AI AppDeveloper FullStack – Applied Statistics, NLP Basic Concepts & Machine Learning Assignments

Applied Statistics & NLP spoc: anilkumar.ganesh@wipro.com TT-Architect Academy

(A) Applied Statistics:

(Make use of Pandas, Numpy)

Ex 1. For a given set of values in *stats.xls* that contains the list of employees, years of experience and their salary write a python script to calculate the mean, mode and median.

Ex2. For the above exercise determine the standard deviation and variance through python scripting.

(B) Natural Language Processing:

(Make use of Pandas, NLTK)

Ex 1. Write a python script that reads the *data_in.csv* from every cell in column labeled as comment and perform sentence tokenization and redirects in to column of *data_out.csv*. Perform the NE Chunking on these sentences.

Ex 2. Write a python script that reads the *data_in.csv* from every cell in column labeled as comment and perform word tokenization and redirects in to column of *data_out.csv*

Ex3. From an input file *data.txt* it is required to identify the POS-Tagging and display it on tree structure.

Ex 4. For a given text file exclude the stop words and perform the Stemming & lemmatization and compare the results.

Ex 5. Create a small dictionary file with required set of words with weightage attached to it with positive and negative numbers. Create a python script that analyzes the given text file and classify it as negative or positive sentiment.

(C) Machine Learning:

Machine Learning spoc: snehalatha.nagamangala@wipro.com _TT-Architect Academy

Machine Learning Linear regression assignments

Write a python script

BEST FIT LINE - Calling SKlearn linear regression

Ex 1. Data:

1. Download the MPG data file from UCI Machine Learning repository https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/
2. Identify target variable and independent variable.
3. Prepare the data file

Univariate Regression

Ex 2. Import relevant python libraries and sklearn linear_model

Ex 3. Split the file into train [80%] and test [20%] data

Ex 4. Apply linear regression

Ex 5. Train the model using the training sets

Ex 6. Display the coefficients coef, intercept and residues

Ex 7.Predict using test data

Ex 8. Perform Accuracy check using the R Square

Ex 9. Display using scatter plot the data points and the best fit line

Multi-variate Regression

Repeat the above steps

BEST FIT LINE - Cost function using un-constrained method - Gradient descent

Ex 1. Use the downloaded data

Ex 2. Convert this data to array

Ex 3. Define the learning rate and no. of iterations as 0.0001 and 1000 respectively along with y-intercept and slope

Ex 4. Create the functions to get the BEST FIT line

1. Compute error for the line given the points
2. Step gradient function
3. Gradient descent

Ex 5. Display using scatter plot the data points and the best fit line

Ex 6. Display the Gradient and y-intercept value in the form y = mx+c

Ex 7. Find the BEST FIT line i.e., m and c of y=mx+c with least error using trial and error method i.e., modify learning rate or iterations or both

Machine Learning KNN assignments

Ex 1. Data:

1. Download the census data file from UCI Machine Learning repository http://archive.ics.uci.edu/ml/machine-learning-databases/haberman/
2. Identify target variable and independent variable.
3. Prepare the data file

Ex 2. Import relevant python libraries and sklearn KNN model

Ex 3. Split the file into train [80%] and test [20%] data

Ex 4. Apply KNN algorithm

Ex 5. Train the model using the training set

Ex 6. Predict using test data

Ex 8. Perform Accuracy check

****