use machine learning as a tool to effectively identify and utilize information.
- is able to critically evaluate existing research related to machine learning.

Assignment specification

1. Group Size=Only 1 (Individual Submission)
2. A Jupyter notebook saved as ipynb (share directly on wise flow). Please use comments and/or markdown cell wherever necessary explanation is required. Additional marks will be given for clean Jupyter notebook and understandable code.
3. Referencing: Any acceptable academic style.

Please address the following questions in your submission.

Problem 1: Conceptual Questions (15 points)

Use Markdown cell in your Jupyter Notebook and explain the following:

1. What are some causes of overfitting? How do we diagnose and treat overfitting in regression models? **(5 points)**
2. What is difference between L1 and L2 regularization? **(5 points)**
3. What is a difference between a parameter and a hyperparameter? How is hyperparameter tuning performed for machine learning models **(5 points)**

Problem 2: Regression Problem (15 points)

The data in the file regression_*housedata.csv* are collected from 1,000 homes being sold in Oslo. The response variable of interest is the Price (price of the house). The input variables are bedrooms, sqft_living (the living space area), sqft_lot (the area of the land the house sits on), floors (the number of levels of the house), sqft_above (area of the house excluding the basement), sqft_basement (basement area). Following 2 tasks need to be performed:

1. Use kNN Regressor and Decision Tree Regressor to build a regression model for prediction of house prices? **(10 points)**
2. Perform Model Evaluation using two metrics: Root Mean Squared Error and Coefficient of Determination? Which of the two regression models is better MLR or Decision Tree? (**5 points)**

Problem 3: Clustering Problem (20 points)

The data in the file *clustering_diabetesdata.csv is* collected from 768 patients tested for diabetes. The dataset consists of following features:

1. Number of times pregnant
2. Plasma glucose concentration)
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)

Following tasks need to be performed:

1. Use k-Means clustering to identify any clusters? **(10 points)**
2. Use Hierarchical clustering to identify any clusters? (**10 points)**

Problem 4: Multi-Layer Perceptron Problem Using Keras (25 points)

The data scientists at one of the retail stores have collected 2019 sales data for different products across various stores in different cities. The data in the file *deep_learning_task_dataset.csv* consists of 5000 datapoints and consists of both input and output variables, the description of which is given in the table below. Divide the dataset into training (80%) and test (20%). You need to predict the sales for test data set.

Table 1: Input and Output Variables for Training Dataset

Variable	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.

Following tasks need to be performed:

1. Pre-processing of dataset. (**10 points**)
2. Define the architecture of your Deep Learning Model. Use markdown cell to explain the architecture of your model. **(5 points)**
3. Training your model. **(5 points)**
4. Accuracy of your predictions. Closer the value of $\mathbf{R^2}$ to 1, higher points you will score. **(5 points)**

Problem 5: CNN Problem Using Keras (25 points)

Build a CNN classifier for the Fashion MNIST dataset. You can use the instructions/commands in Jupyter notebook for downloading the Fashion MNIST dataset from the following link: https://keras.io/api/datasets/fashion_mnist/

Following tasks need to be performed:

1. Pre-processing of dataset. (**10 points**)
2. Define the architecture of your Deep Learning Model. Use markdown cell to explain the architecture of your model. **(5 points)**
3. Training your model. **(5 points)**
4. Accuracy of your predictions. Closer the value of **accuracy** to 1, higher points you will score. **(5 points)**

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