Deakin University

SIT742 Modern Data Science

Unit Assessment Handbook

Gang Li

Trimester 1, 2020

School of Information Technology
Deakin University, Australia

CONTENTS

CHAPTER
ZERO

ASSESSMENT OVERVIEW

Contents

0.1 General Information

2020 *SIT*742 assessment consists of 4 components as detailed below:

Table 0.1: *SIT*742 Assessment Plan

Components	Percentage	Submission Methods	Working Mode
Task 1	25%	Critical Analysis and Report	Individual
Task 2	40%	Project Work	Group
Task 3	05%	Online Quiz in *CloudDeakin*	Individual
Final Examination	30%	Exam	Individual

0.1.1 Where to Get Help?

Students are encouraged to discuss the unit related assessment in *CloudDeakin* unit [**home page**] → [**Discussions**] → [**Student Discussion**].

If you are trying to form a project group for Assessment Task 2, you can post messages at: [**home page**] → [**Discussions**] → [**Finding a project group?**].

Any enquiry about this unit can be posted at: [**home page**] → [**Discussions**] → [**Questions for the Unit Chair**].

0.1.2 Where to Submit?

All tasks should be submitted to *CloudDeakin* by following corresponding specific requirements. You can find the corresponding submission boxes for Task 1 and 2 by following the [**home page**]→ [**Assessment**] → [**Assignments**] link.

0.1.3 Important Dates

Please be aware of the following important dates:

Task 1 **Due Date** *CloudDeakin* Submission, by **23:59pm,** ~~11/04/2020~~ 18/04/2020 **(Week** ~~05~~ 06 **Sunday)**.

Task 2 **Group Sign-Up Due Date** Task 2 group sign-up on *CloudDeakin* by **23:59pm,** 18/04/2020 **(Week** 06 **Saturday)**.

Task 2 **Due Date** *CloudDeakin* Submission, by **23:59pm,** ~~23/05/2020~~ 30/05/2020 **(Week** ~~10~~ 11 **Sunday)**.

Task 3 **Due Date** *CloudDeakin* Online Quiz, in Week 10.

0.1.4 Assignment Results

Task 1 and 2 will be marked based on your submitted *pdf* report and *Jupyter* notebook, while task 3 will be automatically marked online.

- The marking report is expected to be released to CloudDeakin within 14 working days of the due date;
- Within 3 working days after the result is released, any student who wishes to challenge the mark must contact or approach the unit chair during contact hours and bring with them a copy of their assignment and their mark breakdown. Cloud students can email me for this issue.

0.2 General Requirements

Student information and assessment related forms can be found from `http://www.deakin.edu.au/sebe/students/`. Students are required to familiarize with Faculty regulations regarding plagiarism.

1. Any text or code adapted from any source must be clearly labelled and referenced. You should clearly indicate the start and end of any such text/code.
2. **All SIT742 assignments must be submitted as required by their corresponding assessment specifications**. Assignments will not be accepted through any other manner without prior approval. Students should note that this means that email based submissions will ordinarily be rejected.

3. Penalties for late submissions are indicated in the `Unit Guide`. Close of submissions on the due date and each day thereafter for penalties will occur at $11:59$pm local time. Students outside of Victoria, Australia, should note that the local time zone is $UTC+10$, and in Daylight Saving Dates, it will be $UTC+11$.

4. Information regarding assignment extensions is provided in the `Unit Guide & Information` in CloudDeakin. Students must not assume an extension will be granted. Late penalties still apply in the case of a failed application for extension. Thus until an extension is granted students should submit any work completed before the assignment is due. Note that extensions cannot be granted for system outages or encumbrances.

ASSESSMENT TASK

ONE

DATA EXPLORATION: DATA SCIENTISTS SURVEY

Contents

This task contributes 25% of your final *SIT*742 mark. It must be completed individually, and submitted to *CloudDeakin* by **23:59pm,** ~~11/04/2020~~ 18/04/2020 **(Week** ~~05~~ 06 **Sunday)**.

1.1 Background

In 2017, Kaggle (a data science community and competition platform) conducted a survey on a large range of users registered as the data scientist in their platform. The survey data are broadly covered the skill set of the data scientists, the demographic of the data scientists, the feedback of the platform and many other information.

1.2 Task Description

We provide one `Jupyter` notebook `2020SIT742Task1.ipynb` at GitHub-SIT742, together with three data files at the `data` subfolder:

`MCQResponses.csv` The `csv` file contains participants' answers to multiple choice questions. Each column contains the answers of one respondent to a specific question.

`ConversionRates.csv` Currency conversion rates to USD.

`JobPostings.csv` Data scientists job advertising in US with job descriptions, from JobPikr.

You are required to develop a data exploration report by completing the provided `Jupyter` notebook to finish some required analysis, with the exploration data analytics skills as well as visualization skills. Details requirements can be found in the provided notebook, and you need follow the notebook requirements to complete the coding and include the results into the report `SIT742T1Report.pdf`.

1.2.1 Data Exploration

For a data scientist, after obtaining the dataset, the first most crucial task is to obtain a good understanding of the data he or she is dealing with. This includes: examining the data attributes (or equivalently, data fields), seeing what they look like, what is the data type for each field, and from this information, determining suitable numerical/visual descriptions.

In this part of this assessment task, you need to complete the provided notebook coding parts and finish the required analysis in the attributes such as '`education`', '`salary`' and related *demographic* information (70%).

1.2.2 Text analysis

For the job advertisement data `JobPostings.csv`, you are required to write Python code to remove the stop-words, and to extract the high frequency words used in job advertisements.

After that, you can do one self-defined text analysis task to get insight into those advertisement information (30%).

1.3 What to Submit?

Please familiarise yourself with the *General Requirements* (see Section 0.2) on Assignments Submission. By the due date, you are required to submit the following files to the corresponding *Assignment* (Dropbox) in CloudDeakin:

`SIT742Task1.ipynb` Your `Jupyter` notebook solution source file for the data exploration of the data scientists related data. You can fill your name and Deakin ID information at the relevant place in the first markdown cell.

Please follow the `PEP 8` guidelines (Section 3.1) for source code style. Your commenting and adherence to code standards will be considered when marking.

`SIT742T1Report.pdf` This pdf report contains the required source code, the required answers to selected questions, as specified in the notebook file.

No Special Consideration will be granted for this assessment task. Students who have difficulty meeting the deadline because of illness, etc. must apply for an assignment extension no later than the noon on the day prior to the deadline.

ASSESSMENT TASK

TWO

DATA ANALYTICS: FIFA 2019

Contents

This task contributes 40% of your final *SIT*742 mark. It can be done in group of 3 members and submitted to *CloudDeakin* by **23:59pm,** ~~23/05/2020~~ 30/05/2020 **(Week** ~~10~~ 11 **Sunday)**.

2.1 Background

Recently, Kaggle (a data science community and competition platform) released one data set FIFA19, which consists of 18K+ FIFA 19 player with around 90 attributes extracted from FIFA database. Here, we redistribute this data set for this assessment task:

2020T2Data.csv The file contains detailed information about each FIFA 19 player.

More information about this dataset can be found at: `https://www.kaggle.com/karangadiya/fifa19`.

2.2 Task Description

We provide one `Jupyter` notebook `2020SIT742Task2.ipynb` at `https://github.com/tulip-lab/sit742/tree/master/Assessment/2020`, together with **2020T2Data.csv** at the `data` subfolder.

You are required to analyse this dataset using `Jupyter` notebook with **Spark** packages including **spark.sql** and **pyspark.ml**.

2.2.1 FIFA19 Data Analytics

To systematically investigate this dataset, your `Jupyter` notebook should complete the following 3 kinds of analysis (80%):

Part 1 - **Exploratory Data Analysis** data visualization and understanding.

Part 2 - **Clustering Analysis** Identify the inherent clusters among players, and for each cluster, identify its profile.

Part 3 - **Classification Analysis** Build classifiers to predict the '`position_group`' of the player. You are also required to evaluate the performance of at least 3 models using cross-validation.

2.2.2 Project Report

Based on your implementation as required in `Jupyter` notebook, you are required to write a report `SIT742T2Report.pdf` with $1000 - 1500$ words, which should include the following information:

(1) The required report '`Section 1`' to '`Section 3`' (results and analysis) as specified in the notebook.

(2) In the report's '`Section 4`', discuss any findings you can reveal from this data set, such as any rising star? any omni player? etc. (10%)

(3) In the report's '`Section 5`', reflect the project group activities, such as the task distribution and contributions from each group members, and what you have learnt during this project. (10%)

More information about report writing can be found at: `https://www.deakin.edu.au/students/studying/study-support/academic-skills/report-writing`.

2.3 What to Submit?

2.3.1 Important Dates

Please be aware of the following important dates:

Group Sign-Up The group needs to be finalized on *CloudDeakin* by **23:59pm,** 18/04/2020 **(Week** 06 **Saturday)**. If any issue or group correction is needed, please send SIT742 unit chair an email by **23:59pm,** 18/04/2020 **(Week** 06 **Saturday)**.

Final Submission The due date for this assessment task submission is on **23:59pm,** ~~23/05/2020~~ 30/05/2020 **(Week** ~~10~~ 11 **Sunday)**.

2.3.2 Files to Submit

Please familiarise yourself with the *General Requirements* (see Section 0.2) on Assignments Submission. By the due date, you are required to submit the following files to the corresponding *Assignment* (Dropbox) in CloudDeakin:

`SIT742Task1.ipynb` Your `Jupyter` notebook solution source file for the data exploration of the bank marketing data. You can fill your group information at the relevant place in the first markdown cell. Please follow the `PEP 8` guidelines (Section 3.1) for source code style.

`SIT742T2Report.pdf` A $1000-1500$ words report describing and discussing your analysis results, and reflect the project group activities.

No Special Consideration will be granted for this project. Students who have difficulty meeting the deadline because of illness, etc. must apply for an assignment extension no later than the noon on the day prior to the deadline.

APPENDIX
THREE

APPENDIX

Contents

3.1 Code Style: `Pep 8`

`Pep 8` is the de-facto code style guide for *Python* (`https://www.python.org/dev/peps/pep-0008/`). Skim the style guide to gain basic understanding of what is required. Conforming your *Python* code to `PEP 8` is generally a good idea and helps make the code more consistent when working on projects with other developers.

In your assessment task, if the source code or `IPython` notebook is to be included, you are required to format your code so that it meets at least the following major `PEP 8` guidelines:

Comment Please follow the following style for *Python* comments:

1. To explain the functionality of a group of statements, apply block comments before the statements. Indent the comments to the same level as the code.
2. Write documentation strings (i.e. `docstring`) for your function.

Code Lay-out Please follow the following style for *Python* code layout:

1. Blank lines: Surround top-level function and class definition with two blank lines. Use blank lines in functions, sparingly, to indicate logical sections.
2. Indentation: Use four white spaces instead of tab for indentation.

White spaces in expressions and statements Please follow the following style for *Python* while spaces:

1. Surround binary operators with a single space on either side.
2. If operators with different priorities are used, consider add whitespace around the operators with the lowest priority(ies). However, never use more than one space.

You should use:

```
i = i + 1
num += 1
x = x*2 - 1
```

rather than this:

```
i=i+1
num +=1
x = x * 2 - 1
```

String quotes Use either single-quoted or double-quoted strings. Pick one of them and stick to it for consistency. Only use the other one when a string contains single or double quote characters.

Naming Conventions Make sure the naming of your variable follow consistent style: e.g. `lowercase`, `lower_case_with_underscores`, or `mixedCase`.

3.2 Academic Skills

3.2.1 How to Find Papers?

For the assessment task in this unit, you can try to find some related references from highly respected journals and conferences. You can find papers from Scopus, IEEEXplore, ACM Portal, Elsevier ScienceDirect, and DBLP [1].

Search engines like Google, Scopus and CiteSeer are widely being used to find papers, though they do not have a warrant for paper quality. Beside above search method, there are some other tricks which can help you to find the most suitable paper:

People/Group Oriented You could first identify important people/groups in that sub-domain by MS academic, and then find more related papers from their website.

Citation Oriented You could use search engine, like Google Scholar, to find papers with most citations in that sub-domain. In this way, you can find many classical and influential (but maybe not up-to-date) papers in this sub-domain.

[1] DBLP only provides paper titles and sources, you need to download the paper from somewhere else.

Top Conference/Top Journal Oriented You could find a lot of up-to-date papers on your topic from the top conferences or top journals. This would help you understand the state of art of your selected topic. To see which conferences or journals are top ones, you may refer to MS academic, or Google Scholar.

3.2.2 How to Read Papers?

There are some *Research skills* articles available on *How to Read*:

- Michael J. Hanson and Dylan J. McNamee. *Efficient Reading of Papers in Science and Technology*

In general it is unnecessary to understand all the details of all papers you collected, let alone that it is difficult to understand all. When you are reading a paper, you'd better keep in mind to answer following problems:

1. What is the problem concerned and why this problem is important?
2. How is the problem solved, completely solved or partially solved?
3. Does it have any relationship with other papers you have read?

3.2.3 How to Write a Paper?

There are some *Research skills* articles available on *How to Write*:

- Mike Ashby. *How to Write a Paper*

A first-time academic author usually lists everything he collected in a survey. A better way is to find a clue from motivations of different works: What's the goal and what's the problem? What's the first step, what problems it solved and what problems remained? So came the second step…Following your clue to answer these kinds of questions.

More Tips

1. When you are writing your report, please assume that your readers know nothing about your topic.
2. When you are writing your report, please keep your mind clear. It is better to first write down a outline.
3. After finishing your report, please check out your language and logic.

No Plagiarism You should be cautious about the writing. The *TurnItIn* system will be used for all SIT742 assignment submissions. Whenever you are using words and works of others, citations should be made clear such that one can tell which part is actually yours. IEEE Document "Introduction to the Guidelines for Handling Plagiarism Complaints" provides details about how IEEE will identify and handle a plagiarism.