Assessment 3

Practical Introduction to Data Science 2022/23

This assessment is a practical exercise. You should undertake the activities described below and write these up in a report. The report should explain **what** you did, (briefly) **how** you did it (possibly by referring to supplementary code), **why** you made the decisions you made and should also include a **discussion of the results** that you obtained. Can you draw a conclusion? Are the results what you expected?

The report should normally be a **maximum** of 15 pages long. Note that it is possible to obtain a good mark with significantly fewer pages. If your report is long, consider using appendices, or referring to supplementary material.

Since you only have a few weeks to work on this problem and since several of the ideas covered here could have been new to you on this course we are going to work on some fairly straightforward data and try to answer some general questions without worrying *too* much about statistical rigour. In particular, the datasets are not particularly big, which has the advantage that they are manageable to work with, but it has the disadvantage that in some cases, the machine learning algorithms will not have so much data to work with, and so the results that you obtain may not be particularly statistically significant.

You are going to work with two publicly available data sets:

1. Historic weather/climate data from the UK, and
2. Self-reported happiness statistics from the UK

This assessment is fairly open-ended: I encourage you to explore these datasets and see what interesting things you can find. Your primary analysis should use Python or R along with related libraries and packages. You may also use a command line shell such as Bash to help with certain stages. Use of other applications and tools is permitted if justified, but, for example, if you choose to preprocess your data in Excel, it won't be possible to obtain as good a mark as if you used Python, R or Bash (and libraries) to do this.

There is no one "right" answer. In writing up your work, you should try to justify *why* you did things the way you did as well as just describing the steps that you took and the results that you obtained. The process that you go through is as important as the final result, so you should make sure that you clearly describe the steps that you have taken in your report. It is the nature of this kind of work that you will have to make some assumptions and approximations. These should be stated and, where possible, justified.

The assessment is split into four parts. Parts 1, 2 and 3 make up 80% of the mark and are compulsory. Part 4 is split into two options (a) and (b). Please attempt **either** Part 4a **or** Part 4b. The percentages in brackets in the headings below indicate the weight associated with each part.

After each of Parts 1 to 4 have been described below, there are some general hints, which might help you undertake some of the parts of the assessment.

The Data

This is real life data. You will have to do a fair amount of tidying and preprocessing of this data before it can be fed into a learning algorithm. This is an important part of the process and you will get credit for doing this part of the procedure.

Dataset 1 can be obtained from https://www.metoffice.gov.uk/research/climate/maps-and-data/historic-station-data.

If you want to automate the download process, you might find the data in stations.txt (available from Learn) helpful. This file contains a list of strings which correspond to each of the stations, and they can be used in URLs to go directly to the data.
For example, for the station "nairn", you can find the data at:
http://www.metoffice.gov.uk/pub/data/weather/uk/climate/stationdata/nairndata.txt.

Dataset 2 can be obtained from http://www.ons.gov.uk/peoplepopulationandcommunity/wellbeing/datasets/personalwellbeingestimatesgeographicalbreakdown. This contains four sub-datasets (life satisfaction, worthwhile, happiness and anxiety). For the purposes of Part 3 below, you should use the ***happiness*** data. If you choose to do part 4a, then you could look at the other sub-datasets too. This dataset includes data at the level of local government regions, but that is probably more detail than we need for this exercise. It is sufficient to use data at the level of "Region". The file regions.txt (available from Learn) lists the regions and gives a latitude and longitude which can be used to describe their locations. The regions are identified by the Area Code.

Part 1 (25%)

Look at the weather data for each of the weather stations.

Looking only at the weather data (that is, don't include the latitude/longitude of the weather station) can you use a clustering algorithm to see if these stations can be clustered into groups that have "similar" weather? Note that you will have multiple weather readings for each station. You could pick the most recent, an average, an extreme, or you could consider all of the points individually and look for clusters in the individual observations. You should try to justify your choice.

Part 2 (25%)

Now let's turn this into a classification problem.

You should exclude 5 of the weather stations from this set (You could do this by picking the 5 last stations alphabetically).

Can you predict whether they fall in the Northern Third of the UK, Central Third of the UK or Southern Third of the UK using only the weather data? You have latitude data for all the weather stations, so that can give you the classification for each of the weather stations in your training set. To determine the latitude of the lines dividing the UK into three, you should note that the most northerly point has latitude 60.9 and the most southerly point has latitude 49.9.

Part 3 (30%)

The Office for National Statistics[1] collects and publishes data on how happy people think they are. This is self-reported data, so it may well have some inherent biases, but it is good enough for us to draw some general conclusions. Happiness is measured on a scale of 0-10. Let's try to answer the question as to whether the weather affects how happy we are.

Try to join the weather station data with the happiness data. You will only be able to do this in a coarse-grained way, because the places where there are weather stations do not correspond directly to the census areas used in the happiness surveys. You could use bands of latitude to do this, or you could put a grid over the county and look at which grid cell a weather station or census region falls in. Don't worry too much about the fact that weather data and census data don't cover exactly the same time periods; for the purposes of this assessment we can assume that to a first approximation the general climate at a given weather stations is fairly constant over time and that happiness also does not change too much from year to year.

For the final part of this assessment you can try *either* Part 4a *or* Part 4b below.

Part 4a (20%)

Explore this data further in any way that you want.

You might want to:

1. Look for any temporal effects that we neglected earlier: Are people happier in years when the weather is better?
2. Perform clustering or classification on the happiness data on its own.
3. Join these data sets to any other publicly available data set and look for any patterns.
4. Look at an alternative geographical decomposition of the UK from Part 2. What happens if you split the country East to West? What happens if you choose your three regions to have an equal number of weather stations in each region?

[1] https://www.ons.gov.uk

Part 4b (20%)

Try to automate parts 1 to 3 as much as possible. Write scripts and/or programs that follow each of the steps you did to find your answers to parts 1 to 3. Try to parameterise the scripts so that the whole process could be easily re-run with a small difference, say, by dividing the UK into four equal regions instead of three.

Submit your programs/scripts in a zip file (or similar) and include in your report a brief description of how they can be run and what they do.

Some Hints

You almost certainly have more fine-grained data here than you need to answer some of the questions. You will probably want to reduce the data in some way before using it for clustering or classification. You could do this by taking averages, taking extreme values (like maxima) or by selecting a (hopefully) representative subset of the data.

You can pre-process your data in several ways. You can do this manually, but to speed things up, you will probably want to use Python, R, a database system, or the command line. Here are some potentially useful command line examples. You may prefer to use commands from Python or R instead.

Some Useful Command Line Functions

You can use the **curl** command to download data from the command line, e.g.:

```
curl https://www.metoffice.gov.uk//pub/data/weather/uk/climate/stationdata/aberporthdata.txt > aberporthdata.txt
```

You can loop over a number of files with a command like

```
for place in aberporth armagh ballypatrick;
do curl 'https://www.metoffice.gov.uk//pub/data/weather/uk/climate/stationdata/'$place'data.txt' > $place'data.txt';
done
```

You can remove every occurrence of the character '*' from a file with a command like:

```
sed 's/*//g' armaghdata.txt
```

You can use the grep command to match certain patterns. The following command matches every line where the first column contains only numerical digits and the second column (space separated) contains the value '6':

```
grep '^ *[0-9]\+ \+6' armaghdata.txt
```

You can remove the first 7 lines of a file with a command like

```
awk ' NR>7 ' oldfile.txt > newfile.txt
```