Lab 2 - Project Setup and Teamwork contract

(Please sit with your team-mates for the labs!)

Recap: Project Course

Week 1	Week 3
Proposal, EDA & project set-up	Creating an automated pipeline

Remember to write tests!!!

Week 2	Week 4
Finalizing 4 scripts & starting report	Docker & project reproducibility

Thursday TAs

Javier Castillo-Arnemann
Teaching Assistant
Projects 301-307

Kate Sedivy-Haley
Teaching Assistant
Projects 308-315

Wednesday TAs

Gary Zhu
Teaching Assistant
Projects 407-413

Ozum
Teaching Assistant
Projects 401-406, &
414,415

5 minutes

Feedback on Milestone 1

- Level of detail for analysis, key decisions
- Proof-read your proposals and all text!
- Analysis Plan of Action!
- Training & Test Datasets

Milestone 2: What is due this week?

1. Team Contract
2. Script: download and save data
3. Script: reads data, data-cleaning & pre-processing
4. Script: EDA
5. Script: Do the statistical and/or ML analysis
6. Script: Create report

Reminders for Milestone 2

5 minutes

- Use both R and python (at least 1 script for each)

- Update your proposal based on TA feedback!

- Consider about project organization

- Add usage of scripts

README.md

Breast Cancer Predictor

- author: Tiffany Timbers
- contributors: Melissa Lee

Demo of a data analysis project for DSCI 522 (Data Science workflows); a course in the Master of Data Science program at the University of British Columbia.

Introduction

For this project we are trying to answer the question: given tumour image measurements is a newly discovered tumour benign or malignant? Answering this question is important because traditional, non-data-driven methods for tumour diagnosis are quite subjective and can depend on the diagnosing physicians skill as well as experience (Street, Wolberg, and Mangasarian 1993). Furthermore, benign tumours are not normally dangerous; the cells stay in the same place and the tumour stops growing before it gets very large. By contrast, in malignant tumours, the cells invade the surrounding tissue and spread into nearby organs where they can cause serious damage. Thus, it is important to quickly and accurately diagnose the tumour type to guide patient treatment.

The data set used in this project is of digitized breast cancer image features created by Dr. William H. Wolberg, W. Nick Street, and Olvi L. Mangasarian at the University of Wisconsin, Madison (Street, Wolberg, and Mangasarian 1993). It was sourced from the UCI Machine Learning Repository (Dua and Graff 2017) and can be found here, specifically this file. Each row in the data set represents summary statistics from measurements of an image of a tumour sample, including the diagnosis (benign or malignant) and several other measurements (e.g., nucleus texture, perimeter, area, etc.). Diagnosis for each image was conducted by physicians.

Source: https://github.com/ttimbers/breast_cancer_predictor/tree/v1.1

Group Interviews

Questions:

- What stage of the project are you on?
- Each member describes their most important contribution this week
- What is one thing that has worked really well for your team?
- Do you have any blockers/obstacles?
- How are you feeling about the project overall?