Do (wo)men talk too much in films?
Mini Project in Machine Learning

January 27, 20211
Department of Information Technology, Uppsala University

Abstract

This document contains the instructions for the mini project on classification for the course Statistical Machine Learning, 1RT700. The problem is to classify the gender of the two main actors (one male, one female) in Hollywood movies. The training set consists of 1037 films and you will later be given a test set of 387 films. You are expected to (i) try some (or all) classification methods from the course and evaluate their performance on the problem, and (ii) make a decision which one to use and 'put in production' against a test set. Your final prediction will be evaluated and also compared to the performances of the other student groups. You will also write a report about the features that accurately predict who talks most in films. You will document your project by writing a report, which will be reviewed anonymously by your peers. A very well implemented and documented project will earn you a 'gold star' and a higher grade on the report.

0 Requirements

The project is to be done in groups of 3-4 students. All tasks described in this document have to be done in order to pass the project, and of course *all group members have to take part in the project*.

1 Problem and Data: actor classification

The technical problem is to tell which of two actors is male and which is female based on various properties of a film. Although we are predicting the gender of two actors this is a binary classification problem. If actor 1 is male then actor 2 is female and visa-versa. For context of this study, please first read this article: https://pudding.cool/2017/03/film-dialogue/. In this article, the authors were looking at the amount of speaking in films by male and female actors in order to detect gender bias. It turns out then, in children's movies in particular, it is male characters who do most of the talking (figure 1).

You are looking at this data from another perspective, measuring whether male or female lead role is predictable from the amount of dialogue the actors have, the year the film was made, how much money it made and so on.

The training data set `training.csv` consists of an output variable

Lead Is either 'Female' or 'Male'.

The lead is assumed to be the person who speaks most in the film (says the most words). The co-lead is assumed to have the gender male (if lead is female) and female (if lead is male).

The following input variables are provided

Year That the film was released.

Number of female actors With major speaking roles.

Number of male actors With major speaking roles.

Gross Profits made by film.

Total words Total number of words spoken in the film.

Number of words male Number of words spoken by all other male actors in the film (excluding lead and co-lead)

Number of words female Number of words spoken by all other female actors in the film (excluding lead and co-lead)

Number of words lead Number of words spoken by lead.

Difference in words lead and co-lead Difference in number of words by lead and the actor of opposite gender who speaks most.

Lead Age Age of lead actor.

Co-lead Age Age of co-lead actor.

Mean Age Male Mean age of all male characters.

Mean Age Female Mean age of all female characters.

You are expected to use all the knowledge that you have acquired in the course about classification algorithms, to come up with *one* algorithm that you think is suited for this problem and which you decide to put 'in production'. This algorithm will then be tested against a test set made available after peer review.

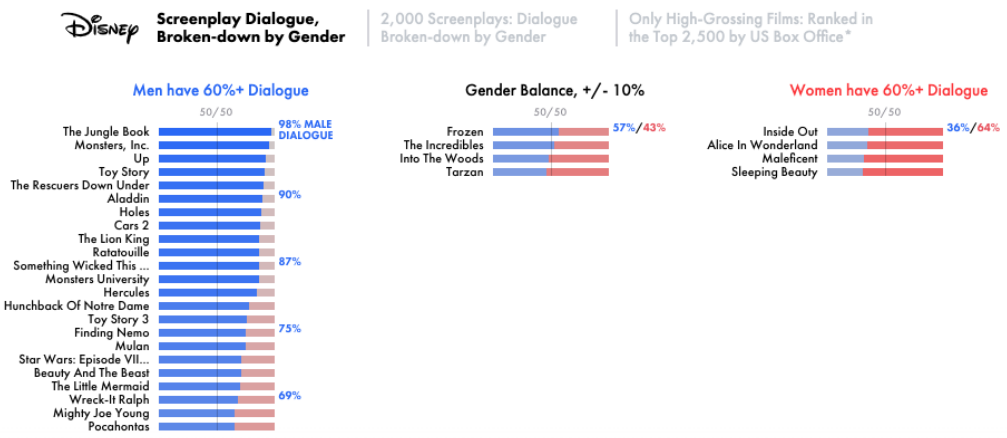

Figure 1: Gender bias in speaking roles in Hollywood films

2 Training

2.1 Methods to explore

The course has (so far[1]) covered the five following 'families' of classification methods:

(i) logistic regression

(ii) discriminant analysis: LDA, QDA

(iii) K-nearest neighbor

(iv) Tree-based methods: classification trees, random forests, bagging

(v) Boosting

In this project, you decide upon *at least* as many 'families' as you are group members, and decide in each 'family' *at least* one method to explore. To be clear, **each group member should independently implement and write about one method.** Who implemented which method should later be clearly written in the contribution statement. All group members should be able to stand for all sections of the report.

2.2 What to do with each method

For *each* method you decide to explore, you should do the following:

(a) Implement the method. We suggest that you use Python, and you may write your own code or use packages (the material from the problem solving sessions can be useful).

(b) Tune the method to perform well.

(c) Evaluate its performance using, e.g., cross validation.
Exactly how to carry out this evaluation is up to you to decide.

Once you have completed the aforementioned tasks, you should with a good motivation (hint: cross validation) select which method you decide to use 'in production' on a test set that will be made available later. **Work on this part of the project together and write the results together**

[1] Deep learning, which will be covered later, is also possible to use for classification. You are of course welcome to explore this as well, in addition to the minimum requirements.

3 Feature importance

Some input variables make better predictions than others. In this task investigate how important the following aspects:

- Words spoken by males and females
- Year of release
- Money made by film

are in predicting the gender. To do this you should try fitting your model including and omitting these variables. Try fitting models including a variable or excluding it and also look at models that include just one variable. Which features are most important in getting a good prediction? Here you can use misclassification error, false positives, false negatives and ROC/AUC to see how well models with or without certain variable perform. For logistic regression, you can also use Akaike Information Criteria (https://machinelearningmastery.com/probabilistic-model-selection-measures/) to test your fit. Do this work together as a group.

Then answer the following questions *based on your analysis*:

- Do men or women dominate speaking roles in Hollywood movies?
- Has gender balance in speaking roles changed over time (i.e. years)?
- Do films in which men do more speaking make a lot more money than films in which women speak more?

Write one paragraph in answer to each question.

Finally, discuss your results about gender and film together in an open and free discussion. No point of view is considered unreasonable in this discussion and you are free to say what you think, in the context of the data. After your group discussion write a joint two paragraph reflection on what you have collectively learnt from this analysis. Again, the conclusions should be based on the data analysis done here, but there is no single correct answer to this.

4 Documentation

You should summarize your work by writing a report, which will be first peer-reviewed by your coursemates (on first hand-in date) before being graded by us.

4.1 What to include in your report

The report should include the following:

(1) A brief introduction to the problem

(2) A concise description of each of the considered methods, and how they are applied to the problem. Even if you used high-level commands, such as `glm()` for logistic regression, you should explain what is 'under the hood' of the command! You can use (and reference) online sources, but use your own words. *Copy and paste from online sources will be identified in our plagiarism check and will lead to automatic failure and (in serious cases) reporting to the University's disciplinary board.*

(3) How the methods were applied to the data (which inputs were used, if the inputs were considered as qualitative or quantitative, how parameters were tuned, etc), including motivations of the choices made.

(4) Your evaluation of how well each method performs on the problem.

(5) Which method you decided to use 'in production', and your (good) arguments for your choice!

(6) Your conclusions.

(7) Your answers to the feature importance task.

(8) Appropriate references.

(9) All code needed to reproduce your reported findings (in an appendix).

In reward of the gold star we will concentrate on the following:

1. Thorough use of hyperparameter tuning and cross-validation.
2. A well thought out 'feature importance' task.
3. Performance on test data (see below).

4.2 How to format your report

Your report should be submitted as a PDF-file following the style used for the prestigious machine learning conference Neural Information Processing Systems (NeurIPS), which also is the style used for this document. In the NeurIPS format, your report should be *no longer than 7 pages* (not counting the reference list and code appendix). Except for the page limitation, you should follow the NeurIPS style closely, including its instructions for figures, tables, citations, etc.

The report should be written in LaTeX. You can access the LaTeX files from the conference webpage https://neurips.cc/Conferences/2020/PaperInformation/StyleFiles. If you prefer not to install a LaTeX compiler on your computer, you can use online services such as Overleaf (https://www.overleaf.com/). In your .tex-file, add the lines
`\makeatletter`
`\renewcommand{\@noticestring}{}`
`\makeatother`
before `begin{document}` to suppress the conference-specific footnote.

When you submit your report for peer review (1st submission), you should *not* include your own names or group name in the report or its filename (since it will be reviewed anonymously by your colleagues)! This is the default setting in the LaTeX template. When you submit the revised report (2nd submission) you should, however, include your names, along with a contribution statement. In LaTeX this is achieved by the `final` option, i.e., use `\usepackage[final]{neurips_2020}`.

The LaTeX template has line numbers in its draft mode. You should not remove these numbers. They can be useful for your reviewers when they want to refer to a specific part of your report (e.g., "the equation on line 54").

5 First submission: peer review

Please follow the instructions on Studium for submission. The first submission is for *anonyomous* peer-review.

Your report will be reviewed by students from other groups. Each student will also receive the report of another group, which you have to review. This means that the peer review is done individually and each group will receive multiple reviews. As a peer reviewer, you are expected to comment on the following aspects of the report:

(I) The subset of methods chosen to explore is sufficiently large (methods from at least as many 'families' as there is group members LIKE).

(II) All tasks (a)-(c) from Section 2.2 are made for each method.

(III) Make an assessment of the technical quality of the proposed solution. Have the considered methods been used in a relevant way to address the problem at hand? Are there any flaws in the reasoning and/or motivations used?

(IV) The report includes everything required from Section 4.1.

(V) The feature importance task is discussed seriously

(VI) The quality of the language in the report is satisfactory.

(VII) The report follows the format requirements (correct template, page limitation, etc.).

The review process is "double blind", meaning that both the project report and the review are anonymous. The review is done by filling out scores in the rubric of the mini-project on Studium and by adding text comments in that rubric. Please follow the instructions on Studium for how to fill in and submit your review.

Of course, you should use a polite and constructive language in your review. (Tip: *think about how you would assess your own report before you submit it!*)

After the review deadline, each group will get the reviews on their report from other students.

6 Second submission: graded

After peer-review *all groups* should resubmit an updated report including:

- Revisions accounting for peer review.
- A clear indication of the method you 'put in to production', i.e. based your submitted predictions on.
- Your names on the report.
- A contribution statement (as a separate document, see below)
- A prediction for the test set (as a separate file, see below)

There is only one week between the peer reviews and the second submission.

6.1 Contribution statements

As single page should clearly state the contributions of each group member, clarifying who contributed to which part, etc. In particular, which method each took responsibility for and who did the work and wrote the other sections. These should be uploaded in a separate section in order to preserve anonymity.

6.2 Model prediction

Upload a `.csv` file with the name `predictions.csv` with the format, e.g.

```
1,0,1,1,1, ... , 0
```

where 1 indicates that your model predicts actor 1 is **female**. This should be a single **row** of comma separated zeros and ones, with no other text. We will evaluate all submissions and publish a top table of best performers (with group name only). The lecturers will present the table at one of the lectures and talk about the best performers.

6.3 Grading

The **second submission** of the report will be graded with one out of four possible grades:

- Fail, if the deadline is missed or the report is far from meeting the criteria. No revision is possible until next time the course is given.
- Revise, if there are only minor issues. A revised version should be handed in before the revision deadline.
- Pass, if the report fulfills *all* criteria.
- Pass with gold star, if the report fulfills all criteria, and *in addition*
  - is written such that a thorough understanding of the methods is conveyed *and*
  - has a technical contribution beyond the minimum requirements. *and/or*
  - performs very well on the test data.

  This will earn you a higher grade for the report and possibly a higher course grade. See the course web page on Studium for details.

7 Third submission: graded

Save us all a lot of work and don't get to this point! However, if you got revise in the second submission, the **third submission** of the report will be graded with one out of two possible grades:

- Fail, if the deadline is missed or the revised report still does not meet the criteria. No more revision is possible until next time the course is given.
- Pass, if the report fulfills all criteria.

Please note that sub-standard reports will not be given the chance to be revised, and gold stars are handed out only at the second submission.

8 Deadlines

See the course web page on Studium.

Good luck!