ECE 4363 – Pattern Recognition
Mohak Kant, R11481106

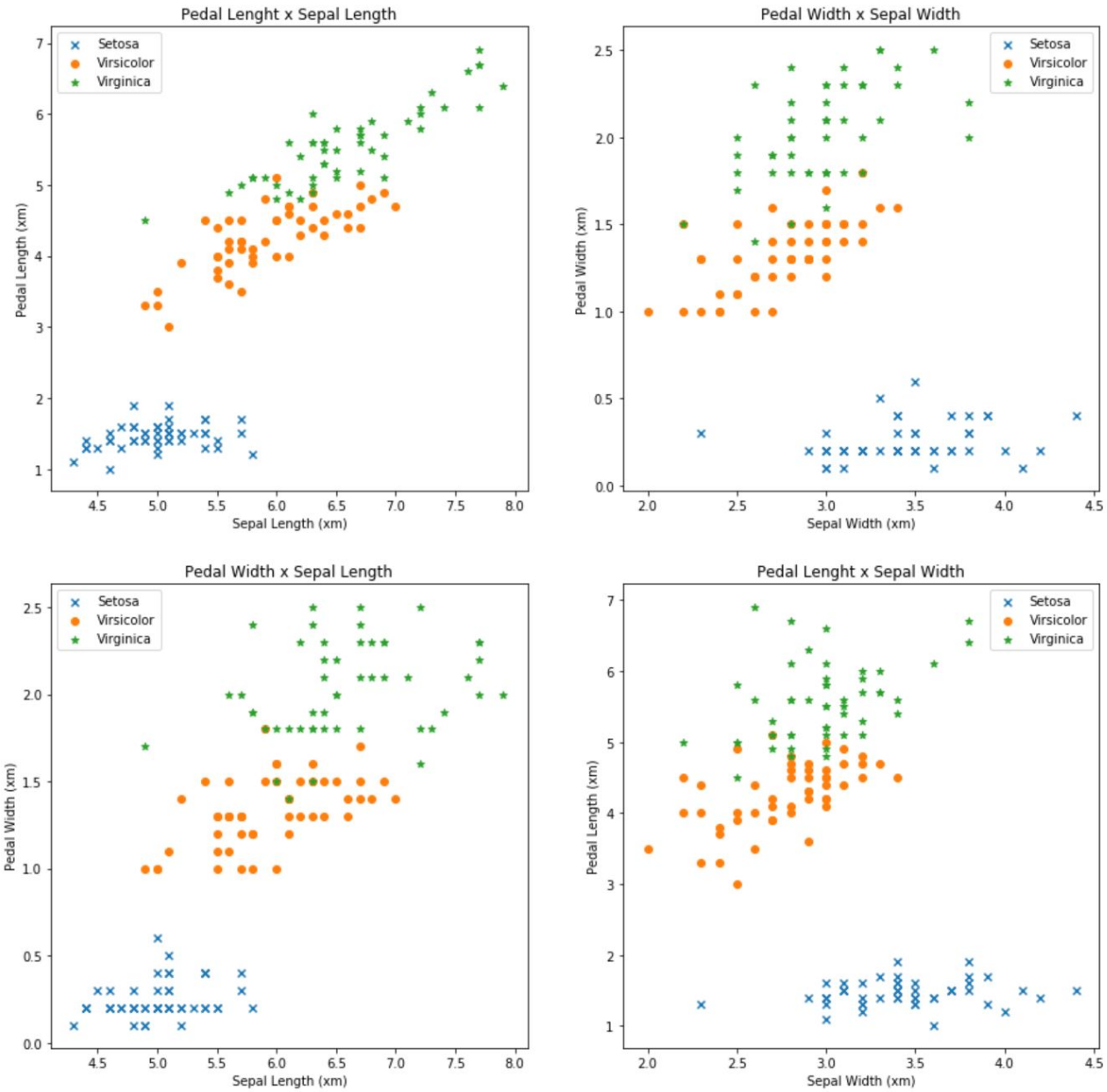
Project 1

1. Dataset & this project

The Fisher's Iris dataset is a multivariate data set consisting of measurements from three species of flowers from the genus Iris. It was introduced by the British statistician and biologist Ronald Fisher.

The features collected by Ronald Fisher, are represented in the image below. The Petal being more commonly known is the flower and the sepal is the leafy part below the flower.

The dataset consists of three different classes - setosa, virginca, and *versicolor*. Each class has 50 data points and 4 feature measurements: petal width, petal length, sepal width, and sepal length. The following are the plots for the three different classes:



Intuition: It can be seen from the plots that Setosa is a likely candidate for being linearly separable while the other two classes have quite a bit of overlap. Looking at these plots, I would assume that selecting only two features at a time, the Batch Perceptron will probably not work as well while trying to separate Versicolor and Virgicia but Least-Squares will probably still find a boundary. However, when using all the 4 features, the Batch perceptron might converge as these classes may be separable in a higher dimensional space.

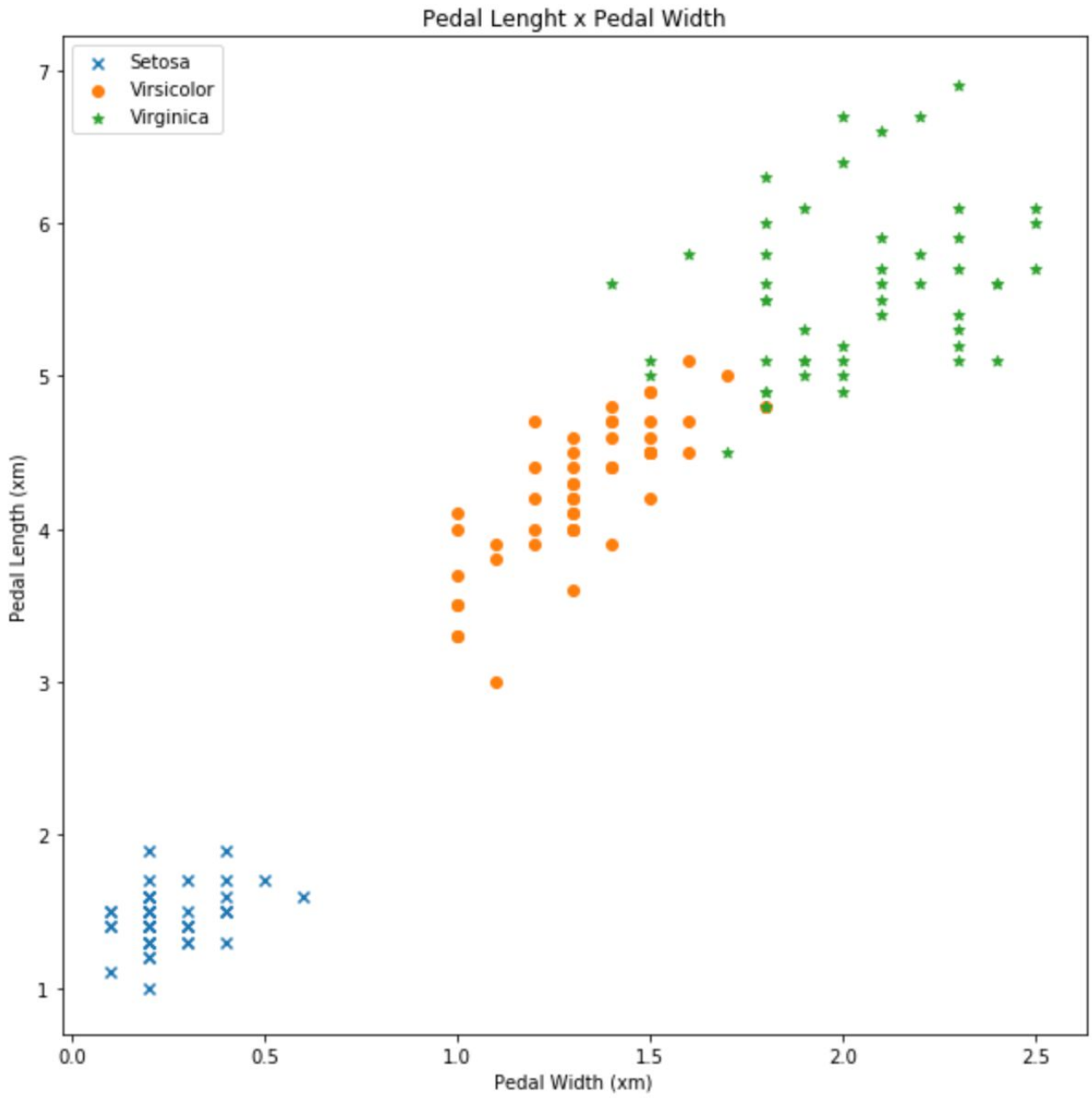
This project's goal consists of exploring and visualizing the data and later using statistical knowledge, mainly that of a Linear Discriminant Algorithm, called the Perceptron and Least-Squares to try to seperate the distinct classes present in the dataset.

2. Statistics

	Sepal Length	Sepal Width	Petal Length	Petal Width
Minimum	4.3000	2.0000	1.0000	0.1000
Maximum	7.9000	4.4000	6.9000	2.5000
Mean	5.8433	3.0573	3.7580	1.1993
Variance	0.6811	0.1887	3.0955	0.5771
Within-Class Variance	0.2597	0.1130	0.1814	0.0410
Between-Class Variance	0.4214	0.0756	2.9140	0.5360

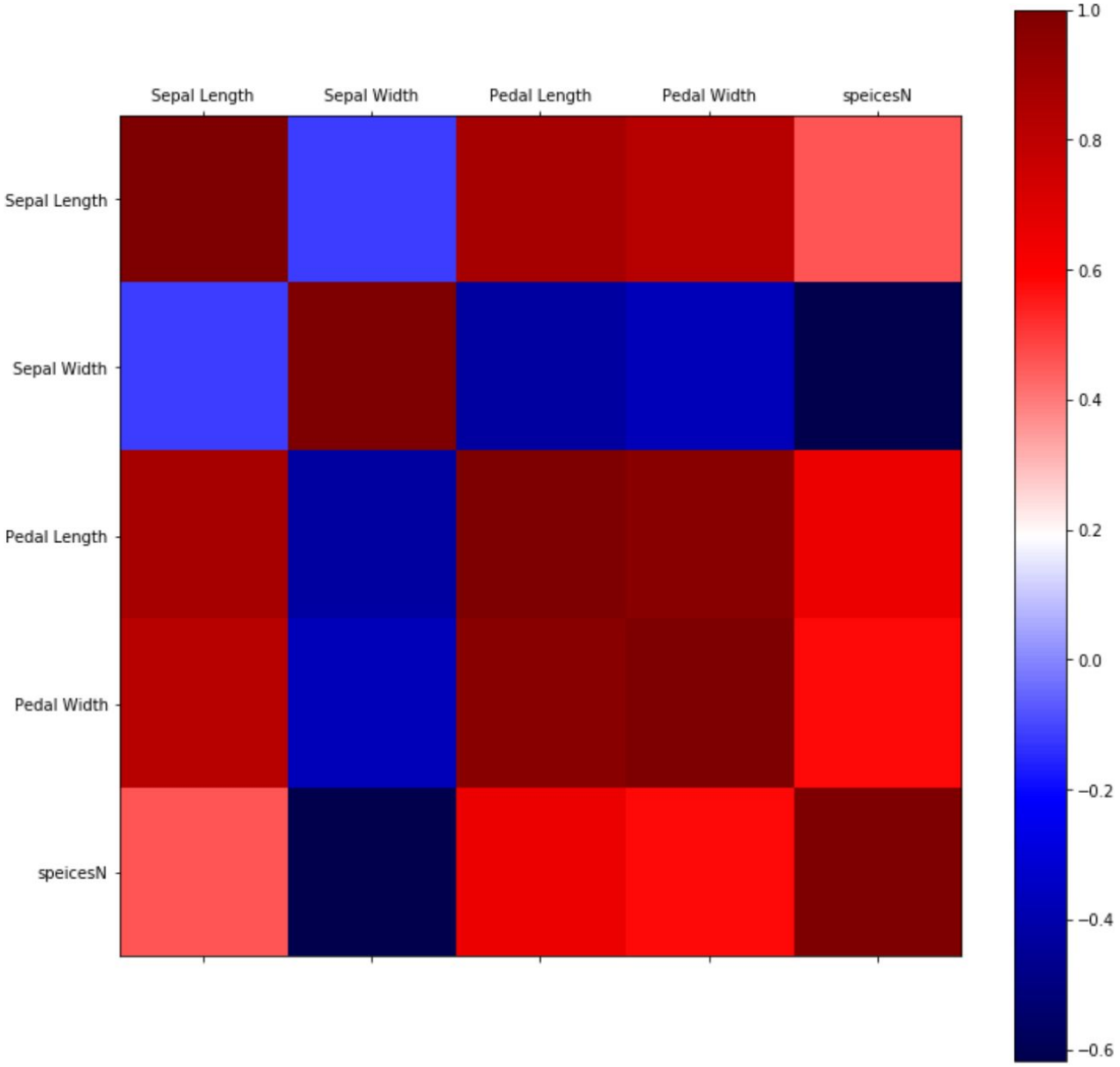
Table 1: statistics for studying the samples

From the statistics above, Petal Length and Petal Width have the least Within-class variance and comparatively (compared to other features) have a higher between class variance. Thus, the different classes based on these two features should be clustered in itself and away from others.



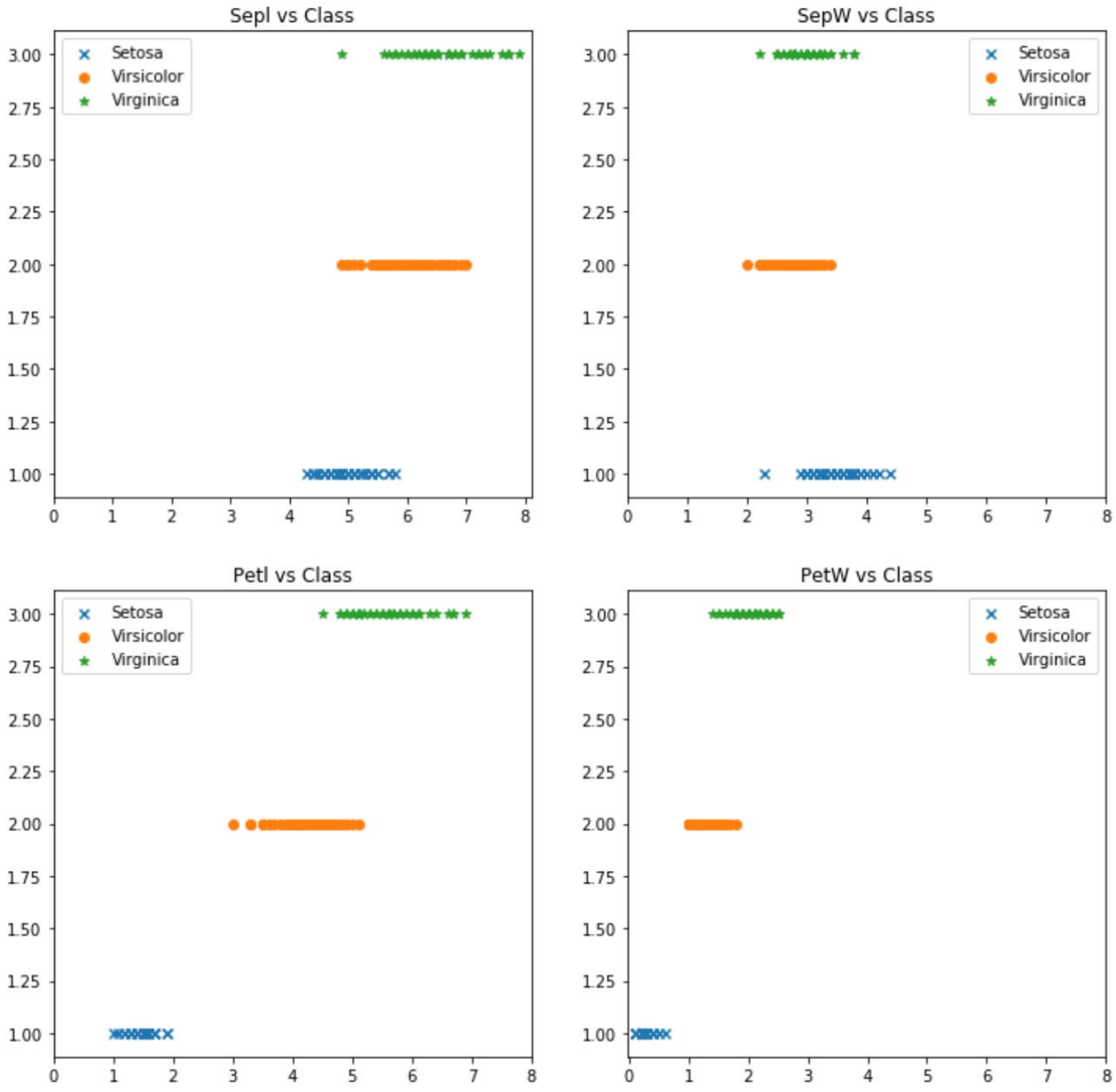
Also, here the Variance is the sum of the Within-Class Variance and the Between-Class Variance.

3. Correlation coefficients



The correlation map demonstrates the linear dependence of the features. The darker red and the darker blue regions formed by the intersection of their corresponding features demonstrate correlation coefficients close to +1/-1(imagine an uphill or a downhill line), signifying lying along lines rather than clusters. Whereas, lighter blue regions with coefficients close to 0, signify linear independence and clustering and thus would be better candidates for discrimination amongst classes.

4. Features VS class labels



From these plots, is is evident that Petal Length and Petal Width are better candidates for classification as they are more separable by straight lines along the X-axis. Moreover, the class Setosa seems to be totally linearly separable.

5. Classification tasks

5.1. Setosa Vs. (Versicolor + Virginica), all features.

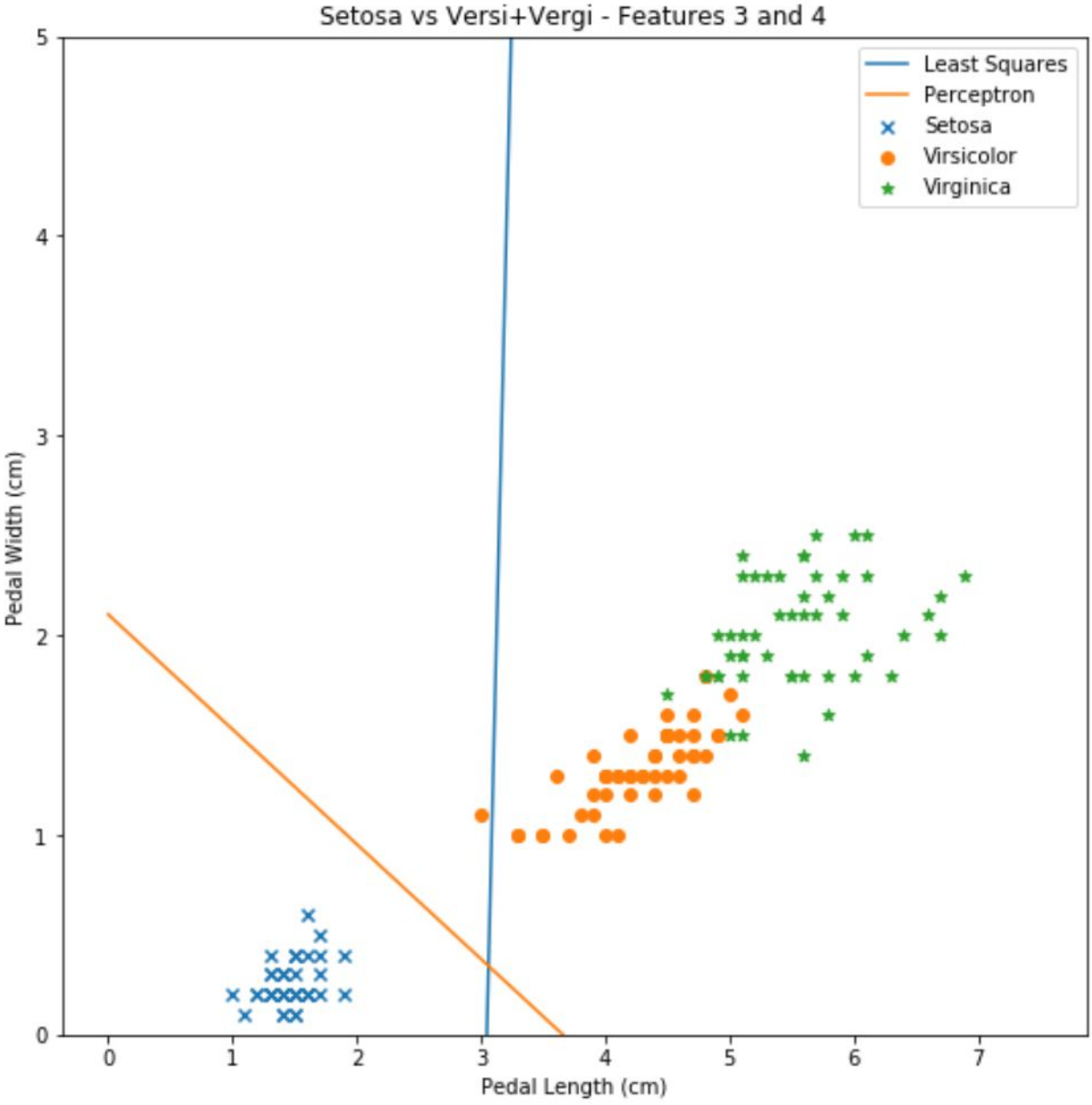
- Perceptron - epochs programmed: 1000
- Perceptron - Epochs taken to reach abs(w - w_previous) < 10^{-6}: 8
- Perceptron - Did it converge? Yes
- Least-squares - misclassifications: 0
- computed weight vector:

w	Perceptron	least-squares
w_1	11.01	0.13206
w_2	27.34	0.4857
w_3	-38.37	-0.44931
w_4	-17.6	-0.11495
$w_{5 =} w_{L+1}$	5.5	-0.76355

5.2. Setosa Vs. (Versicolor + Virginica), features 3 and 4 (Petal Length, Petal Width).

- Perceptron - epochs programmed: 1000
- Perceptron - Epochs taken to reach abs(w - w_previous) < 10^{-6}: 8
- Perceptron - Did it converge? Yes
- Least-squares - misclassifications: 1
- computed weight vector:

w	Perceptron	least-squares
w_1	-5.41	-0.50266
w_2	-9.41	0.019669
$w_{3 =} w_{L+1}$	19.8	1.5321


Setosa vs Versi+Vergi - Features 3 and 4
Least Squares
Perceptron
Setosa
Virsicolor
Virginica
Pedal Width (cm)
Pedal Length (cm)
0
1
2
3
4
5
6
7

5.3. Virginica Vs. (Setosa + Versicolor), all features.

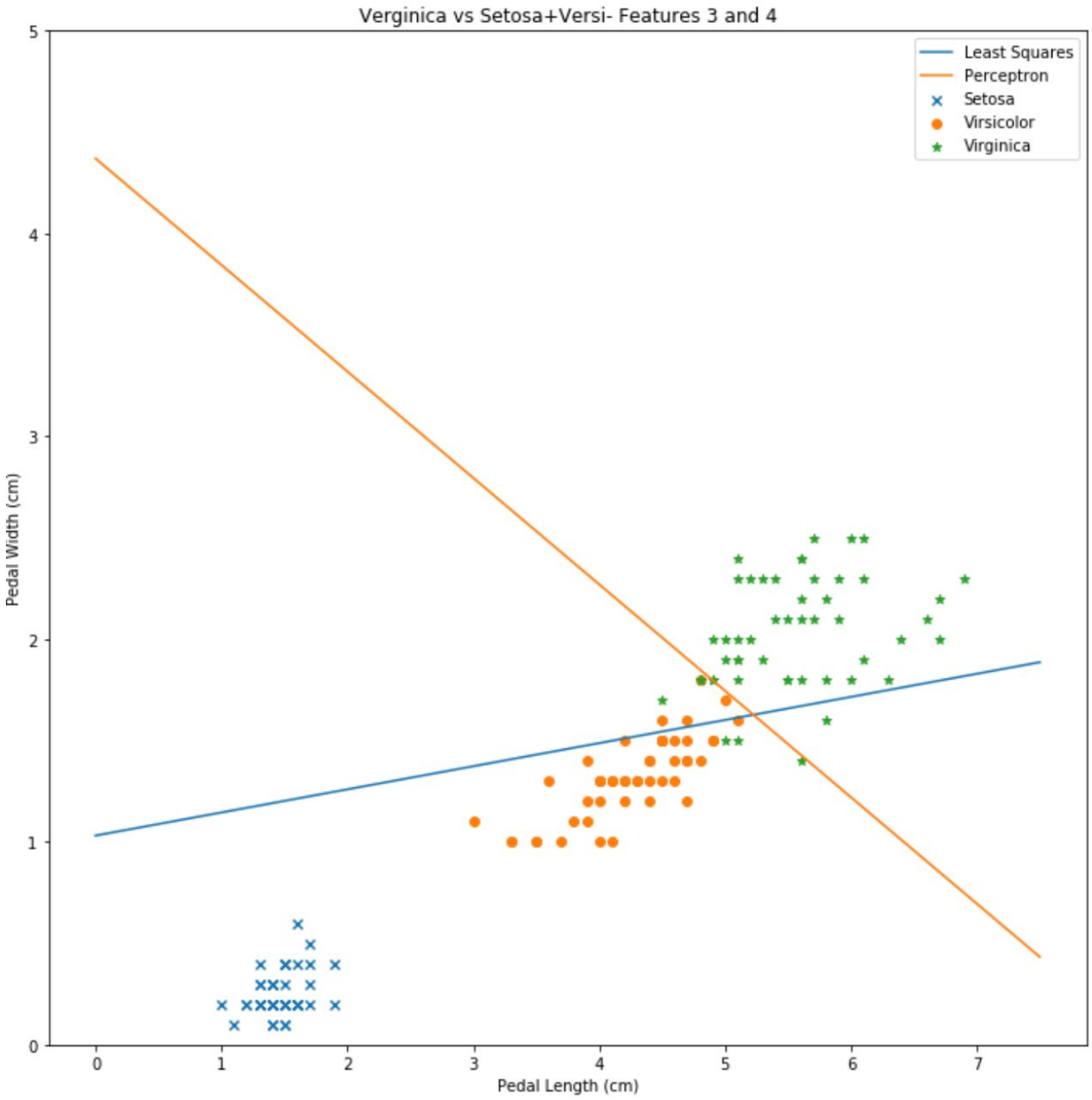
- Batch perceptron - epochs programmed: 10000
- Batch Perceptron - Did it converge? no
- Least-squares - misclassifications: 11
- computed weight vector:

w	Perceptron	least-squares
w_1	-63.04	-0.091752
w_2	-76.64	0.40554
w_3	111.92	0.0079758
w_4	137.38	1.1036
$w_5 = w_{L+1}$	-168.2	-2.3906

5.4. Virginica Vs. (Setosa + Versicolor), features 3 and 4 (Petal Length, Petal Width).

- Perceptron - epochs programmed: 10000
- Perceptron - Did it converge? no
- Least-squares - misclassifications: 8
- computed weight vector:

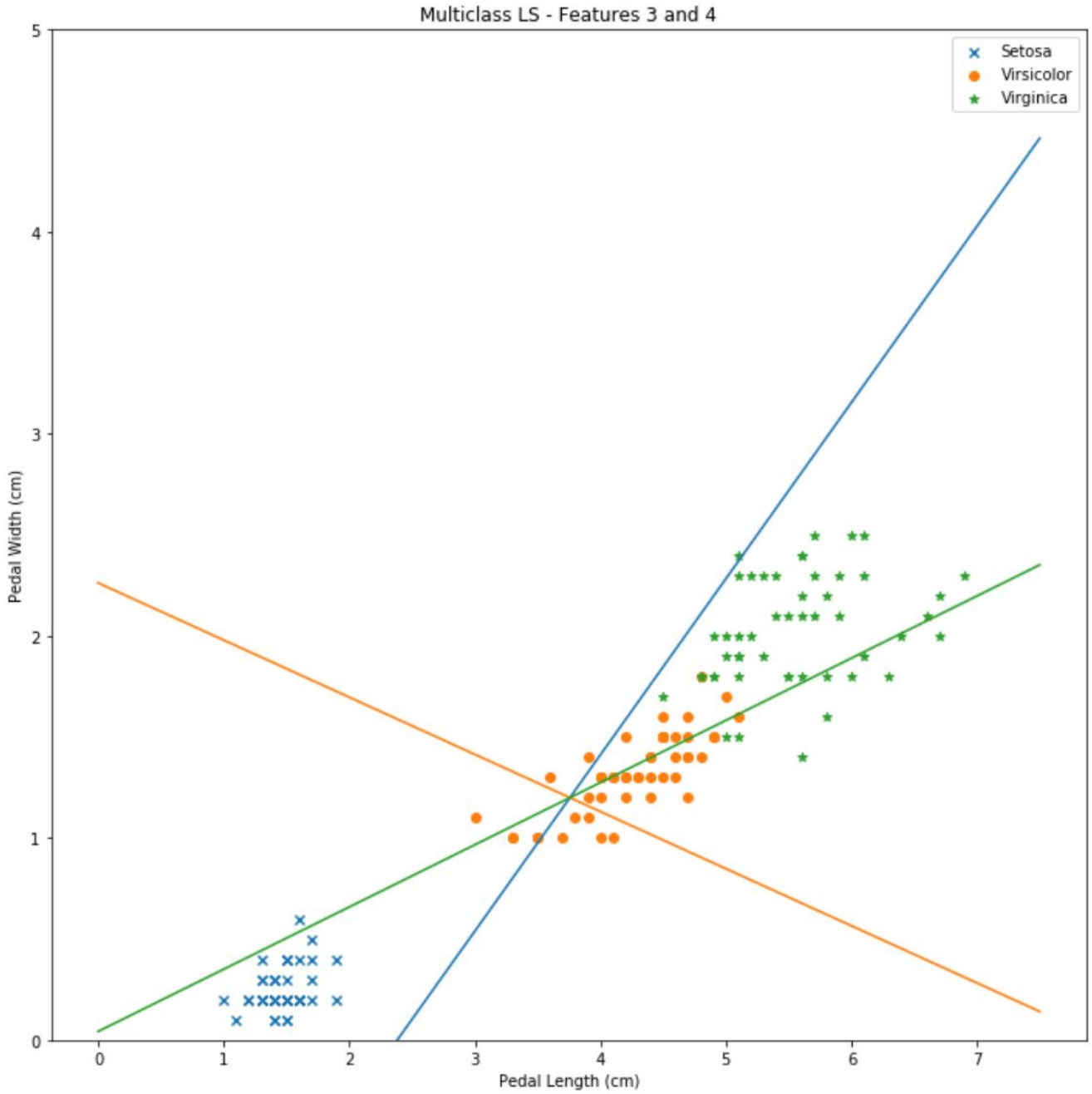
w	Perceptron	least-squares
w_1	23.74	-0.5026
w_2	45.21	0.01966
$w_3 = w_{L+1}$	-197.6	1.53206



6. Multiclass least-squares, features 3 and 4 (Petal Length, Petal Width).

- misclassifications: 34
- computed weight vector:

w			
w_1	-0.25132	0.3243	-0.0730
w_2	0.00983	-0.65008	0.640255
$w_{3=}w_{L+1}$	1.26603	-0.10584	-0.16018


Multiclass LS - Features 3 and 4
Setosa
Virsicolor
Virginica
Pedal Width (cm)
Pedal Length (cm)
0
1
2
3
4
5
0
1
2
3
4
5
6
7

References

Correlation Coefficients:
https://www.dummies.com/education/math/statistics/how-to-interpret-a-correlation-coefficient-r/