Resources / Labs (/COMP9321/22T1/resources/72107) / Week 8 (/COMP9321/22T1/resources/72114) / Regression & Clustering

Regression & Clustering

Prerequisites:

It is assumed that you will install and take a look at the following packages in python before heading to activities:

- sklearn (http://scikit-learn.org/stable/)
- (http://flask.pocoo.org/) pandas (https://pandas.pydata.org/)

This lab makes use of the iris dataset (https://github.com/mysilver/COMP9321-Data-Services/blob/master/Week10_Regression_and_Clustering/diet.csv) . This dataset has four features including sepal_length, sepal_width, petal_length, and petal_length of three species of flowers: setosa, versicolor, and virginica.

Another dataset you will use in this lab is diet dataset (https://github.com/mysilver/COMP9321-Data-Services/blob/master/Week10_Regression_and_Clustering/diet.csv) . This data set contains information on 78 people using one of three diets. with the following columns:

Variable	**Description**	**Data type**
Person	Participant number	Numeric
gender	Gender, 1 = male, 0 = female	Binary
Age	Age (years)	Numeric
Height	Height (cm)	Numeric
preweight	Weight before the diet (kg)	Numeric
Diet	Diet (3 different kinds of diets named 1,2,3)	Numeric
weight6weeks	Weight after 6 weeks (kg)	Numeric

Activity-1:

Description : Create a model for weight prediction based on diet and person information

Steps :

1. Load the diet dataset
2. Split the dataset into test and train datasets; 70% of the data should be used for training the model and the rest for testing
3. Train a LinearRegression (http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html) regression model by fitting on the train dataset;
4. Based on the trained model, predict the weights of people in the test dataset;
5. Print the predictions and the real weights
6. Print the mean square error (http://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html) for your predictions

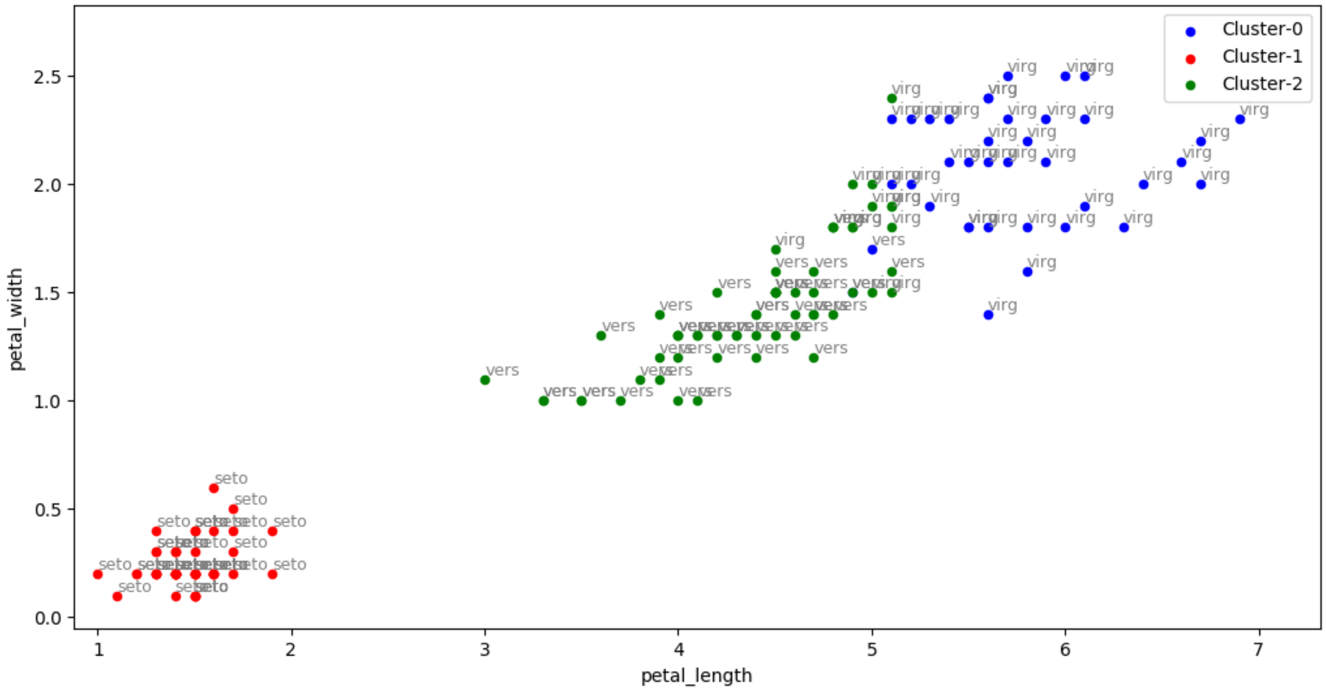
(https://github.com/mysilver/COMP9321-Data-Services/blob/master/Week10_Regression_and_Clustering/activity_1.py)

Activity-2:

Description : Using K-Means (https://en.wikipedia.org/wiki/K-means_clustering) split the iris dataset into 3 clusters

Steps :

1. Load the diet dataset
2. Drop the 'species' column; this is required because clustering is an unsupervised method.
3. Use K-means (http://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html) to cluster the data into 3 clusters; because we know that there are 3 different species of flowers in this dataset
4. Plot the clusters based on what you have learnt in Visualisation Lab. Plot a scatter chart using x=petal_length', y='petal_width' (https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.DataFrame.plot.scatter.html) for each cluster
5. Label each data point with the true label of flower class.



(https://github.com/mysilver/COMP9321-Data-Services/blob/master/Week10_Regression_and_Clustering/activity_2.py)

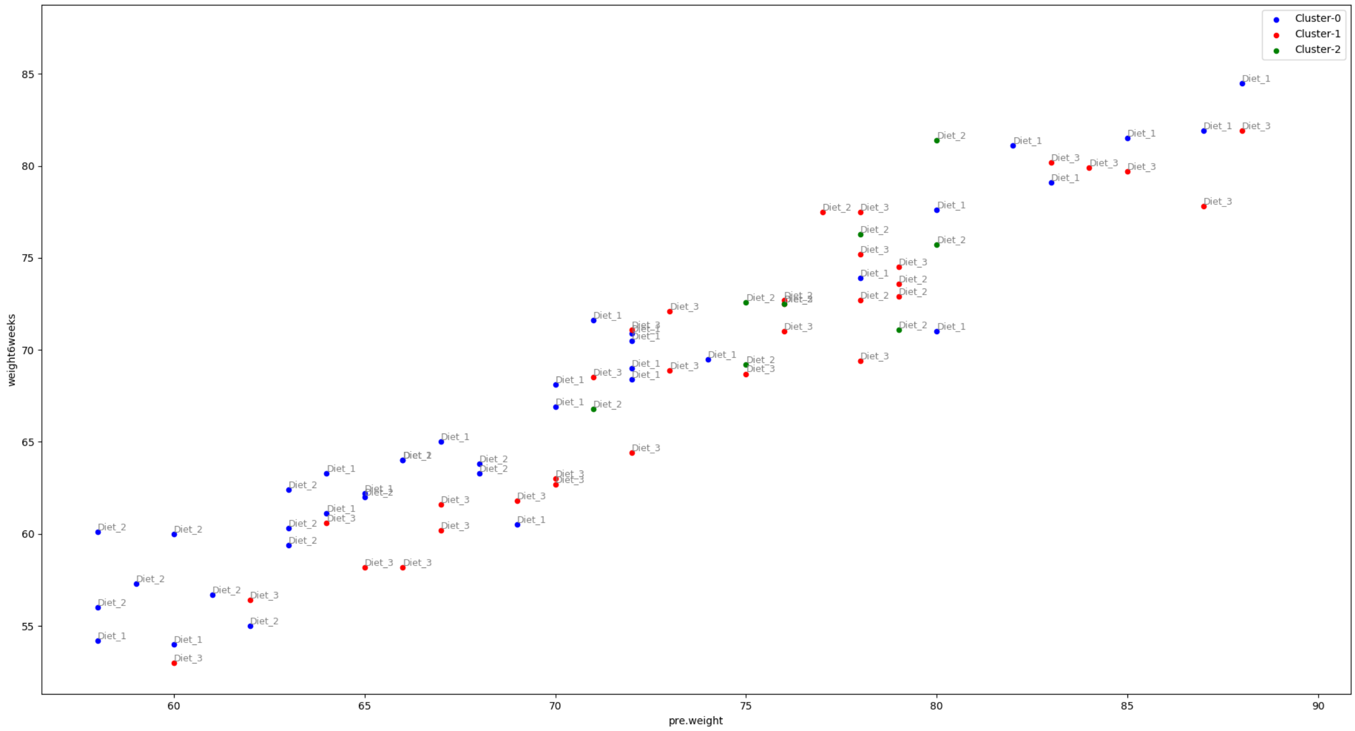
Activity-3:

Description : Using AgglomerativeClustering (http://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html) split the diet dataset into 3 clusters based on the diet types

Steps :

1. Load the diet dataset

2. Drop the 'Diet' column; this is required because clustering is an unsupervised method.
3. Use AgglomerativeClustering (http://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html) to cluster the data into 3 clusters; because we know that there are 3 different types of diet in this dataset
4. Plot the clusters based on what you have learnt in Visualisation Lab. Plot a scatter chart using x=pre.weight', y='weight6weeks' (https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.DataFrame.plot.scatter.html) for each cluster
5. Label each data point with the true label of diet.
6. Change the Clustering algorithm to KMeans; which one is better for this problem?



(https://github.com/mysilver/COMP9321-Data-Services/blob/master/Week10_Regression_and_Clustering/activity_3.py)

Resource created about a month ago (Monday 14 March 2022, 03:04:57 PM), last modified 15 days ago (Monday 11 April 2022, 04:09:23 PM).

Comments

(/COMP9321/22T1/forums/search?forum_choice=resource/74091)

(/COMP9321/22T1/forums/resource/74091)

Add a comment

Solomon Rachamim (/users/z5375417) 15 days ago (Tue Apr 12 2022 08:41:50 GMT+0800 (China Standard Time))

for activity 2:

> 1. Load the diet dataset

do you mean the iris dataset?

thanks

Reply