

Representation Learning

ACML Assignment 4

Lavinia Pulcinella (i6233926) Enrique Barrueco (i6242336)

November 2020

1 Exercise 1

The problem described in Exercise 1 can be represented as a **Relational Representation**. That is, IMDB-like information involves all kinds of different information about a movie. For the problem at hand, to predict whether how much a movie will make on the opening weekend there are all kinds of information which can help the learning process.

- Table 1: **Production** containing information (i.e. attributes) on the production of the movie.
Attributes: MovieTitle, DirectorName, Genre, Release, Number of Awards, etc..
- Table 2: **ActorAwards** contains information on the actors starring in the movie and how many awards they won.
Attributes: ActorName, Number of Awards (AA)
- Table 3: **DirectorAwards** contains awards information on directors.
Attributes: DirectorName, Number of Awards (DA)
- Table 4: **ActorInfo** containing actor filmography.
Attributes: ActorName, MovieTitle
- Table 5: **DirectorInfo** containing director filmography.
Attributes: MovieTitle, DirectorName

Assume a movie is labelled as positive (**pos**) if it has boxed great on the first weekend. For instance, a simple hypothesis can be that a movie is profitable if directed by a specific director or a specific actor stars in it, namely

- `pos(Movie) :- DirectorInfo(MovieTitle, "Director Name", _)`
or
- `pos(Movie) :- ActorInfo(MovieTitle, "Actor Name", _)`

Awards won (x) by either directors and actors are usually an indicator of the good outcome of a movie. So an Hypothesis could be of the form of:

- `pos(Movie) :- Production(MovieTitle, _ , "GenreType", ReleaseYear, _), ActorAwards(ActorName, Number of Awards (AA)), DirectorAwards(DirectorName, Number of Awards (DA)), sum(AA, DA) > x`

Movie is positive if an actor is starring in a Action movie (for instance) released after 2015 (as special effects became better).

- `pos(Movie) :- Production(MovieTitle, _, Action, ReleaseYear, Y, _), ActorInfo(MovieTitle, "actor name"), Y > 2015`

This Relational Representation can be "down-played" to Multi-Instance by having only one table containing all the information. However, if for example a director or an actor win another awards or they make a new movie, in a relational representation setting each table gets updated only once. Otherwise, in a multi-instance representation all entries for the corresponding actor or director need to be changed.

2 Exercise 2

The problem presented in **Exercise 3** can be represented as an **Attribute Value** problem. That is it could be represented in a single table that could look like the following:

Title	Known Sender	Email Domain	Money Mention	Time Sent	Class
Nutricional Info	No	yindex	Yes	02:40	Spam
money fo u	Yes	notscamming	Yes	15:43	Spam
re:Dinner	Yes	gmail	No	22:40	Ham
Free money	No	hotmail	Yes	07:03	Spam
Student Discount	No	maastrichtuniversity	Yes	12:23	Ham
Course Registration	no	maastrichtuniversity	No	12:14	Ham
Inheritance Awaiting	no	dreri23	Yes	16:23	Spam

These would be some of the possible hypothesis that could be formulated:

- Spam :- SpamOrHam(., No, ., Yes, 03:45, .)

If the sender is not know, there is mentions of money in the email and it was sent in the middle of the night (indicating remoteness from the sender) then classify it as Spam

- Ham :- SpamOrHam(., ., maastrichtuniversity, ., .)

In words: If the email comes from a domain of Maastricht University always regard it as ham no matter the title, whether the sender is known or not, no matter if money is mentioned or the time received.

- Spam :- SpamOrHam('Inheritance', No, ., ., ., .)

If the subject of the email includes a mention of an Inheritance waiting for the receiver from an unknown sender where money is mentioned then the mail should be classified as spam.

- Ham :- SpamOrHam(., Yes, ., ., Yes)

If the email comes from a familiar email address talking about money then don't classify it as spam.

The problem could also be represented in a lower level down the hierarchy, that is by using a Boolean representation where each column of a huge table would represent whether a number of words are mentioned for example, the size of this table gets out of hand when more than 2 words are considered, as the combination of possible words make the possible combinations unfeasible.