Bachelor of Informational Technology – Intelligent Systems

Module

TEK305 Machine Learning

Due date for submission

(see Wiseflow)

Module leader and e-mail Per Lauvås

Tomas Sandness | per.lauvas@kristiania.no

Teacher and e-mail

Arvind Keprate | arvindke@oslomet.no

Learning outcomes

After successfully completing the course the student:

Knowledge

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

Skills

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

General competence

The student ...

- use machine learning as a tool to effectively identify and utilize information.
- 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. For Problem 3, the csv file of predictions must be submitted.
4. Referencing: Any acceptable academic style.

Please address the following questions in your submission.

Problem 1: Regression Problem (20 points)

The data in the file *Regression_housedata.csv is* 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 Multilinear Regression (MLR), and kNN 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 kNN Regressor? (**10 points)**

Problem 2: 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 a 2 hours in an oral glucose tolerance test
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:

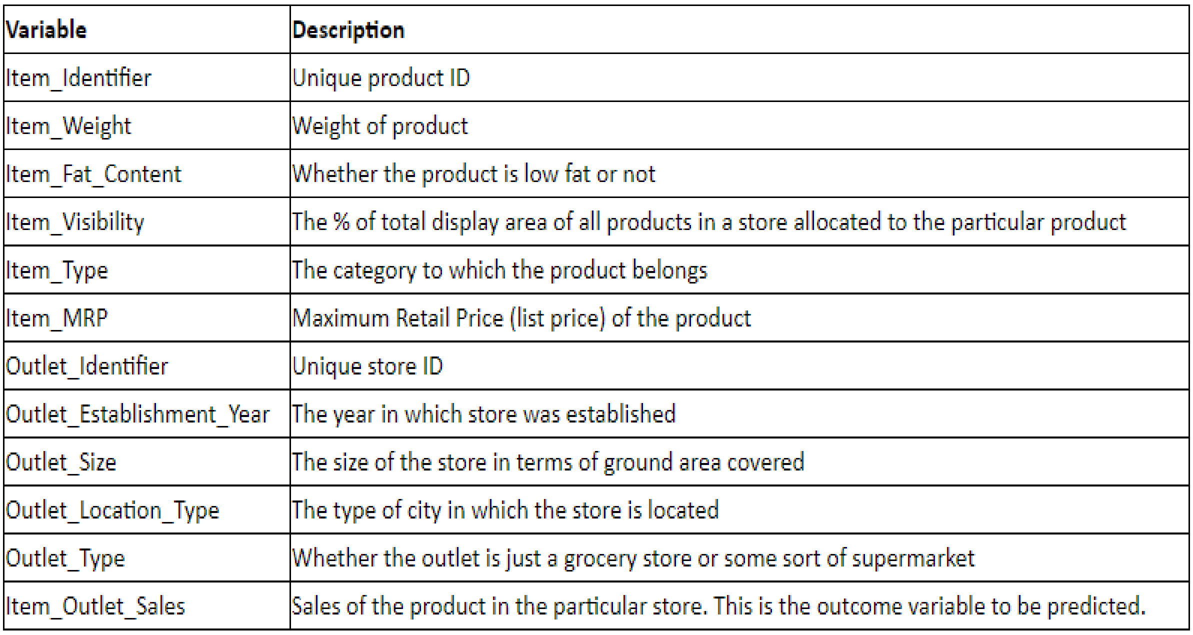
1. Use k-Means clustering to identify any clusters? **(10 points)**

2. Use Hierarchical clustering to identify any clusters? (**10 points)**

Problem 3: Deep Learning Problem Using Keras (60 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 consists of training (5000) and testing (3523) datasets (see the *Deep_learning_task_train.csv* and *Deep_Learning_task_test.csv*). The training dataset has both input and output variables, the description of which is given in the table below. 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.

Your model performance will be evaluated on the basis of your prediction of the sales for the testing data (*Deep_Learning_task_test.csv*), which contains similar variables as training dataset except for the variable (Item_Outlet_Sales) to be predicted. Your submission needs to be in the format as shown in *sample_submission.csv*.

The exam evaluator has the actual sales for the test dataset, against which your predictions will be evaluated. The evaluator will use the $\mathbf{R^2}$ value to judge your response.

Following tasks need to be performed:

1. Pre-processing of dataset (both *Deep_Learning_task_train.csv* and *Deep_Learning_task_test.csv*). (**10 points**)

2. Define the architecture of your Deep Learning Model. Use markdown cell to explain the architecture of your model. **(15 points)**

3. Training your model. **(10 points)**

4. Generate predictions for the test dataset using the trained model and save predictions in a csv file (to check the format, refer to the *sample_submission.csv* file provided). **(10 points)**

5. Accuracy of your predictions. Closer the value of $\mathbf{R^2}$ to 1, higher points you will score. **(15 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