Assignment 3: Summative Practical Assessment in Machine Learning Module of the Data Analytics Course

Choose questions to obtain 60 marks for this assignment 3. No question is compulsory. Please read the individual question for instructions. Individual questions will carry different marks mentioned next to the question number.

(40 marks will be awarded for the previously answered assignment 1 & 2 that have been already uploaded to UGVLE. By answering additional questions, it is possible to accumulate extra marks, capped from 100. The total marks for the machine learning assignments will be from 15/100 of the final mark).

Final assignment mark for the machine learning assignments (1,2,3) is given below:
= 0.15 * min(100, [assignment 1 mark + assignment 2 mark + assignment 3 marks])

Deadline is indicated in the UGVLE.
Question 1: 30 marks
Question 2, 3, 4, 5, 6, 7, 8 : 10 marks each

Question 1 - [30 marks]

Task 1: Read and understand the Online Passive Aggressive journal paper (plus the NIPS conference paper if you require further explanation), in particular, **sections 1,2,3 and 10 (compulsory reading sections) of the paper**. You may read other sections if you wish to do so (optional).

Task 2: Implement the Online Passive Aggressive Algorithm in Python without referring to the in-built implementations of the PA classifier inside machine learning libraries like Scikit Learn.

Task 3: Download the Breast Cancer WIS Consin dataset
(URL: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic))
and its attribute description from the UCI Data repository:

Data:

http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data

Description:

http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names

<u>Handling Missing Data</u>: Save the data file as datafile.csv. Open it with a simple text editor (e.g. Notepad++) and replace the question mark with 0 (zero as an integer value).

Attribute Information: (class attribute has been moved to the last column)

#	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 - 10
3.	Uniformity of Cell Size	1 - 10
4.	Uniformity of Cell Shape	1 - 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class:	(2 for benign, 4 for malignant)

Load the data as variable “data”.

From the above attribute description, remove the 1st attribute, which is a unique identifier. Replace the 11th attribute, the class number (2 or 4), with (-1,+1). Benign means not infected with cancer. Malignant means possible cancer.

Using the above description, come up with the following X input matrix:

X = data[2:10]

Y = data[11]

Task 4: Separate the data into training (2/3) and testing (1/3). Report training and testing accuracy for C=1, iterations={1,2,10}.

Output:

Record a video describing the source code and the algorithm. Input your Youtube video's link to the text box available at the Assignment submission link in the UGVLE.

Upload the source code, data, predictions, results (train, test) and your conclusion as a zip file with a Word/pdf document in it. Use the UGVLE link provided. Mention the link to your Youtube video in the submitted report (word/PDF document) as well.

Question 2 - [10 marks]

1. Apply ID3 and SVM algorithms to the Breast Cancer (numerical) and CIFAR - 10 (image) datasets, which can either be loaded using sklearn and keras or downloaded using the links given below.
   Breast Cancer (numerical) : [Please refer Question 1 - Task 03]
   CIFAR - 10 (Canadian Institute For Advanced Research) (image):
   Dataset consists of 60000 images, each of 32x32x3 colour images having ten classes, with 6000 images per category. The dataset consists of 50000 training images and 10000 test images.
   The classes in the dataset are airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck.

   You can download the CIFAR dataset from:
   https://www.cs.toronto.edu/~kriz/cifar.html
   or

   by loading it with the help of the Keras library**.**

   > from keras.datasets import cifar10

2. Differentiate the results obtained in part 1 above.

3. Evaluate the results obtained by applying ID3 and SVM algorithms to the Breast Cancer dataset using the performance metrics for classification given below:
   a. Training and Testing Accuracy (Generalization Error)
   b. Sensitivity & Specificity
   c. Precision & Recall
   d. Confusion Matrix
   e. F1 Score
   f. ROC Curve

Hint: You can refer to these links for more details about performance metrics for classification -
https://towardsdatascience.com/a-practical-guide-to-seven-essential-performance-metrics-for-classification-using-scikit-learn-2de0e0a8a040

https://towardsdatascience.com/popular-machine-learning-performance-metrics-a2c33408f29

https://medium.com/analytics-vidhya/classification-performance-metric-with-python-sklearn-d8342ac25898

https://builtin.com/data-science/evaluating-classification-models

https://www.kaggle.com/vipulgandhi/how-to-choose-right-metric-for-evaluating-ml-model

Output:

Upload the source code, predictions and your answers as a zip file. Use the UGVLE link provided. Record a video describing the source code and the algorithm. Input your Youtube video's link to the text box available at the Assignment submission link in the UGVLE.

Question 3 - [10 marks]

The following question relates to the Principal Component Analysis.

1. Execute this Colab Notebook (URL:https://colab.research.google.com/drive/10BL-SMMYaeldrq7VT7FA9hsR9Z-_yiXJ?usp=sharing). Moreover, document the code by understanding the implementation of the PCA algorithm.

   Hint: You can refer to this online tutorial : https://pub.towardsai.net/principal-component-analysis-pca-with-python-examples-tutorial-67a917bae9aa

2. Write answers to the following questions
   a. What is PCA?
   b. How can we use PCA for dimensionality reduction and visualize the classes?
   c. Why do we need to normalize data before feeding it to any machine learning algorithm?

3. Apply PCA algorithm to **Breast Cancer dataset [Please refer Question 1 - Task 03]**

   a. Reduce the dimensionality into 2 features.
   b. Visualize the data to check whether it helps to distinguish between a patient who has breast cancer or not.

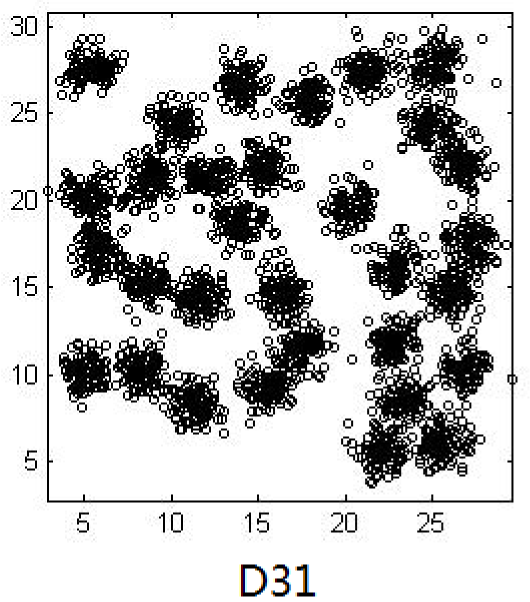
Output: Record a video describing the source code, the algorithm and your opinion on whether PCA helps to distinguish between a patient who has breast cancer or not. Input your Youtube video's link and colab notebook link to the text box available at the Assignment submission link in the UGVLE.

Upload the source code/notebook, data and your answers with a conclusion as a zip file. Use the UGVLE link provided. Mention the link to your Youtube video in the submitted notebook as well.

Question 4 - [10 marks]

The following question relates to the Self Organizing Maps (SOM):

1. Download the D31 clustering dataset, which has a 2D distribution as given in the following figure. Use this dataset for answering the following questions.
   URL to D31 dataset: http://cs.joensuu.fi/sipu/datasets/D31.txt



2. Implement SOM algorithm from scratch in a Colab notebook with documentation using markdown.
3. By using the 2D dataset, explain how the different regions get excited. Explain how the weights **w** in the SOM algorithm change with these input vectors.
4. Create a Youtube video describing the answers to the above questions, geometric data 2D representation.

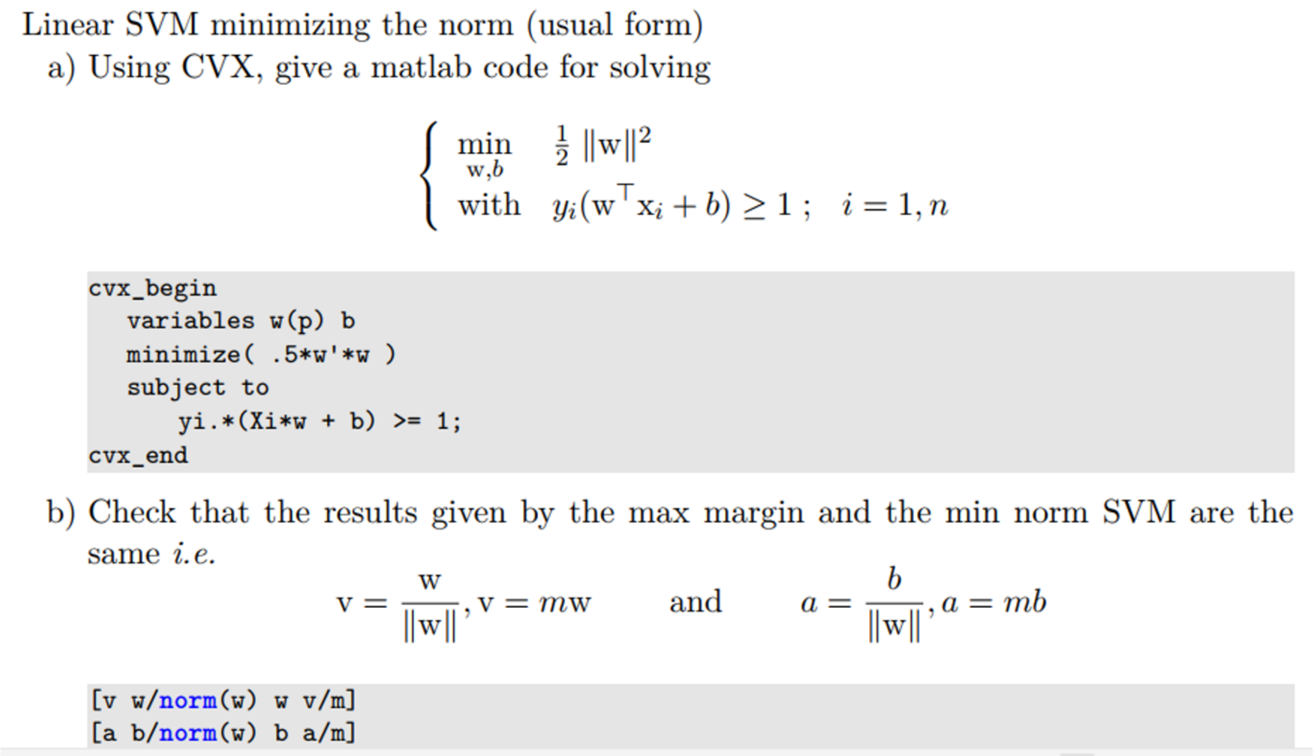
Output:

Record a video describing the source code, the algorithm and answers to the questions given. Input your Youtube video's link and colab notebook link to the text box available at the Assignment submission link in the UGVLE.

Upload the source code/notebook, data and your answers as a zip file. Use the UGVLE link provided. Mention the link to your Youtube video in the submitted notebook as well.

Question 5 - [10 marks]

1. Implement a support vector machine by using an optimizer like CVX (http://cvxr.com/cvx/), where the SVM formulation is available in the slides.
2. Use Breast Cancer dataset mentioned above [Please refer Question 1 - Task 03] to evaluate this implementation.



Linear SVM minimizing the norm (usual form)

a) Using CVX, give a matlab code for solving

$$\begin{cases} \min\limits_{\mathrm{w},b} & \frac{1}{2}\,\|\mathrm{w}\|^2 \\ \text{with} & y_i(\mathrm{w}^\top \mathrm{x}_i + b) \geq 1 \,; \quad i = 1, n \end{cases}$$

```
cvx_begin
  variables w(p) b
  minimize( .5*w'*w )
  subject to
     yi.*(Xi*w + b) >= 1;
cvx_end
```

b) Check that the results given by the max margin and the min norm SVM are the same *i.e.*

$$\mathrm{v} = \frac{\mathrm{w}}{\|\mathrm{w}\|}, \mathrm{v} = m\mathrm{w} \qquad \text{and} \qquad a = \frac{b}{\|\mathrm{w}\|}, a = mb$$

```
[v w/norm(w) w v/m]
[a b/norm(w) b a/m]
```

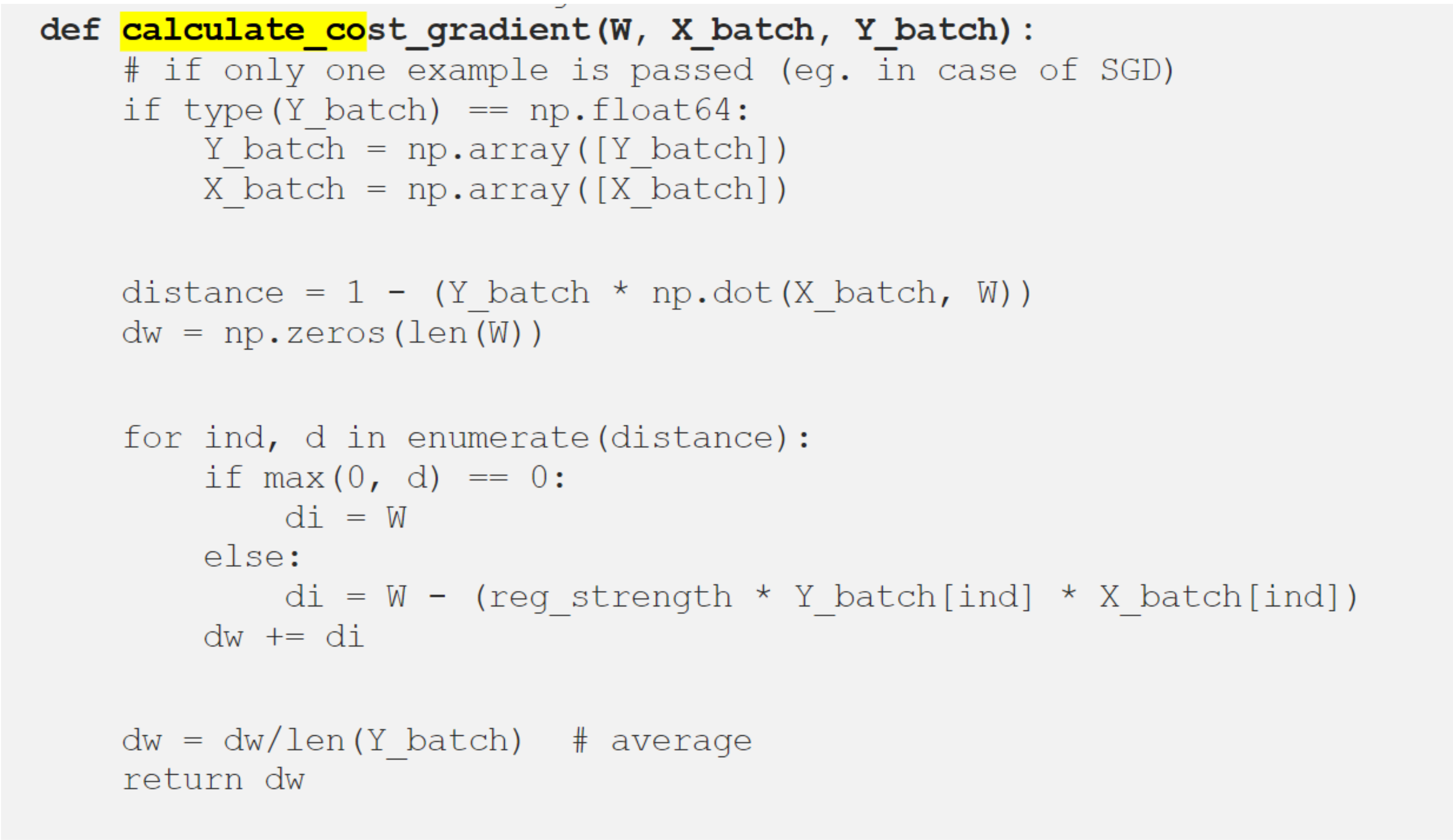
Hint: You can refer to the tutorial such as the following.
URL: https://towardsdatascience.com/svm-implementation-from-scratch-python-2db2fc52e5c2

Output:

Upload the source code, predictions and your answers as a zip file with a Word/pdf document in it. Use the UGVLE link provided. Record a video describing the source code and the algorithm. Input your Youtube video's link to the text box available at the Assignment submission link in the UGVLE.

Question 6 - [10 marks]

1. Implement the Support Vector Machine algorithm from scratch by considering the gradient of the formulation as mentioned in the slides.
2. Use the Breast Cancer dataset mentioned above [Question 1 - Task 03] to evaluate your implementation.



```
def calculate_cost_gradient(W, X_batch, Y_batch):
    # if only one example is passed (eg. in case of SGD)
    if type(Y_batch) == np.float64:
        Y_batch = np.array([Y_batch])
        X_batch = np.array([X_batch])

    distance = 1 - (Y_batch * np.dot(X_batch, W))
    dw = np.zeros(len(W))

    for ind, d in enumerate(distance):
        if max(0, d) == 0:
            di = W
        else:
            di = W - (reg_strength * Y_batch[ind] * X_batch[ind])
        dw += di

    dw = dw/len(Y_batch)  # average
    return dw
```

Hints: Refer tutorials such as this : https://towardsdatascience.com/svm-implementation-from-scratch-python-2db2fc52e5c2 *(Use the browser in incognito mode)*

Output:

Record a video describing the source code and the algorithm. Input your Youtube video's link to the text box available at the Assignment submission link in the UGVLE.

Upload the source code, data, predictions, results and your conclusion as a zip file with a Word/pdf document in it. Use the UGVLE link provided. Mention the link to your Youtube video in the submitted report (word/PDF document) as well.

Question 7: **[10 Marks]**

Convolutional Neural Networks

1. Build a Convolutional Neural Network based visual classifier for the MNIST dataset, consisting of digits from 0-9. Exemplar architectures including but not limited to AlexNet, VGG, ResNet, Inception, Xception, NASNet.
2. Describe the architecture and the operations: ReLU, Convolution Operation, Max-Pool and other operations depending on your choice.

Hints: You may refer to the tutorials such as these:
URL: https://victorzhou.com/blog/keras-cnn-tutorial/,
https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/

Output:

Record a video describing the source code, the algorithm and answers to the questions given. Input your Youtube video's link to the text box available at the Assignment submission link in the UGVLE.

Upload the source codes, data and answers as a zip file with a Word/pdf document. Use the UGVLE link provided. Mention the link to your Youtube video in the submitted report (word/PDF document) as well.

Question 8: [10 Marks]

1. Write a report on the Design Patterns used in Machine Learning. Present the report as a set of slides and record a video to be uploaded to Youtube.com.

Hints: https://towardsdatascience.com/design-patterns-in-machine-learning-b73eea4882cd
Developer’s Gang of Four Design patterns are described here:
https://www.tutorialspoint.com/design_pattern/design_pattern_overview.htm

Output:
Record a video describing your report as a presentation. Input your Youtube video’s link to the text box available at the Assignment submission link in the UGVLE.

Upload the report (PDF/Word), presentation as a zip file. Use the UGVLE link provided. Mention the link to your Youtube video in the submitted report (word/PDF document) as well.