Pattern Recognition

ECE 4363 / ECE 5363

Project 1

- Load the dataset "fisheriris" into the workspace. For those programming in Python, the dataset is provided in the attached Excel file.
  - Study the dataset in terms of (a) Number of classes, (b) Number of features, and (c) What the data represents, i.e., gain some intuition about the problem domain. Based on your study, would you expect the features to perform well in this problem?

There is three classes in the data, three species of this Iris flower apparently. There are four features, sepal length, sepal width, pedal length and pedal width. Which appear to be measurements of the flower. The sepal is the outer green petals that protect the flower before it blooms, they form the bud I suppose you would say and therefore you might expect the bigger sepals to have bigger flowers although the data at a cursory glance seems a bit mixed on that assumption. The length seems to be related while the width does not.

There does seem to be some differences between the classes between measurements, for example below you can see the difference in petal width between Setosa and Versicolor is significant with Setosa's being around .2-.4 and Versicolor's being around 1.3-1.5 and it is like this for several of the measurements and classes and this would lead me to believe that we would be able to classify these well with the methods we have at our disposal.

0.2	setosa
0.3	setosa
0.3	setosa
0.2	setosa
0.6	setosa
0.4	setosa
0.3	setosa
0.2	setosa
0.2	setosa
0.2	setosa
0.2	setosa
1.4	versicolor
1.5	versicolor
1.5	versicolor
1.3	versicolor
1.5	versicolor
1.3	versicolor
1.6	versicolor
1	versicolor
1.3	versicolor
1.4	versicolor
1	versicolor
1.5	versicolor

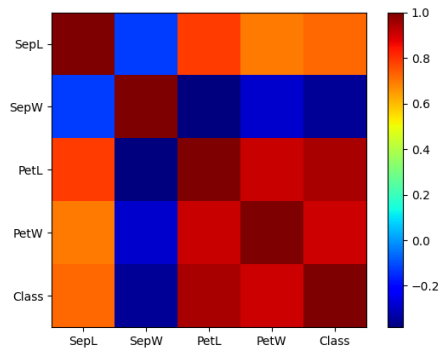
- Compute the following quantities for each feature. Do you observe anything of interest from these statistics?

	Sepal Length	Sepal Width	Pedal Length	Pedal Width
Minimum	4.3	2.0	1.0	0.1
Maximum	7.9	4.4	6.9	2.5
Mean	5.8	3.1	3.8	1.2
Variance	0.7	0.2	3.1	0.6
Within-Class Variance	Setosa: 0.1, Versicolor: 0.2, Virginica: 0.3			
Between-Class Variance	Setosa: 2.3, Versicolor: 0.1, Virginica: 1.5			

The Between-Class Variance of Versicolor being so low seems alarming. This seems to say that Versicolor is not very different than the other two and therefore difficult to classify. I will bring up IID now because this seems to be the first place where it comes into question. Some of the Within-Class Variances being bigger than that Between-Class Variance makes it seem like you will have a harder time telling different Versicolors apart than telling it from other classes.

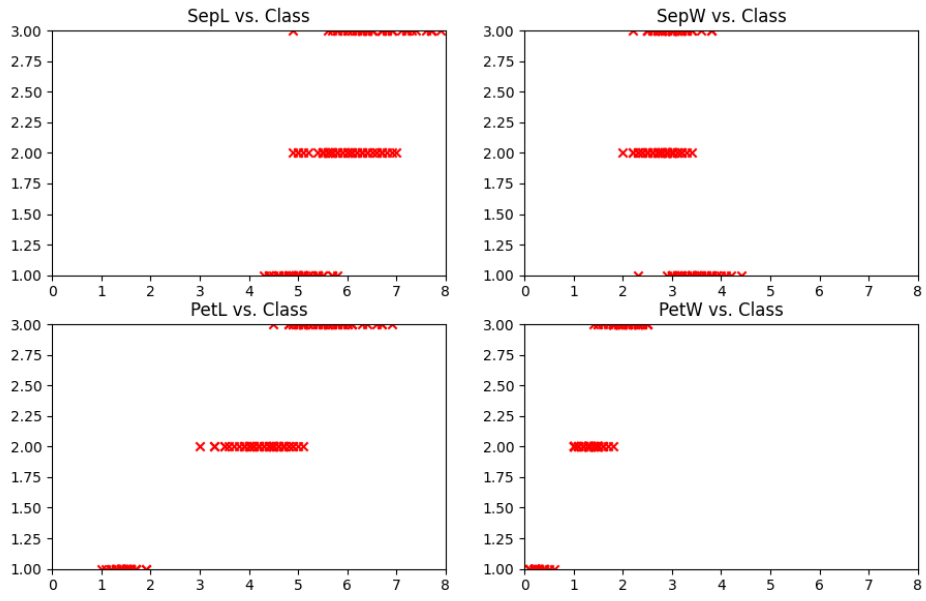
- Compute and display the correlation coefficients exactly as shown below (left figure). Do you observe anything interesting from this display?

I observe something very interesting from this display which is that petal length and petal width are highly correlated, that is a serious problem for IID which is what our classifiers rely on to be effective. The measurements are not independent. Even petal length and sepal length are highly correlated as I suggested earlier. Although petal width and sepal width are not which also supports my earlier observations.



- Display each of the four features versus the class label, exactly as shown below (right figure). What can you state about how well the features may perform in classification?

I mean yeah, you can already see how only a couple of them are linearly separable which is going to be a problem for Perceptron and expect some misclassification from Least Squares. Petal Length and Petal Width appear linearly separable between Setosa and Versicolor although not with Virginica and Versicolor.



- Perform the following classification tasks.

Setosa Vs. Versi+Virigi	All Features	Batch_Perceptron and LS
Setosa Vs. Versi+Virigi	Features 3 and 4 Only	Batch_Perceptron and LS
Virgi Vs. Versi+Setosa	All Features	Batch_Perceptron and LS
Virgi Vs. Versi+Setosa	Features 3 and 4 Only	Batch_Perceptron and LS
Setosa Vs. Versi Vs. Virigi	Features 3 and 4 Only	Multiclass LS

- For each case, report whether the method converged. If so, report (a) No. of epochs, (b) Computed weight vector, (c) No. of training misclassifications, and whenever appropriate, (d) plot of feature vectors, as well as the computed decision boundary.

Setosa Vs. Versi+Virigi	All Features	Batch Perceptron and LS

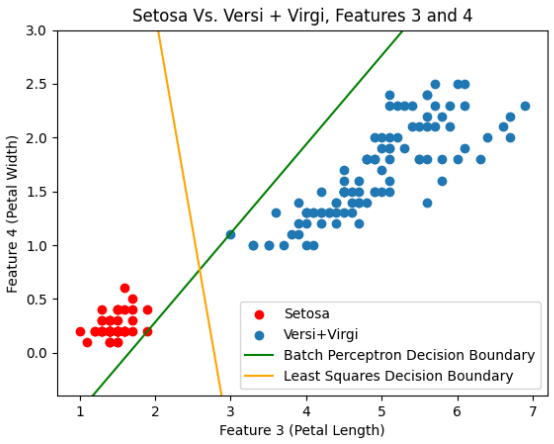
19 epochs for Batch Perceptron to converge. 0 misclassifications. Too many dimensions to plot.

```
Setosa Vs. Versi + Virgi, All Features
Epochs:  19
Weights:  [[-0.11290307345389104]
 [0.18494541330520786]
 [-0.190499357902406]
 [0.3497683882858727]
 [0.26628557457771096]]
Weights (LS):  [[ 0.08288849]
 [ 0.34562494]
 [-0.41679651]
 [-0.18430551]
 [-0.15787014]]
Misclassifications:  0
```

Setosa Vs. Versi+Virigi	Features 3 and 4 Only	Batch_Perceptron and LS

255 epochs. 0 misclassifications.

```
Setosa Vs. Versi + Virgi, Features 3 and 4
Epochs:  255
Weights:  [[-0.32271161869398424]
 [0.3901866248701473]
 [0.5352014598032334]]
Weights (LS):  [[-0.45199065]
 [-0.11189485]
 [ 1.25747866]]
Misclassifications:  0
```



Virgi Vs. Versi+Setosa	All Features	Batch Perceptron and LS

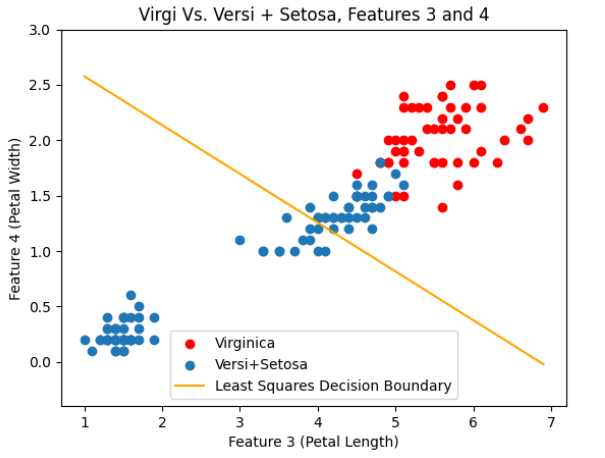
Perceptron never converged. 36 misclassifications with Least Squares. Too many dimensions to plot.

```
Virgi Vs. Versi + Setosa, All Features
Epochs:  1000
Weights:  [[0.11330213021589283]
 [0.18031505479030083]
 [-0.28815270114650426]
 [-0.4115099249770564]
 [0.8122615420350183]]
Weights (LS):  [[ 0.03074673]
 [-0.36132918]
 [-0.29673611]
 [-0.38089105]
 [ 2.43280256]]
Misclassifications:  36
```

Virgi Vs. Versi+Setosa	Features 3 and 4 Only	Batch Perceptron and LS

Perceptron never converges because they are not linearly separable. Least Squares gets 37 misclassifications although that can be modified a bit if you bias it a bit, there's twice as many samples in the one class than the other which is part of what's biasing it so far the other way.

```
Virgi Vs. Versi + Setosa, Features 3 and 4
Epochs:  1000
Weights:  [[-0.11166296941321384]
 [-0.09876033747494113]
 [0.6895414905804511]]
Weights (LS):  [[-0.20922401]
 [-0.47502808]
 [ 1.43369154]]
Misclassifications:  37
[[-0.20922401]
 [-0.47502808]
 [ 1.43369154]]
```



Setosa Vs. Versi Vs. Virigi	Features 3 and 4 Only	Multiclass LS

Least Squares has a pretty hard time with this because Petal Length and Petal Width are highly correlated especially between Versicolor and Virginica. 50 misclassifications

```
Setosa Vs. Versi Vs. Virgi, Features 3 and 4
Weights (LS):  [[-0.22599533  0.12138332  0.104612  ]
 [-0.05594743 -0.18156661  0.23751404]
 [ 1.12873933  0.08810644 -0.21684577]]
Misclassifications:  50
```