accenture

consulting | technology | outsourcing

Accenture Management Consulting

UCM VII Modelling Week
Propagation models

High performance. Delivered.

Problem description

Introduction

- **Social media** play a central role within the customer relationship management (CRM). More precisely. Threats and opportunities arise due to the ability of the users to share content quickly on social networks:
  - **Threat**: The emergence of **negative buzz** about a brand might turn into non-negligible reputational problems if spreading of information is significant.
  - **Opportunity: Viral communication**, ensuring fast and massive propagation, allows high performance marketing actions at very low cost
- Predictive **models of content propagation** within social networks enable **risk anticipation and opportunities identification** that might have a significant impact in a company's activity.

Problem description

- General goal: to **develop models of content propagation within social networks**. The general problem is rather complex and must have into account multiple factors. We will restrict ourselves to **Twitter** (wwww.twitter.com) which allows public access to data.

- The proposed problem is to **identify whether a given content (hashtag) will be used by a user,** as a function of how this hashtag or others have been used before. The following data will be available:

  – A **target table**. Containing user ID, hashtag ID, and a binary target variable: 1 if the user uses the hashtag, 0 if not. This table will split in train and test datasets.

  – A **directed graph** of (a part of) the twitter network, including follower and followees relationships.

  – A **table of users**. including all users in the graph and basic information about the user.

  – A **table of hashtags**. Including details of the hashtag: message ID, user_id, datetime of the message, number of tweets, followers and friends of the user at the moment in which the message was sent.

Guidelines to solve the problem

- The problem may be tackled by different means. A possible schema is described below, though it is highly encouraged to explore alternative methodologies:
  - Build a **modelling table** to predict the use of hashtags by each user **using a binary classification approach**. We can use the table of binary events, enhancing it with aggregated information on how the user has used the hashtag before and how propagation has taken place between the friends of the user using information from the graph, users, and hashtags tables.
  - Use the **train portion** of the previous table to build a binary classification model. Assess the goodness of the model with the **test data**.
  - The best model will be chosen in terms of the area under ROC curve on the test data.

- **SAS code is provided** to build an initial version of the table with basic features described in the next slide. Building the training table is very time consuming and the available time is short). The code may be edited to include improved variables that might enhance the predictive power of the model.

Mathematical approach - data transformation

- Predicting the use of a hashtag H by a user U depends on several factors:
  - **U User profile**:
    - How many users follow U?
    - How many users are followed by U?
    - How does U use twitter? (number of messages, number of previous hashtags, etc.)
  - **H Hashtag profile**:
    - How many times has the hashtag been used so far?
    - Has it been used previously by influent users?
  - **U neighbourhood**
    - Has the hashtag been used by his followers / followees?
    - How much do his followees use the hashtag depending on the type of followee?
  - Other factors to be defined. (one of the objectives of the problems is **finding out these additional factors** and integrate them in the predictive model).

Mathematical approach - modelling

- The desired model predicts a **binary target** variable:

$$y = \begin{cases} 1 & \text{if the user } i \text{ uses the hashtag } j \\ 0 & \text{if the user } i \text{ doesn't use hashtag } j \end{cases}$$

- To model the event starting with the proposed table, the models should allow capturing **non-linear effects as well as interactions** between the explicative factors defined before.
- We propose to use **logistic regression models with interactions**, **neural networks** or **decision trees**.
- Other alternative approaches will be discussed with the students.

Data model