

**MLDM Assignment**

**Machine learning and data mining course: 364-2-1651**

**Name: Inbal Karibian ID: 315861773**

**Name: Adi Yogev Golbari ID: 307849869**

**Name: Arad Peleg ID: 206068801**

**Part 1:** Theoretical question

* Showing that the maximum likelihood estimates for a univariate Normal distribution:

**תמונה שמכילה קו, גופן, טקסט, עלילה

התיאור נוצר באופן אוטומטי**

**תמונה שמכילה טקסט, כתב יד, גופן, מספר

התיאור נוצר באופן אוטומטי**

**תמונה שמכילה טקסט, כתב יד, גופן, מספר

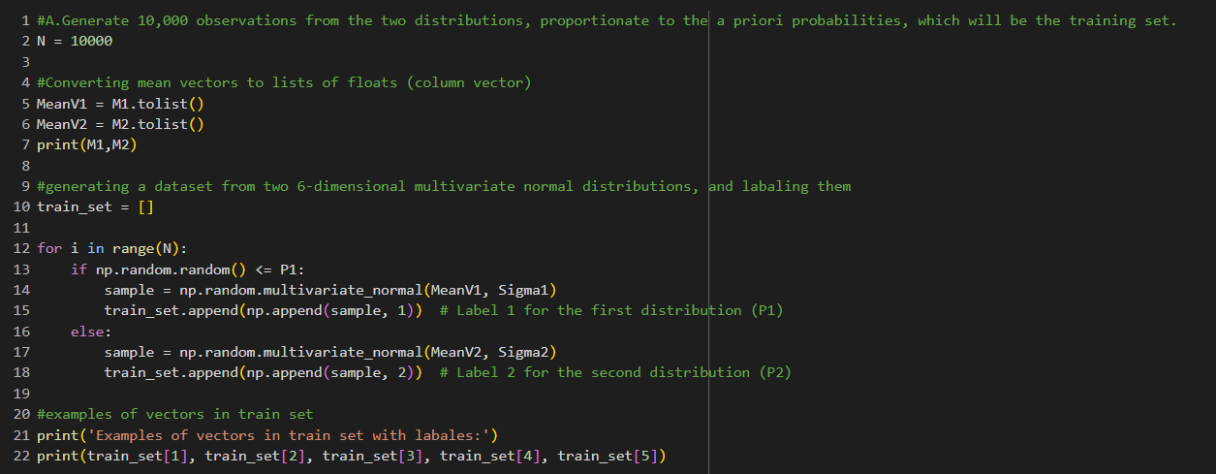
התיאור נוצר באופן אוטומטי**

**Part 2:** Question with data and software

**Note:**

* we uploaded the data to git-hub (arad’s) repository for easy access in google collab. because we are three ppl and its more comfortable to work like this as a team on the same not book.
* Environment: google collab

1. Generating 10,000 observations from the two distributions:



1. Computing the MLE estimators for each of the class conditional parameters and comparing them to the true values:

**Code:**

**תמונה שמכילה טקסט, צילום מסך

התיאור נוצר באופן אוטומטי**

**תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי**

**Outputs:**

**תמונה שמכילה טקסט, גופן, תפריט, צילום מסך

התיאור נוצר באופן אוטומטיתמונה שמכילה טקסט, גופן, תפריט, צילום מסך

התיאור נוצר באופן אוטומטי**

**תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי**תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

1. Generating validation set data with 2,000 observations:

**תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי**

* Building helping functions: cross-validation, optimization and estimation (evaluation)
* In cross validation we are using CV-10 over the training set to estimate the model accuracy and generalization error in RN and MLP(NN):

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט, צילום מסך

התיאור נוצר באופן אוטומטי

תמונה שמכילה טקסט, צילום מסך

התיאור נוצר באופן אוטומטי

1. Fitting a random forest to the data. Using the validation set to compare several forest configurations to choose the best performing one:

**Code:**

**תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי**

**Outputs:**

* RF best tree and accuracy:

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

* RF manual cross validation means and variance of accuracy and generalization error (CV-10):

תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי

1. Fitting a neural network to the data. Using the validation set to determine the number of neurons to use for the network:

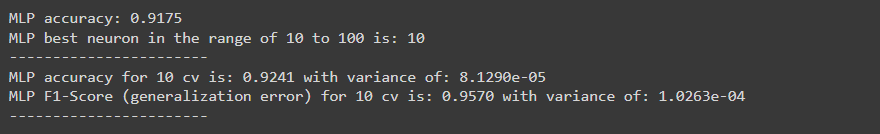
**Code:**

**תמונה שמכילה טקסט, צילום מסך, גופן

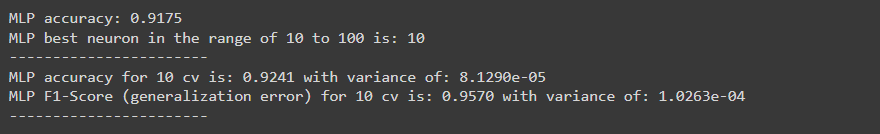
התיאור נוצר באופן אוטומטי**

**Outputs:**

* MLP (NN) best neuron and accuracy:



* MLP (NN) manual cross validation means and variance of accuracy and generalization error (CV-10):



1. overfitting phenomenon as a function of training set size. Fitting the model with training sets of size 𝑁 = 10, 20, 30, … ,1,000. Plotting the test and training error as a function of 𝑁.

**Note:**

We choose the model MLP because it has higher accuracy.

**Code:**

**תמונה שמכילה טקסט, צילום מסך, גופן

התיאור נוצר באופן אוטומטי**

**Output:**

תמונה שמכילה טקסט, צילום מסך, גופן, עלילה

התיאור נוצר באופן אוטומטי

**Explanation:**

The graph presents the F1 scores for both training and testing datasets across different training data sizes. with small training data sizes, the model overfits we know that because of the high training F1 score and a lower testing F1 score. As the training data size increases, the testing F1 score improves and becomes more stable, indicating better adaptability. Eventually, the training and testing F1 scores meet, suggesting that the model is no longer overfitting and is learning to adapt well to unseen data. This shows the importance of having sufficient amount of training data to develop a good model.