|  |  |
| --- | --- |
| **Project Case** |  |
| COMP7117  Artificial Neural Network |
| **Computer Science** | **E202-COMP7117-TD01-00** |
| ***Valid on*** *Even Semester Year 2019/2020* | **Revision 00** |

1. Seluruh kelompok tidak diperkenankan untuk:

*The whole group is not allowed to:*

* + - Melihat sebagian atau seluruh proyek kelompok lain,

*Seeing a part or the whole project from other groups*

* + - Menyadur sebagian maupun seluruh proyek dari buku,

*Adapted a part or the whole project from the book*

* + - Mendownload sebagian maupun seluruh proyek dari internet,

*Downloading a part or the whole project from the internet,*

* + - Mengerjakan soal yang tidak sesuai dengan tema yang ada di soal proyek,

*Working with another theme which is not in accordance with the existing theme in the matter of the project,*

* + - Melakukan tindakan kecurangan lainnya,

*Committing other dishonest actions,*

* + - Secara sengaja maupun tidak sengaja melakukan segala tindakan kelalaian yang menyebabkan hasil karyanya berhasil dicontek oleh orang lain / kelompok lain.

*Accidentally or intentionally conduct any failure action that cause the results of the project was copied by someone else / other groups.*

1. Jika kelompok terbukti melakukan tindakan seperti yang dijelaskan butir 1 di atas, maka **nilai kelompok** yang melakukan kecurangan (menyontek maupun dicontek) akan di – **NOL** – kan.

*If the group is proved to the actions described in point 1 above, the score of the group which committed dishonest acts (cheating or being cheated) will be “Zero”*

1. Perhatikan jadwal pengumpulan proyek, segala jenis pengumpulan proyek di luar jadwal tidak dilayani.

*Pay attention to the submission schedule for the project, all kinds of submission outside the project schedule will not be accepted*

1. Jangan lupa untuk melihat kriteria penilaian proyek yang ditempel di papan pengumuman, atau tanya asisten anda.

*Don’t forget to look at the project assessment criteria that posted on the announcement board, or ask your teaching assistant.*

1. Persentase penilaiaan untuk matakuliah ini adalah sebagai berikut:

*Marking percentage for this subject is described as follows:*

|  |  |  |
| --- | --- | --- |
| **Tugas Mandiri**  *Assignment* | **Proyek**  *Project* | **UAP**  *Final Exam* |
| 40% | 60% | - |

1. Software yang digunakan pada matakuliah ini adalah sebagai berikut:

*Software will be used in this subject are described as follows:*

|  |
| --- |
| **Software**  *Software* |
| Microsoft Visual Studio Code  Python 3.6  Scipy  Scikit  TensorFlow 1.10 |

## Ekstensi file yang harus disertakan dalam pengumpulan tugas mandiri dan proyek untuk matakuliah ini adalah sebagai berikut:

*File extensions should be included in assignment and project collection for this subject are described as follows:*

|  |  |
| --- | --- |
| **Tugas Mandiri**  *Assignment* | **Proyek**  *Project* |
| PY | PY |

## Soal

*Case*

**Red Wine Online Shop**

**Red Wine Online Shop** is e-commerce based in US which focuses on the quality issues of red wine. To make their job more efficient, **Red Wine Online Shop** wants to build a model that could predict the overall quality of the red wine from their ingredients. **Red Wine Online Shop** also wants to cluster customer’s intention from customer behavior in the website. Therefore, as a programmer, you are asked to help them build the application based on the existing dataset.

* + - 1. **Clustering (Self-Organizing Map)**

First, **Red Wine Online Shop** wants to **group the user** based on the **similarities** in user **behaviors**. To do that, you are going to use **Kohonen Self-Organizing Map** technique to **cluster the data**.

1. **Dataset Description**

**Content**

The given dataset contains **3,632 data** consisting of **behavior information** of all **Red Wine Online Shop** users in the web page.

**Feature Description**

The table below shows the feature descriptions in the dataset.

Table . Table of features descriptions for clustering

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Column** | **Description** | **Possible Value** |
| **Features** | Administrative | Administrative Value | Number |
| Administrative Duration | Duration in Administrative Page | Number |
| Informational | Informational Value | Number |
| Informational Duration | Duration in Informational Page | Number |
| Product Related | Product Related Value | Number |
| Product Related Duration | Duration in Product Related Page | 0 to 3 |
| Bounce Rates | Bounce Rates of a Web Page | Number |
| Exit Rates | Exit Rates of a Web Page | 0 to 3 |
| Page Values | Page Values of Each Web Page | Number |
| Special Day | Special Days Rate like Valentine etc. | String |
| Month | Month of the Year | String |
| Operating System | Operating System Used | 0 to 8 |
| Browser | Browser Used | 0 to 13 |
| Region | Region of the User | 1 to 9 |
| Traffic Type | Traffic Type in the Web Page | 1 to 20 |
| Visitor Type | Types of the Visitor | String |
| Weekend | Weekend or Not | True or False |
| Revenue | Revenue Will be Generated or Not | True or False |

1. **Feature Selection**

Instead of **using the actual value** for the clustering, you are asked to create **features derived** from the **actual data**. The features requested are:

Table . Required features and derivation formula

|  |  |
| --- | --- |
| **Feature** | **Derivation Formula** |
| Special Day Rate | if (Special Day is “HIGH”):  Special Day = 2  elif (Special Day is “NORMAL”):  Special Day = 1  elif (Special Day is “LOW”):  Special Day = 0 |
| Visitor Type | if (Visitor Type is “Returning\_Visitor”):  Visitor Type = 2  elif (Visitor Type is “New\_Visitor”):  Visitor Type = 1  elif (Visitor Type is “Other”):  Visitor Type = 0 |
| Weekend | if (Weekend is “TRUE”):  Weekend = 1  elif (Weekend is “FALSE”):  Weekend = 0 |
| Product Related Duration | Duration in Product Related Page |
| Exit Rates | Exit Rate of a Web Page |

1. **Feature Extraction**

After the five new features are extracted, you are asked to use **Principal Component Analysis** (**PCA**) to both clean the data and reduce the dimensionality even further.

The steps that you want to take are as follows:

1. **Select the features** as defined in the Feature Selection section
2. **Normalize** the data
3. Analyze the data with **Principal Component Analysis** to obtain the new components
4. Take the **highest 3 principal components** as the input of your neural network
5. **Architecture**

You are to **create your own architecture design** that will be **able to solve the given problem**. Consider the following when building your architecture:

* **Number of input nodes** required
* **Number of clusters**

These considerations will be **accounted for in the grading process**.

1. **Training**

The training procedure of the neural network are as follows:

* + - 1. **Epoch** forthetrainingsis **5000**
      2. **For each data** in the dataset, **find the winning node** by using **nearest distance**
      3. **Update the neighbor around** the winning node in a square pattern
      4. **Update the weight** of the network

1. **Visualization**

**After the training is complete**, use **matplotlib** to **visualize the clusters generated by the self-organizing map**.

1. **Classification**

In addition to **cluster similar user behaviors**, **Red Wine Online Shop** also wants you to create an application that can help them **classify** the **grade** of the wine they sell. In order to do that, you will be given a dataset labeled with **overall quality of the wine** which depends on various ingredients of the wine.

* 1. **Dataset Description**

**Content**

The given dataset contains **1,599** **data of the wine** which are labeled with **overall quality of the wine** organized by **UCI Machine Learning**. The overall quality is based on the ingredients of the wine, with “**Great**” being the highest-quality grade, “**Good**” is second best, “**Fine**” is the middle grade, “**Decent**” is one grade bellow, and “**Fair**” being the lowest-quality grade.

**Feature Description**

The table below shows the feature descriptions in the dataset.

Table . Table of feature descriptions for classification

|  |  |  |  |
| --- | --- | --- | --- |
| **Category** | **Column** | **Description** | **Possible Value** |
| **Features** | Fixed Acidity | Most acids involved with wine or fixed or nonvolatile (do not evaporate readily) | Number |
| Volatile Acidity | The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste | Number |
| Citric Acid | Found in small quantities, citric acid can add 'freshness' and flavor to wines | Number |
| Residual Sugar | The amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and | Number |
| Chlorides | The amount of salt in the wine | String |
| Free Sulfur Dioxide | The free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents | String |
| Total Sulfur Dioxide | Amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 | Number |
| Density | The density of water is close to that of water depending on the percent alcohol and sugar content | String |
| pH | Describes how acidic or basic a wine. | String |
| Sulphates | A wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and | Number |
| Alcohol | The percent alcohol content of the wine | Number |
| **Output** | Wine Overall Quality | The overall quality of the wine. | “Great”, “Good”, “Fine”, “Decent”, and “Fair” |

* 1. **Feature Selection**

To **classify** the **overall rating** of the hospital, **Red Wine Online Shop** does not want to use all features. They will select **only a few features** for you to build the application. Instead of using the **actual value**, you are tasked to create **features derived** from the actual data. The features requested are:

Table . Required features and derivation formula

|  |  |
| --- | --- |
| **Feature** | **Derivation Formula** |
| Volatile Acidity | Volatile Acidity Rate |
| Chlorides | Chlorides Rate |
| Free Sulfur Dioxide | if (Free Sulfur Dioxide is “High”):  Free Sulfur Dioxide = 3  elif (Free Sulfur Dioxide is “Medium”):  Free Sulfur Dioxide = 2  elif (Free Sulfur Dioxide is “Low”):  Free Sulfur Dioxide = 1  else:  Free Sulfur Dioxide = 0 |
| Total Sulfur Dioxide | Total Sulfure Dioxide Rate |
| Density | if (Density is “Very High”):  Density = 0  elif (Density is “High”):  Density = 3  elif (Density is “Medium”):  Density = 2  elif (Density is “Low”):  Density = 1 |
| pH | if (pH is “Very Basic”):  pH = 3  elif (pH is “Normal”):  pH = 2  elif (pH is “Very Acidic”):  pH = 1  else pH = 0 |
| Sulphates | Sulphates Rate |
| Alcohol | Alcohol Rate |

While the **output of the system** will be:

Table 5. Required system output

|  |
| --- |
| **Label Name** |
| Quality (Wine overall quality) |

* 1. **Feature Extraction**

Due to the **large number of features** that need to be considered in building the neural network, you want to **simplify the data** to make your network trains faster. While **reducing the complexity of the data is important**, **preserving the variance and relationship between the data is also important**. To solve those problems, your approach in **reducing the dimensionality** of the data is by using **Principal Component Analysis (PCA)** technique.

The steps that you want to take are as follows:

1. **Select the features** as defined in the Feature Selection section
2. **Normalize the data**
3. **Analyze** the data with **Principal Component Analysis** to obtain the new components
4. Take the **highest 4 principal components** as the **input of your neural network**
   1. **Architecture**

You are to **create your own architecture design** that will be **able to solve the given problem**. Consider the following when building your architecture:

* **Number of input nodes** required
* **Number of output nodes** (classes) required
* **Whether hidden layer is required** **or not** (whether the case is a linearly separable case or not)

These considerations will be **accounted for in the grading process**.

* 1. **Training**

The training of the neural network is done with **70% of the dataset picked randomly**. The training is done with **gradient descent** as the optimization formula for **5,000 epochs**. In addition, during the training, **20% of the dataset** should be used as the **validation dataset**.

The training procedure are as follows:

1. **Initialization**

The initialization step needs to be run once before starting the training iteration:

1. Take the **output** of the **Principal Component Analysis** as the **features**
2. **Initialize** the **weights** and **biases** **randomly**
3. **Iteration**

For **5,000 epochs**, repeat the following:

1. **Calculate the error** by comparing the output of the neural network to the target in the dataset using **mean squared error** (**MSE**)
2. **Update** the **weights and biases** using **gradient descent optimization**
3. **For every 100 epochs**, **print** the **current error** and **epoch number** to the console
4. **After reaching the 500th epoch**, **calculate the validation error** by passing the validation dataset. After that, **record the validation error** and **save the model to file**
5. **For** **every 500 epochs**, **get the new validation error** by passing in the validation dataset. If the **validation error is lower** than the previous validation error, **save the model to file**. If the **validation error is higher**, **do not save the model**
   1. **Evaluation**

The neural network is to be **evaluated** based on the accuracy with **10% of the dataset** **after the training process** finished. The **accuracy** is calculated as follows:

|  |
| --- |
|  |

**Reference**

* The dataset is obtained from Kaggle (<https://www.kaggle.com/roshansharma/online-shoppers-intention>) by Roshan Sharma. The dataset has been heavily cleaned and modified for the purpose of this case.
* The dataset is obtained from Kaggle (<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>) by UCI Machine Learning. The dataset has been heavily cleaned and modified for the purpose of this case.