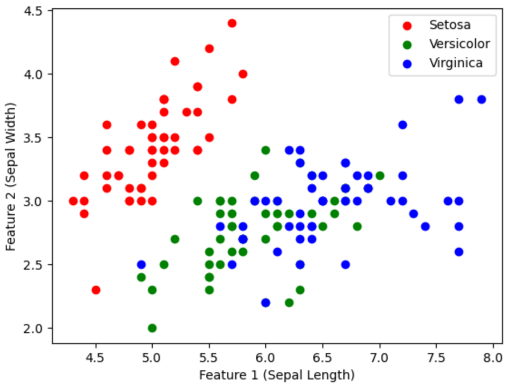
Pattern Recognition

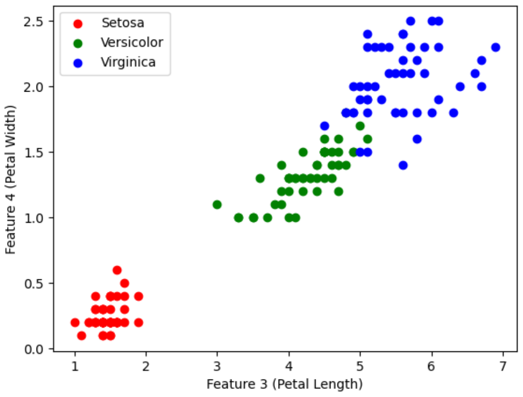
ECE 4363 / ECE 5363

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?

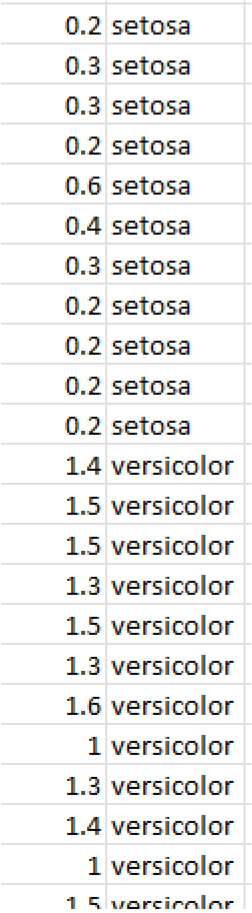
There is three classes in the data, three species of this Iris flower apparently. There are four features, sepal length, sepal width, pedal length and pedal width. Which appear to be measurements of the flower. The sepal is the outer green petals that protect the flower before it blooms, they form the 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.





Above I've plotted the sepal length against the sepal width and petal length against petal width for the 3 features which you can see is linearly separable for Setosa versus the other two, but not as much between the other two especially with the first two features which looks almost completely mixed together and therefore just not distinguishable between the classes. I would expect the second to features to perform well and the first two not to, especially the second feature.

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.

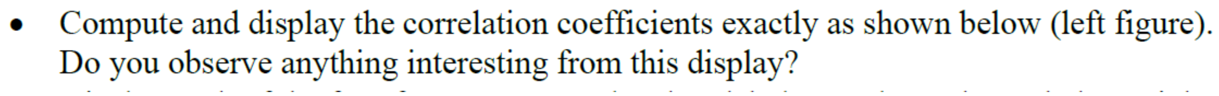


0.2	setosa
0.3	setosa
0.3	setosa
0.2	setosa
0.6	setosa
0.4	setosa
0.3	setosa
0.2	setosa
0.2	setosa
0.2	setosa
0.2	setosa
1.4	versicolor
1.5	versicolor
1.5	versicolor
1.3	versicolor
1.5	versicolor
1.3	versicolor
1.6	versicolor
1	versicolor
1.3	versicolor
1.4	versicolor
1	versicolor
1.5	versicolor

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

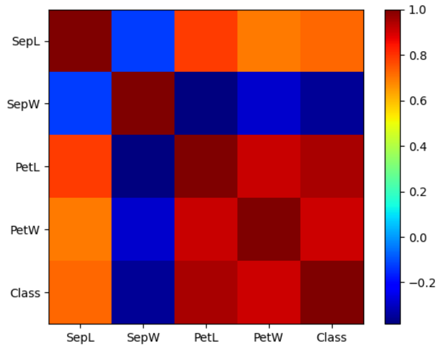
	Sepal Length	Sepal Width	Pedal Length	Pedal Width
Minimum	4.3	2.0	1.0	0.1
Maximum	7.9	4.4	6.9	2.5
Mean	5.8	3.1	3.8	1.2
Variance	0.7	0.2	3.1	0.6
Within-Class Variance	0.26	0.11	0.18	0.04
Between-Class Variance	0.42	0.08	2.91	0.54

The sepal size is actually bigger than the petal size which is interesting and matches up with the pictures of the Irises that I've seen so that makes sense. The variance being so low for some measurements seems bad because the less variance there is in general the harder it is to tell things apart logically. Especially sepal width has very low variance, but petal length has very high variance so that seems good. The within-class variance being small also seems like a good thing because that implies that there is a smaller target to hit, just like the larger between-class variances imply that there is a lot of space between the targets. Especially with petal length the between-class variance is huge while the in-class variance is very small that should make a great predictor. On the other hand, the within-class variance being higher than the between-class variance for sepal width means that there is more difference within a class than between them and therefore it will be impossible to tell them apart with that metric. It seems that you want the between-class variance to be big, the in-class variance to be small, and the difference between them to be big.

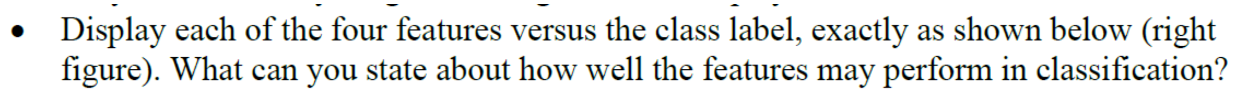


- Compute and display the correlation coefficients exactly as shown below (left figure). Do you observe anything interesting from this display?

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.

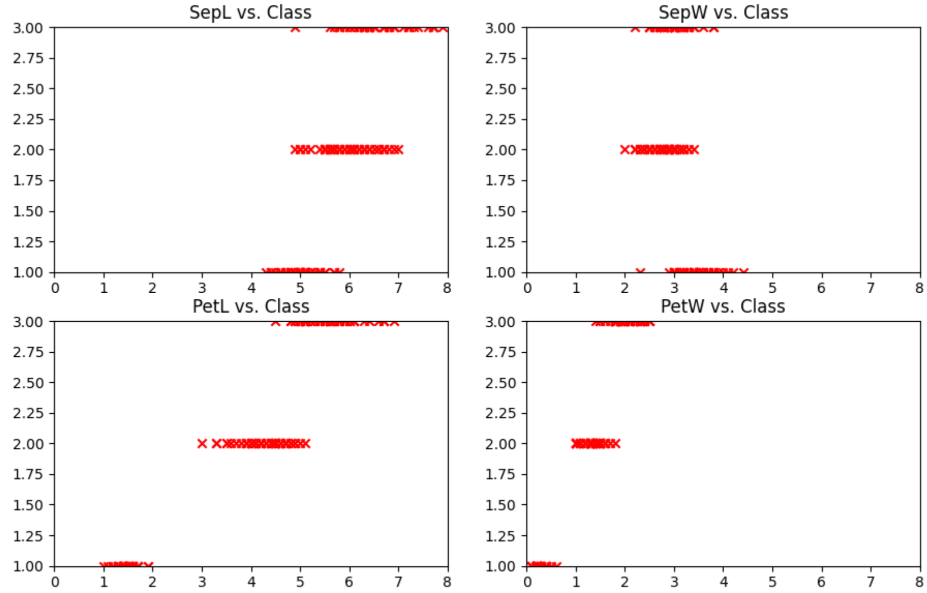


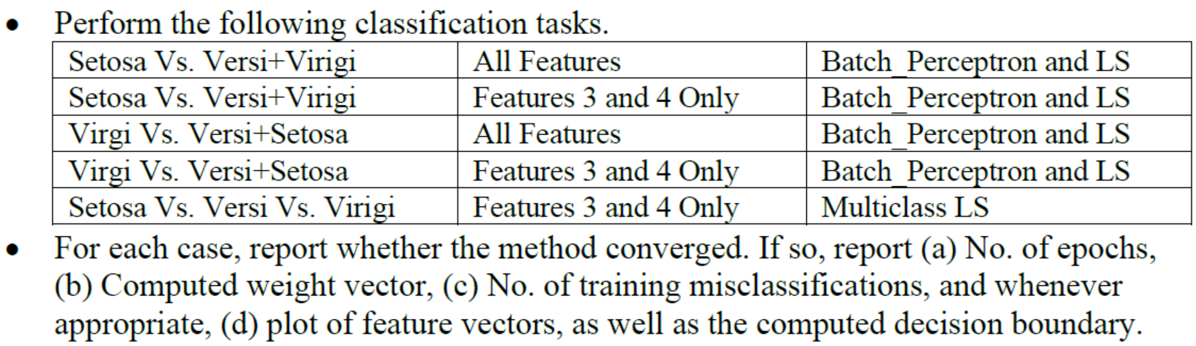
Petal length and class are also highly correlated so it seems that will be our greatest predictor as shown before with variances. Class and sepal width are also very lowly correlated.



- 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.

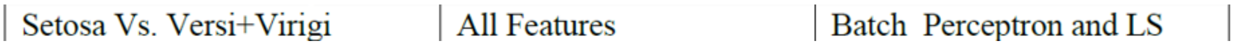




- Perform the following classification tasks.

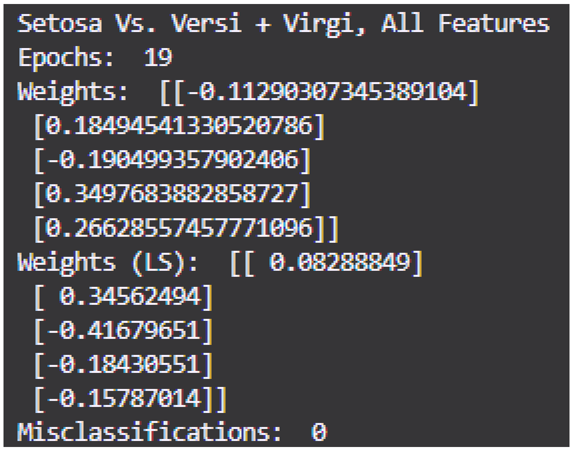
Setosa Vs. Versi+Virigi	All Features	Batch_Perceptron and LS
Setosa Vs. Versi+Virigi	Features 3 and 4 Only	Batch_Perceptron and LS
Virgi Vs. Versi+Setosa	All Features	Batch_Perceptron and LS
Virgi Vs. Versi+Setosa	Features 3 and 4 Only	Batch_Perceptron and LS
Setosa Vs. Versi Vs. Virigi	Features 3 and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

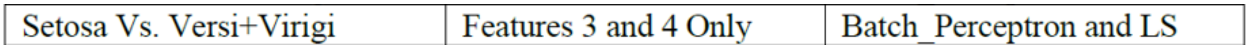


Setosa Vs. Versi+Virigi	All Features	Batch_Perceptron and LS

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

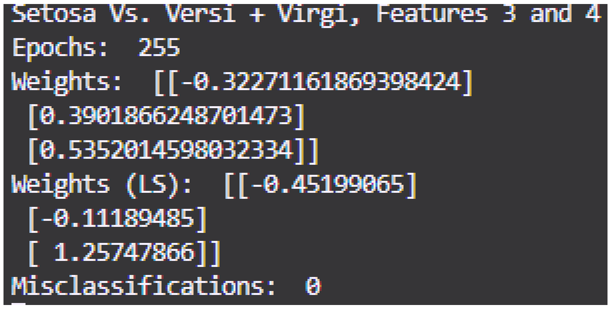


```
Setosa Vs. Versi + Virgi, All Features
Epochs:  19
Weights:  [[-0.11290307345389104]
 [0.18494541330520786]
 [-0.190499357902406]
 [0.3497683882858727]
 [0.26628557457771096]]
Weights (LS):  [[ 0.08288849]
 [ 0.34562494]
 [-0.41679651]
 [-0.18430551]
 [-0.15787014]]
Misclassifications:  0
```

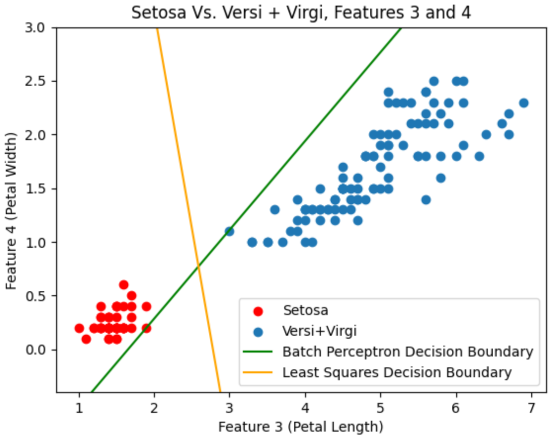


Setosa Vs. Versi+Virigi	Features 3 and 4 Only	Batch_Perceptron and LS

255 epochs. 0 misclassifications.

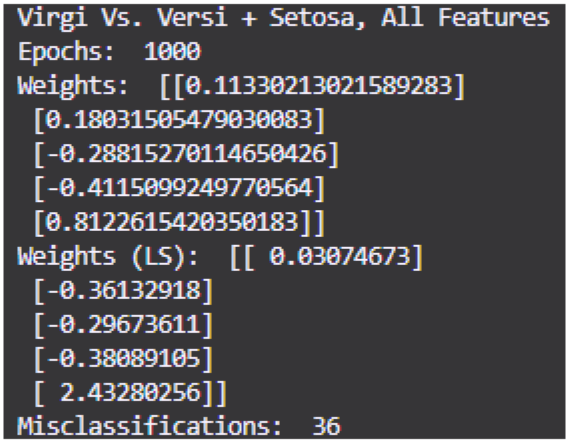


```
Setosa Vs. Versi + Virgi, Features 3 and 4
Epochs:  255
Weights:  [[-0.32271161869398424]
 [0.3901866248701473]
 [0.5352014598032334]]
Weights (LS):  [[-0.45199065]
 [-0.11189485]
 [ 1.25747866]]
Misclassifications:  0
```



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

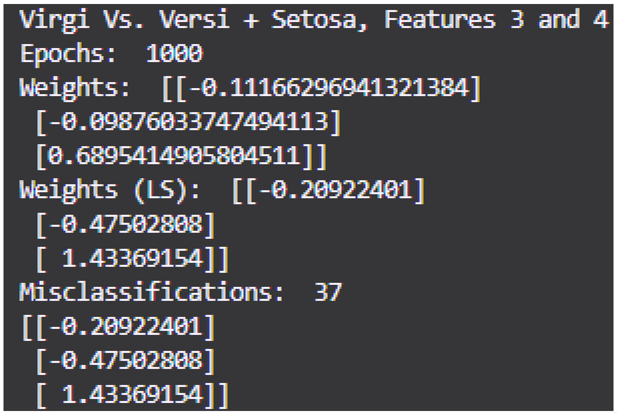
Perceptron never converged. 36 misclassifications with Least Squares. Too many dimensions to plot.



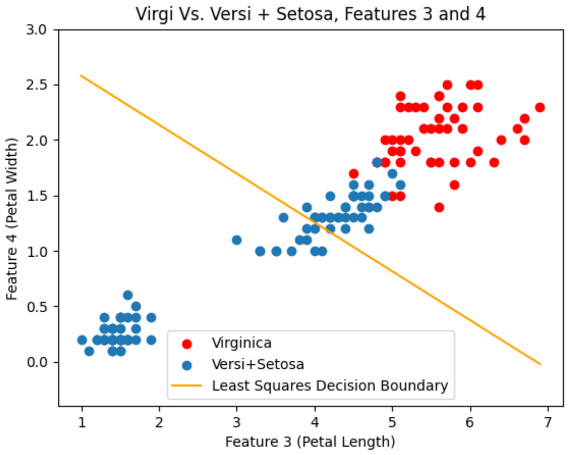
```
Virgi Vs. Versi + Setosa, All Features
Epochs:  1000
Weights:  [[0.11330213021589283]
 [0.18031505479030083]
 [-0.28815270114650426]
 [-0.41150992497705 64]
 [0.8122615420350183]]
Weights (LS):  [[ 0.03074673]
 [-0.36132918]
 [-0.29673611]
 [-0.38089105]
 [ 2.43280256]]
Misclassifications:  36
```

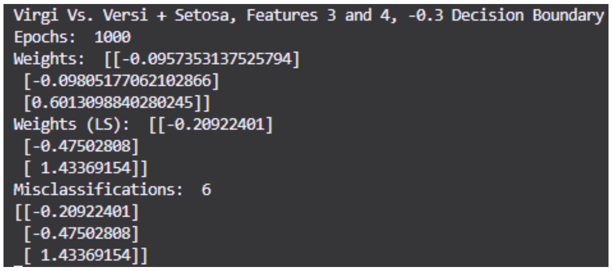
Virgi Vs. Versi+Setosa	Features 3 and 4 Only	Batch Perceptron and LS

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. You can see with the second plot I did below it where I moved the decision boundary 0.3 that the misclassifications went down and you can see visually that it's a better decision boundary there.

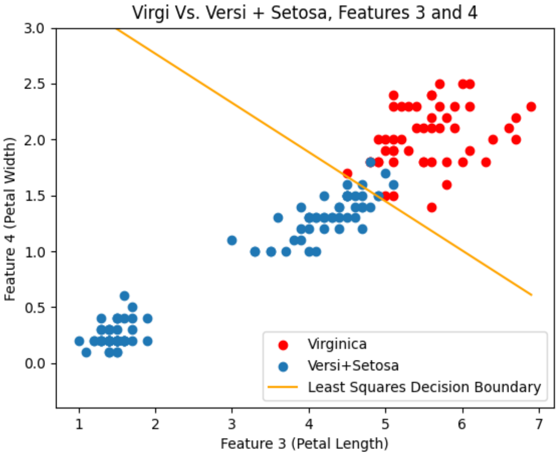


```
Virgi Vs. Versi + Setosa, Features 3 and 4
Epochs:  1000
Weights:  [[-0.11166296941321384]
 [-0.09876033747494113]
 [0.6895414905804511]]
Weights (LS):  [[-0.20922401]
 [-0.47502808]
 [ 1.43369154]]
Misclassifications:  37
[[-0.20922401]
 [-0.47502808]
 [ 1.43369154]]
```



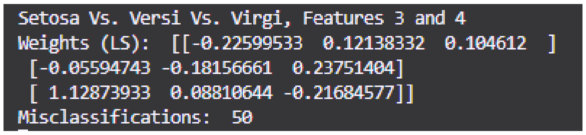


```
Virgi Vs. Versi + Setosa, Features 3 and 4, -0.3 Decision Boundary
Epochs:  1000
Weights:  [[-0.0957353137525794]
 [-0.09805177062102866]
 [0.6013098840280245]]
Weights (LS):  [[-0.20922401]
 [-0.47502808]
 [ 1.43369154]]
Misclassifications:  6
[[-0.20922401]
 [-0.47502808]
 [ 1.43369154]]
```

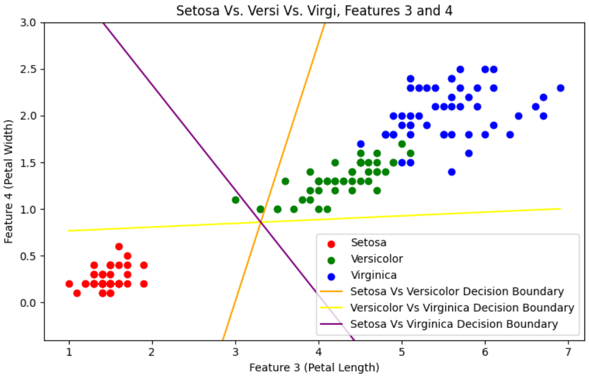


Setosa Vs. Versi Vs. Virigi | Features 3 and 4 Only | Multiclass LS

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
```


Setosa Vs. Versi Vs. Virgi, Features 3 and 4
Feature 4 (Petal Width)
Feature 3 (Petal Length)
Setosa
Versicolor
Virginica
Setosa Vs Versicolor Decision Boundary
Versicolor Vs Virginica Decision Boundary
Setosa Vs Virginica Decision Boundary