|  |
| --- |
| Intro |
|  |  |
|  | The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis. |
|  |  |
|  | Multivariate is analysis on various different outcome variables - the end result |
|  | Linear discriminant analysis tried to find a common theme or difference between 2 or more variables |
|  |  |
|  | collected the data to quantify the morphologic variation of Iris flowers of three related species |
|  |  |
|  | The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other. |
|  |  |
|  | Use of Data Set |
|  |  |
|  | Based on Fisher's linear discriminant model, this data set became a typical test case for many statistical classification techniques in machine learning such as support vector machines[5]. |
|  |  |
|  | Classification is how we know where a new observation should be categorised. |
|  | Machine Learning is how a computer learns about new features and where and how they should be used. It invovles pattern recognition and computational learning theory |
|  | Support Vector Machines are the ability of machine learning to classify new material based on their commonality with other things it knows about, like how a person learns based on experiences and common words etc. |
|  |  |
|  | The use of this data set in cluster analysis however is not common, since the data set only contains two clusters with rather obvious separation. One of the clusters contains Iris setosa, while the other cluster contains both Iris virginica and Iris versicolor and is not separable without the species information Fisher used. This makes the data set a good example to explain the difference between supervised and unsupervised techniques in data mining: Fisher's linear discriminant model can only be obtained when the object species are known: class labels and clusters are not necessarily the same.[6] |
|  |  |
|  | Cluster analysis is the grouping or data into similar categories, important for machine learing. |
|  | Data Mining is how we delve into the data and make it usefull for future use. |
|  | The cluster analysis does not work as the 3 flowers from the same species appear as almost seperate species due to the difference between the setosa being so different to the 2 others virginica and versicolor. |
|  | Cluster analysis works by taking the means of each type and making a formula to categorize each input. |
|  |  |
|  |  |
|  |  |
|  | Therefore, the three species of Iris (Iris setosa, Iris virginica and Iris versicolor) are separable by the unsupervising procedures of nonlinear principal component analysis. |
|  |  |
|  | Clustering is unsupervised learning, you dont know about the variables so you have to create the variable into different groups |
|  | Classification is supervised learning - you know about the variables and are able to classify each new variable into a suitable group. |
|  |  |
|  | So the interesting thing about the data set is that if you used the supervised learning you would get the wrong result from your data set as each of the 3 species are so different. |
|  |  |
|  | Reference: |
|  | https://en.wikipedia.org/wiki/Iris\_flower\_data\_set |
|  | http://dataconomy.com/2015/01/whats-the-difference-between-supervised-and-unsupervised-learning/ |
|  |  |
|  | https://issuu.com/denhams/docs/statistical\_analysis\_report |
|  | Document describing using histrograms and other graphs the importance of the relative parameters in finding the difference and similarities betwen each data set |
|  |  |
|  | https://www.techopedia.com/definition/32880/iris-flower-data-set |
|  | Nice explanation |