SURF **SURF**

First Progress Report

Name: Nishanth Bhaskara **July 7, 2016**

A) Motivation

Machine learning is a field of computer science that draws its origins from pattern recognition, computational learning theory, and artificial intelligence. The basic idea is for a program to be able to learn from its environment and solve problems without being explicitly programmed to do so. The exact way this works is by examining inputs and drawing comparisons between them in order to predict the result with a certain degree of probability, as opposed to solving the problem in a deterministic way. This process saves both time and resources when approaching certain problems in industry and science.

Machine learning is having more and more use in the modern world. In the last decade, several fields have experienced a humongous growth in the volume, complexity, and lack of clarity of simulated and measured results. An example of this can be seen in astronomy. Panoramic digital detectors cause the amount of processed synaptic sky curves to double every 12-18 months. Now, the main issue that arises is the inability to process this huge amount of data and draw valid conclusions in a realistic amount of time.

This is where machine learning comes in. Using Google's powerful TensorFlow library, we will try to solve these emerging problems in industry using machine learning techniques. Some of these techniques we will investigate and use are deep neural networks and convolutional neural networks.

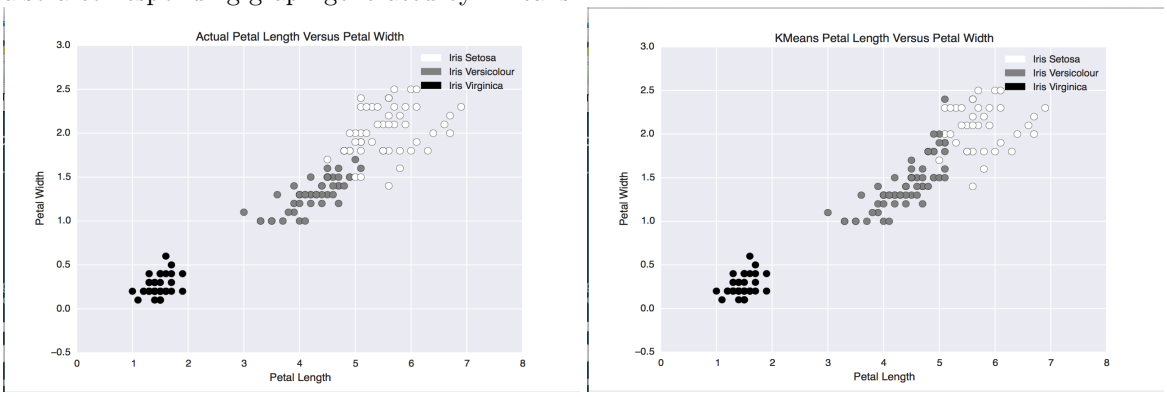
B) Current Work

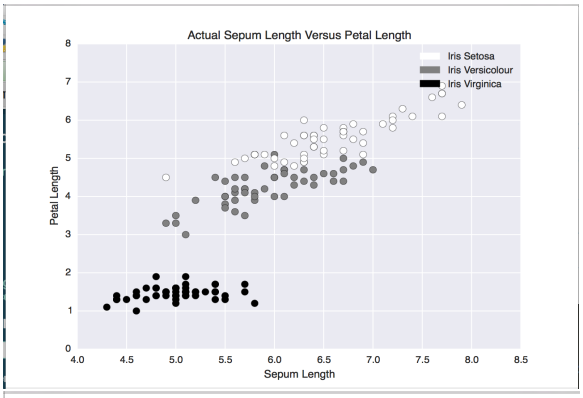
Currently, we are working with smaller data sets in order to gain mastery of TensorFlow first, and then scale up after we have done so. For this, we are using the Iris Data Set measured in the 1930s which has data of different physical features of flowers.

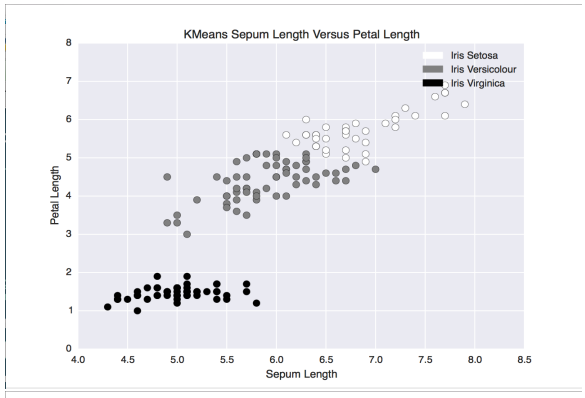
We are given 150 data points, each of which are 4 - dimensional vectors, and a corresponding class which corresponds to the type of flower that it is. The goal is to use the Tensor Flow library and the KMeans algorithm, deep neural networks, or convolutional neural networks to be able to cluster this data set into 3 groupings, and try to achieve accurate classification.

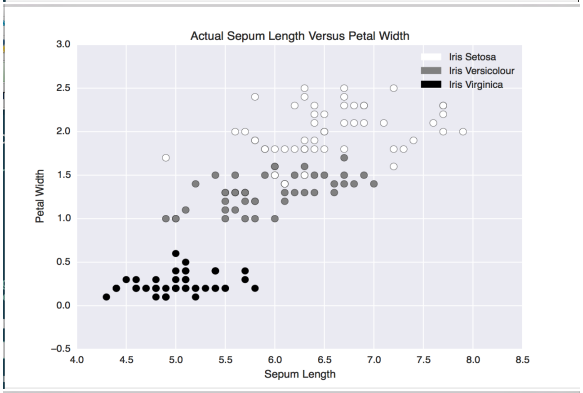
Part I: KMEANS Results

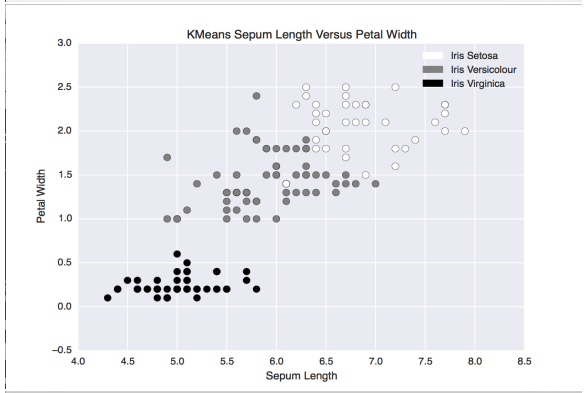
The result is best depicted by graphs. As we are only able to display 2 dimensions at a time and we have 4 dimensional data points, we need $\binom{4}{2} = 6$ graphs to depict our results. For each of these 6 graphs, there is also a corresponding graph generated by kMeans.

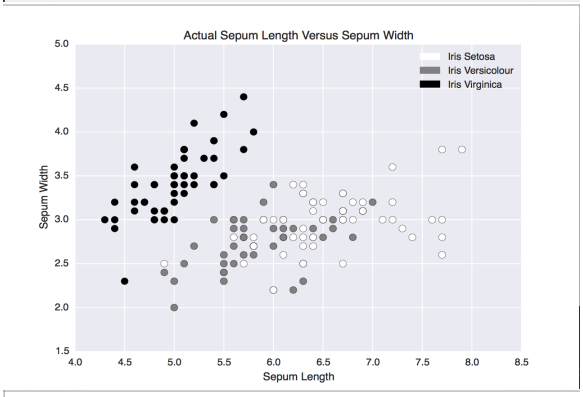


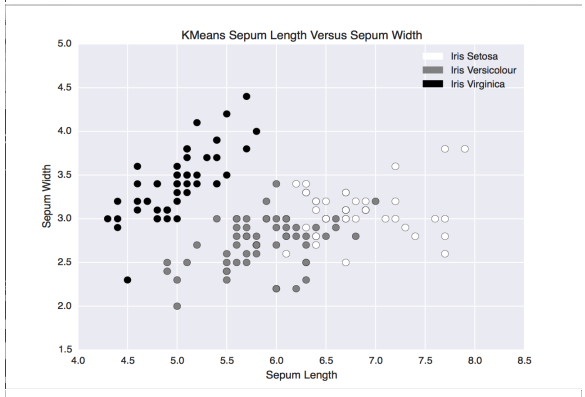

Actual Sepum Length Versus Petal Length
Iris Setosa
Iris Versicolour
Iris Virginica
Petal Length
Sepum Length

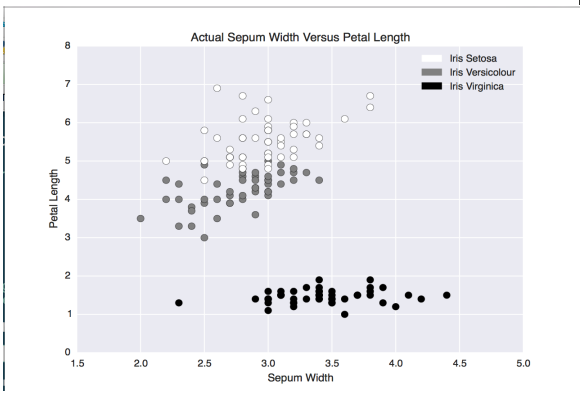

KMeans Sepum Length Versus Petal Length
Iris Setosa
Iris Versicolour
Iris Virginica
Petal Length
Sepum Length

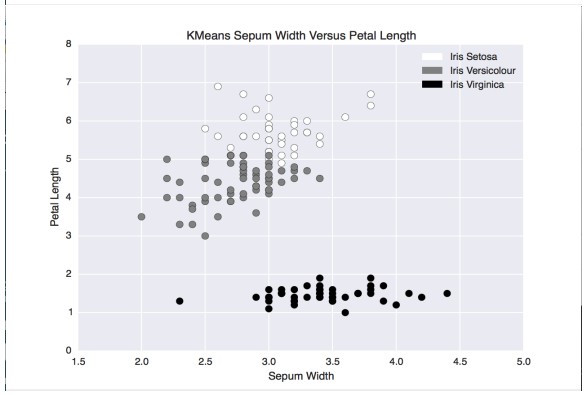

Actual Sepum Length Versus Petal Width
Iris Setosa
Iris Versicolour
Iris Virginica
Petal Width
Sepum Length

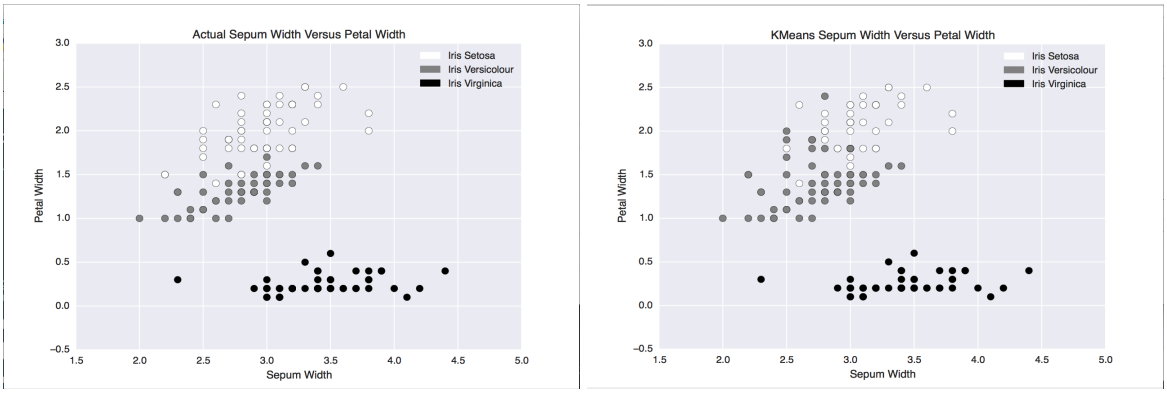

KMeans Sepum Length Versus Petal Width
Iris Setosa
Iris Versicolour
Iris Virginica
Petal Width
Sepum Length


Actual Sepum Length Versus Sepum Width
Iris Setosa
Iris Versicolour
Iris Virginica
Sepum Width
Sepum Length


KMeans Sepum Length Versus Sepum Width
Iris Setosa
Iris Versicolour
Iris Virginica
Sepum Width
Sepum Length


Actual Sepum Width Versus Petal Length
Iris Setosa
Iris Versicolour
Iris Virginica
Petal Length
Sepum Width


KMeans Sepum Width Versus Petal Length
Iris Setosa
Iris Versicolour
Iris Virginica
Petal Length
Sepum Width




```
The accuracy for kMeans for the first class is 1.0
The accuracy for kMeans for the second class is 0.96
The accuracy for kMeans for the third class is 0.96
```

The results were surprisingly good. As expected, the Iris Virginica had a success rate of 100% as it can be linearly separated by a hyperplane from the other two classes (as we can see in the graphs). Despite the fact that Iris Versicolour and and Iris Setosa were not as clearly separable, our algorithm was still able to achieve an accuracy of 96%.

PART 2: Neural Networks Results
Below we have the confusion matrix for our results

	Actual Class 1	Actual Class 2	Actual Class 3
Predicted Class 1	8	0	0
Predicted Class 2	0	14	0
Predicted Class 3	0	1	17

Below is the accuracy of our DNN on our test set.

```
Accuracy: 0.933333
```

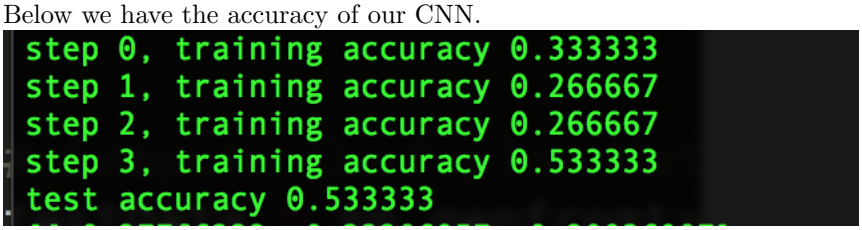
Lastly, below we have the probabilities the classifier gave to each sample being in a class. In red is the correct class, and if there is no blue, that means the highest percentage also matched with the correct class. If there is blue text, it corresponds to the largest percentage/the classification that the DNN gave the sample.

Sample	Class 1	Class 2	Class 3	Correct Class
1	0.00255513	0.963146	0.0342992	1
2	6.48E-06	0.184529	0.815465	2
3	0.992481	0.0075186	1.13E-09	0
4	0.00165332	0.963993	0.0343533	1
5	0.00277461	0.918272	0.0789532	1
6	0.00380918	0.982331	0.01386	1
7	0.998071	0.00192946	5.75E-12	0
8	0.00012001	0.551666	0.448214	2
9	0.00522768	0.977295	0.0174774	1
10	2.44E-07	0.0207774	0.979222	2
11	3.47E-07	0.0387548	0.961245	2
12	0.992982	0.00701811	1.42E-09	0
13	2.34E-06	0.0456054	0.954392	2
14	0.00188548	0.852341	0.145773	1
15	0.00647622	0.975394	0.0181302	1
16	0.996285	0.00371502	9.37E-11	0
17	0.0422964	0.95599	0.00171316	1
18	0.994022	0.00597789	7.13E-10	0
19	0.995915	0.00408537	1.41E-10	0
20	6.72E-07	0.0203202	0.979679	2
21	0.996408	0.00359225	8.12E-11	0
22	0.00080587	0.827074	0.172121	1
23	1.56E-06	0.0597539	0.940245	2
24	0.00011481	0.482809	0.517077	1
25	0.00886843	0.981563	0.00956857	1
26	0.00130325	0.940863	0.0578341	1
27	0.997549	0.00245088	1.59E-11	0
28	0.00128333	0.971974	0.0267427	1
29	7.31E-08	0.0134275	0.986572	2
30	0.00433767	0.986691	0.00897091	1

PART 3: Convolutional Neural Networks Results
Below we have the confusion matrix for our results.

	Actual Class 1	Actual Class 2	Actual Class 3
Predicted Class 1	8	0	0
Predicted Class 2	0	0	0
Predicted Class 3	0	14	8

Below we have the accuracy of our CNN.



Lastly, below we have the probabilities the classifier gave to each sample being in a class. In red is the correct class, and if there is no blue, that means the highest percentage also matched with the correct class. If there is blue text, it corresponds to the largest percentage/the classification that the CNN gave the sample.

Sample	Class 1	Class 2	Class 3	Correct Class
1	0.371671	0.128967	0.499363	1
2	0.347279	0.124356	0.528366	2
3	0.451697	0.133066	0.415237	0
4	0.374397	0.126433	0.49917	1
5	0.367205	0.131802	0.500992	1
6	0.367615	0.127403	0.504982	1
7	0.490878	0.126206	0.382916	0
8	0.352531	0.124637	0.522832	2
9	0.377021	0.128422	0.494557	1
10	0.332063	0.124131	0.543805	2
11	0.337877	0.121939	0.540184	2
12	0.489055	0.132413	0.378533	0
13	0.357668	0.128142	0.51419	2
14	0.362174	0.132888	0.504938	1
15	0.359206	0.13208	0.508714	1
16	0.47979	0.12973	0.39048	0
17	0.377297	0.132242	0.490461	1
18	0.463917	0.128534	0.407548	0
19	0.474379	0.130029	0.395592	0
20	0.348093	0.129438	0.522469	2
21	0.4654	0.128958	0.405641	0
22	0.369799	0.127585	0.502616	1
23	0.346053	0.12286	0.531087	2
24	0.347834	0.127693	0.524473	1
25	0.37169	0.130374	0.497936	1
26	0.365526	0.126312	0.508163	1
27	0.492564	0.125909	0.381527	0
28	0.362047	0.124957	0.512996	1
29	0.347554	0.123151	0.529295	2
30	0.366439	0.127011	0.506551	1

We obviously see that our DNN had a much higher accuracy than our CNN. We can attribute this to the fact that the small amount of features in our data set don't work well with the CNN methodology. In this case obviously a normal fully connected neural network works much better.

C) How Current Work Influences Future Work
This was very small scale experimentation as we only had around 150 data points of which each only had 4 features. Eventually we scale up to data sets with hundreds of thousands of data points and as many as 50 features. Being able to confidently use TensorFlow as well as mastery of neural networks and other machine learning techniques is essential to progress. So far I have not many any challenges that stopped me from making progress, but I expect that a problem in the coming month will be optimizations to the code and smart choices in the learning models we choose so that our results are not only accurate, but relatively fast/efficient to achieve. As far as resources, my laptop is the only necessary tool.

D) Conclusion
I am on track with the project as shown by the progess made so far, I expect to be able to meet my project goals.