INFO411: Data Mining and Knowledge Discovery

Project 8

Instructions:

This task is a real-world data mining problem. You are required to prepare a set of presentation slides which must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) your proposed data mining approach and methodology; (3) the strengths and weaknesses of your proposed approach; (4) the performance measures that can evaluate your data mining results; (5) the results and a brief discussion. Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (propose approaches, and discuss their strengths and weaknesses)
- Results (Figures and tables of data analysis)
- Discussion (discovered knowledge from data mining)

Task: Marathon Times

Background:

The length of a marathon race is 42.195 km. Participants usually include both genders and a wide range of age groups. Both age and gender tend to affect completion times, but the age effect is not necessarily monotonic (people generally tend to improve before their times start to deteriorate with increasing age).

For a large event, event organisers usually try to stagger the starting times to avoid too much jostling and overtaking. Elite athletes cross the start line first, so that their "net" time is very similar to the "gun" or "official" time', where

net time (= time of crossing finish line − time of crossing start line
gun time = time of crossing finish line − time starting gun is fired

Through a process of self-seeding or different start groups based on previous race results, slower runners tend to have a bigger gap between net time and gun time. Therefore the length of this gap could potentially be used as an inverse measure of ability. Data for the Melbourne Marathon (one of the largest events held in Australia) are available from

`http://melbournemarathon.com.au/General/Previous-Results`

However, as the download interface is complex, results for years 2013–2015 have been provided for you.

Requirements:

1. You are required to build two different models to predict net time using gender, age group, and the difference between official (gun) time and net time for years 2013–2015. You can choose which year or years to use for this project, keeping in mind that performances might vary between years due to factors like weather.

2. Compare the performance of the two models, and discuss the relative advantages and disadvantages.

3. Present some visualisations of the raw data and the fitted values.

4. Present details of your preferred model, including a discussion of the relative importance of the effects of age and gender.

5. Discuss whether the difference between official (gun) time and net time appears to be a useful proxy measurement of ability.