Project 2 **Report** **Shivam Kumar** (160010010) CS 386 September 7, 2018

1 Visualize the Iris dataset using scatterplots

After creating some scatter plots among different features and pairplot of the dataset, it became clear that petal length and petal width of Setosa are smaller from that of other two species. These two features are also correlated among themselves as can be seen from correlation matrix. As a result, in the scatter plot of petal length with petal width, we can see that Setosa is separated from other two species. Looking at other two species, there is an indistinguishable boundary between Versicolor and Virginica.

Observing correlation matrix, we can see a high correlation among sepal length, petal length, and petal width. But these three features are not significantly correlated with the sepal width.

If our goal has been to separate Setosa from the two other species, then petal length and petal width will be considered the ideal features.

2 Visualize the Iris dataset using boxplots and histograms

From histograms and box-plots, we can see that petal width and petal length of Setosa are separate from that of other species and there is also a separation between Viginica and Versicolor. Sepal length distributions of different species are also different but they are not separate from each other. Sepal width distributions of different species are almost overlapping.

We can see a trend in petal width, petal length and sepal length of flowers of different species, as we are moving from Viginica to Setosa, the length of these features is decreasing.

3 Visualize the Iris dataset using 3D-plots

We can see that all the three species are clustered. Setosa is clustered in a small space and its cluster is separate from the other two species. The clusters of Versicolor and Virginica are also separate but the boundary between both clusters is not clear.

I have chosen sepal length, petal width and petal length as our three axes because the clusters formed by taking these three features are more separate than taking any other set of three features.

4 Analyses of the k-means algorithm

When we ran k-means algorithm by taking all features present in our dataset with k = 3, we got three clusters similar to the clusters of three species in the original dataset. After running k-means algorithm several times, we saw that algorithm is taking 5-10 iterations to converge.

If there is no change in assignment of data points then it guarantees that there will be no change in future iterations also. So we can stop there. We have chosen this as our stopping criteria. Some other stopping criteria we could have chosen are following.

1. We will stop looping when J is not changing much, like if difference between J of i'th iteration and (i+1)'th iteration is less than 0.001
2. We will stop looping if centers are not changing.
3. We can also set a maximum number of iterations combined with above two methods.

5 Compare k-means and agglomerative clustering

We have observed the number of points assigned to each classes by different algorithms and compared them with true classes. We saw that these numbers are similar in both algorithms but they differ significantly with the actual number of data points in each classes. Though the number of points assigned to each classes are similar, k-means is giving comparatively better results than the another.

We made the following observations from the above 3D plots of result of both algorithms and true labels:

1. All points belonging to Setosa, are grouped in a single class by both algorithms.
2. Some points belonging to Versicolor, are assigned to the other class by both algorithms.
3. The results produced by both of the algorithms are very similar.

Effect of initial configuration for k-means - When we ran k-means algorithm by taking different initial centroids, we saw no difference between results produced by them. After limiting maximum number of iterations to 5, the 'random' initialization(init='random) was producing worse result compared to other initializations. Therefore, we concluded that if we choose centriods badly then, it will take more iterations to make better clusters.

Effect of initial configuration for agglomerative clustering - After running this algorithm with different linkages, we observed different assignments to some points of Versicolor and Virginica. After further exploration, we saw that the 'complete' linkage strategy is giving us significantly worse result compared to other two linkages.

One method to select which linkage to use is apply all three linkages and analyze which linkage is giving us best results.

6 Select K

Some methods are common to all clustering algorithms (for both k-means and agglomerative clustering):

1. Elbow method : Run k-mean algorithm for different values of k and then choose p as our k if error value is decreasing sharply till p'th iteration.
2. While choosing k, we make sure that k « m (where m = number of data points).
3. Visualizations of data points are used to get an idea about number of clusters.
4. Use other informations about the dataset if they are available. For example, suppose you have data of weights of people and you want to make one cluster for men and one for women. Here you should start experimenting with k = 2.

Special method for k-means - Run agglomerative clustering algorithm on a smaller random sample of orginal data : Agglomerative clustering algorithm is more expensive in terms of time than k-means algorithm. But it gives us an idea about the number of clusters present in dataset.

Special method for agglomerative clustering - Dendrogram formed from the dataset is used to get an idea about choosing a nice k.

Attachments

ai-project-2.ipynb, ai-project-2-code.py