CS3PD18 Python & Data Science Applications Coursework Assignment

Deadline 12pm December 14th 2018.

Tom Thorne t.thorne@reading.ac.uk

Important information

Please include your student ID number in the name of the file containing your work.

By submitting your work electronically or on paper you are confirming that you have read AND agree with the following statement

I certify that this is my own work and that the use of material from other sources has been properly and fully acknowledged in the text. I have read the University's definition of plagiarism (as defined in my course handbook). I understand that the consequences of committing plagiarism, if proven and in the absence of mitigating circumstances, may include failure of the Year or Part or removal from membership of the University.

If you are unable to complete your work on time because of any physical or mental health, legal, personal or family problem then please contact the School Senior Tutor as soon as possible and complete an Extenuating Circumstances Form (download and complete a form from the essentials website https://student.reading.ac.uk/essentials/_the-important-stuff/rules-and-regulations/extenuating-circumstances.aspx). Under normal circumstances your form should be submitted before the coursework deadline.

Assignment

Data

The data are available on Blackboard under the Assignment heading. The files on Blackboard contain comma separated values (CSV) with a header describing briefly each column. These are described in more detail below. You **must** use these versions of the data sets.

Bike journey data

This data set contains anonymised bike trip data from the Los Angeles Metro Bike Share, available here: https://bikeshare.metro.net/about/data/. **Use the metro.csv data file provided on Blackboard**. Each row corresponds to a single bike journey. The columns in the data are (taken from https://bikeshare.metro.net/about/data/):

trip_id Locally unique integer that identifies the trip

duration Length of trip in minutes

start_time The date/time when the trip began, presented in ISO 8601 format in local time

end_time The date/time when the trip ended, presented in ISO 8601 format in local time

start_station The station ID where the trip originated

start_lat The latitude of the station where the trip originated

start_lon The longitude of the station where the trip originated

end_station The station ID where the trip terminated

end_lat The latitude of the station where the trip terminated

end_lon The longitude of the station where the trip terminated

bike_id Locally unique integer that identifies the bike

plan_duration The number of days that the plan the passholder is using entitles them to ride

trip_route_category "Round Trip" for trips starting and ending at the same station or "One Way" for all other trips

passholder_type The name of the passholder's plan

Seed shape data

This data set contains measurements of seeds from a number of different plant species. Each row corresponds to a single seed. The columns are:

area The area of the seed.

perimeter The length of the perimeter of the seed.

compactness A measure of the area of the seed relative to the perimeter.

length The length of the seed.

width The width of the seed.

asymmetry A measure of the asymmetry of the seed.

groove length The length of the groove in the seed.

Data sourced from UCI machine learning repository, originally by *M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik, S. Zak* in 'A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images', in: Information Technologies in Biomedicine, Ewa Pietka, Jacek Kawa (eds.), Springer-Verlag, Berlin-Heidelberg, 2010, pp. 15-24.

Tasks

For the assignment you need to produce a Jupyter notebook analysing the data provided. **You should break your code down into blocks and document your notebook using Markdown sections to explain what you are doing and why**.

Using the provided data, complete the two tasks below, using any of pandas, matplotlib, seaborn, numpy, scipy, scikit-learn and networkx.

Task 1 – Bike journey data

- Load the `metro.csv` file into a pandas data frame.
- Find a sensible way to remove the missing values from the data frame, and explain why you have chosen this method.
- Explore the distribution of the duration variable. You should produce a plot visualising the distribution, and calculate and discuss briefly statistics of the variable.
- Produce a plot showing how the *distribution* of duration relates to passholder type.
- Perform an appropriate statistical test to check if the mean duration is different between One Day Pass and Flex Pass passholders. What assumptions have you made by using this test?
- Convert the start_time and end_time columns to date objects if they are not already.
- Create a new column in the data frame that gives the hour of the day that each journey started on.
- Explore how the duration variable varies between each journey starting hour of the day, creating a plot to visualise this.
- Explore how the distribution of the duration variable varies between each day of the week, creating a plot to visualise this.
- Calculate the total numbers of passholders of each type travelling on each week day. Discuss the results.

Consider the data as a network of stations, with edges having weights corresponding to the total number of journeys made between them (at any time).

- Produce a visualisation of the network and discuss the output.
- Calculate statistics of the network, plot them where relevant, and discuss the results.

Task 2 – Seed shape data

- Load the `seeds.csv` file into a pandas data frame.
- Use scikitlearn to fit a Gaussian mixture model to the data, with 2 components. Describe Gaussian mixture models and interpret the results.
- Generate a scatter plot of compactness against groove length, showing the resulting cluster membership.
- Use an appropriate criterion (or several) to compare Gaussian mixture model clusterings with between 1 and 10 components. Discuss the results and suggest the number of clusters you would choose and why.

Building the notebook

Use the Anaconda Python distribution to write a Jupyter notebook. Anaconda is available to download for free, for Windows, Linux and Mac here - https://www.anaconda.com/download/.

You must create a Python 3.6 or Python 3.7 notebook, and **only use the packages included in Anaconda 5, Python 3.6 or Python 3.7 versions** in your notebook. If you have a good reason to use a Python package not included in Anaconda, please contact the lecturer (t.thorne@reading.ac.uk) first to check before using it.

Submission

Please include your student ID number in the name of the file containing your work.

Your notebook should be submitted on Blackboard Learn, under the Assignments section, as one archive containing a .ipynb notebook file, as well as a .html HTML version of the notebook. These can both be saved from the Jupyter interface under File -> Download as.

The deadline for submission is **12pm on December 14th 2018**.

Hints

- Use the pd.to_datetime function on a column to convert it into a datetime object.

- datetime objects have an hour attribute giving the hour of the day, and dayofweek attribute giving the day of the week.
- pandas has a function crosstab that will count the number of occurences of each pair in two Series.