SM1/SC1 Project Suggestions

In all cases you should discuss your plans with Song Liu and Anthony Lee/Feng Yu to make sure you are covering the requirements for SM1 and SC1.

Each suggested project also has a contact for more information. They can give some project-specific advice on how to get started.

Tennis match prediction

Data

- WTA
- ATP

Basic model

- Bradley–Terry Model

To achieve better predictions, you will have to use additional explanatory variables / covariates.

You can use a frequentist or a Bayesian approach.

There are numerous articles on this model and various extensions.

If you want to use another sport, you may have to extend the model to accommodate draws (I don't recommend that you do this).

Contact for more information: Anthony Lee

Chicago Crime Dataset

Data

- Chicago City Data Portal
- Description

This is a dataset contains all the crimes (time/location/type/arrested) happened in Chicago since 2001. This is a typical example of dataset contains temporal (time) and spatial (space) information. There are several tasks you can do on this:

1. Given time/location/type of crime, can you predict if the suspect is arrested by police?
2. Given a specific location (a circle defined by coordinate X,Y with radius 5km), can you predict how many crimes (of specific types) happens during a specific time window (say Oct-2010)?

Contact for more information: Song Liu

Large Hadron Collider Data

Data

- CERN Data Portal

This is a simulated experimental data at Large Hadron Collider, CERN. The simulator generated observed events of "collisions" and your task is to predict whether such an event is a background or a signal (Higgs Boson) event. All features of events have physical meanings which may (or may not) help with your prediction. A description of those features without using too much physics knowledge can be found here:

- Description

This dataset is also used for a Kaggle challenge.

Contact for more information: Song Liu

Finance Data

Data

- Yahoo Finance

Yahoo Finance is one possible source of historical financial time series data.

We will mostly focus on the stock price history of large US companies, in the S&P 500 index. The current constituents of the S&P 500 index can be found here, among many other sources. It may be cumbsersome to manually download stock prices of all 500 companies, so we may have to find better data source.

One possible project is porfolio optimisation, i.e. to select the best asset distribution subject to certain constraints so that the expected return is maximized while financial risk is minimized. It is important to first construct the covariance matrix for the rates of return on the assets in the portfolio. The optimisation problem itself may be constrained by many factors, e.g. whether short selling is allowed. Transaction cost is another consideration that will impact on the frequency of adjustment to the portfolio.

Contact for more information: Feng Yu