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

January 19, 2023
Version 3.2

Department of Information Technology
Uppsala University

Abstract

This document contains the instructions for the 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.

1 Workflow

Before you start with the project, you should **read the whole instructions carefully**. Below we outline the general workflow of the project. For details of each part, see the specific sections.

1. Work on your report:
   - You work on your project in groups of 3-4 students.
   - Check out the problem formulation and the provided data set described in Section 2.
   - Answer the questions of the data analysis task in Section 3.
   - Implement and tune the methods as described in all subsection of Section 4.
   - When writing your report, follow the documentation guidelines in Section 5.
2. First submission:
   - As a group you submit your report.
   - After the report submission each student gets assigned to peer review one report from another group. Read the report and make useful comments on how to improve it. See Section 6.
3. Second submission:
   - Every group has received multiple feedback from the peer review and has to revise their report. Re-upload the revised version of your report. See Section 7.
   - Upload a contribution statement, see Section 7.2.
   - Upload your model predictions, see Section 7.3.
   - After the deadline your submissions will be graded by the teachers.

4. Third submission (Section 8):
    - If your report grade after the second submission by the teachers is a revise, then you have to update your report according to the feedback and re-upload it.
    - If you failed the peer review (no submission or too little/un-useful comments), you get assigned a second additional peer review and eventually additional tasks to compensate the missed deadline.

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*.

There are deadlines on each part of this workflow. For the exact dates, see the course web page on Studium.

2 Problem and data: Lead actor classification

2.1 Problem formulation

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.

You are looking at a subset of this data from another perspective: measuring whether male or female lead role is predictable from a number of features.

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

2.2 Data set

The training data set `training.csv` consists of an output variable **Lead**, see the first row in Table 1. 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). This output variable has to be predicted from a number of features, see the remaining rows in Table 1.

3 Data analysis task

The first step—before we start to build a model—is always to take a closer look at the data by analyzing some statistics of the data set. In this step you can already gain some insights into the data which helps you interpret the results of your methods later in the project.

Look into the provided data set by e.g. plotting the individual features. Based on your analysis, answer the following questions:

- 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 concise answers to each question and support your findings with evidence (statistics, plots, ...). Discuss the results. Additionally you can explore the correlation of features, outliers, range of values, and many more aspects.

Table 1: Available label (first row) and features (remaining rows) in the data set.

Feature Name	Description
Lead	Is either 'Female' or 'Male'.
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 if lead is male).
Number of words female	Number of words spoken by all other female actors in the film (excluding lead if lead is female).
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.

4 Implementation of Methods

4.1 Methods to explore

The course has (so far) 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 of the list method above.** 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.

Note, the model family of deep neural networks (DNNs) is not a part of the required model families above as it is covered later in the course. You are welcome to explore DNNs, but each member still has to fulfill the minimum requirement of implementing and describing one of the model families (i)-(v) listed above.

Additionally, you have to compare to a naive classifier. This can be a classifier that always predicts `male` or `female` or random labels as the lead actor in your analysis and compare its performance with your methods.

4.2 Before you start

As first step after the data analysis from section 3 you should agree on the following points in your group before you start to work individually on the different models:

- Select an evaluation protocol which everyone uses. For example to split in train and validation or use k-fold cross-validation.
- Choose one or multiple metrics which you want to evaluate your model. For example you can use accuracy, balanced accuracy, f1-score, recall, precision, and many more.

- Identify any pre-processing steps for the data.
- Discuss which methodological approach will be used for model tuning. For example grid search, random search, or similar.
- Additionally you can think about the input features. Should you use all of them or only a subset or a combination of different features.

4.3 What to do with each method

This task can be done individually.
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..

4.4 Model selection

Once you have developed the models (individually), you should identify *as a group* which model you consider the 'best' one for the task. You decide which model to use 'in production' on the test set. This test set will be made available after the first submission. Write up the results together.

5 Documentation

You summarize your work by writing a report, which will be first peer-reviewed by your fellow student in the course (first submission date) before being graded by the teachers (second submission date).

5.1 What to include in your report

The report has to include the following:

(1) An abstract to summarize the problem and your findings.

(2) A brief introduction to the problem.

(3) Your data analysis including answers to the questions and plots.

(4) The model development, including:

  (a) A concise mathematical description of each of the considered method, and how they are applied to the problem. You should *not* describe the code commands that you use but the underlying mathematical concept.

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

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

  (d) Model selection: Which method you decided to use 'in production', and your arguments for your choice!

(5) A conclusions.

(6) Appropriate references.

You can use (and reference) online sources, but use your own words. Note: Copy and paste from online sources or past reports will be identified in our plagiarism check and will lead to automatic failure and (in serious cases) reporting to the University's disciplinary board.

5.2 How to format your report

Important points:

- Use the template for the Neural Information Processing Systems (NeurIPS) conference with line numbers in its draft mode for the first submission and without line numbers for the second (final) submission.
- Include all points from the list in section 5.1.
- Obey the page limit of 7 pages, excluding appendix and references. Your report must be fully understandable within those 7 pages. This page limit includes the title and abstract. Do not include a table of contents. A violation of the page limit will not be accepted and the report has to be resubmitted.

Details:

- Your report has to be submitted as a PDF-file following the style used for the prestigious machine learning conference 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 has to be written in LaTeX. You can access the LaTeX files from the conference webpage https://neurips.cc/Conferences/2022/PaperInformation/StyleFiles. Make sure to copy the style file and and link it in your main file as done in the original template. 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.
- (A) For the first submission (including peer review), 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. However, add the number of group members at the end of the abstract (for peers to evaluate the number of methods).
  (B) For the second submission, you should include your names. In LaTeX this is done by the `final` option, i.e., use `\usepackage[final]{neurips_2022}`.
- The LaTeX template has line numbers in its draft mode. You should not remove these numbers for the first submission. 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").
- Make sure all plots are readable and support your statements. General rule: font size in figures should roughly be equal to the font size in the text.

6 First submission

6.1 Submission details

- The first submission is for *anonyomous* peer-review.
- Since the report is submitted anonymously, you need to specify the number of group members in your group manually. Do this by adding “Number of group members: K”, where K is the number of group members in your group, as the last sentence in the *abstract*.
- You have read the final version of the report from start to end; made sure it is readable; and all parts listed in section 5.1 are included.
- The report does not contain material copied from elsewhere (all reports are checked for plagiarism using Urkund).

6.2 Peer review:

The peer-review is a mandatory part of the examination. For the grading criteria of the peer-review please see the Grading section below.

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) Before the implementation, the data is properly analyzed and the questions from Section 3 are answered.

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

(III) All tasks (a)-(c) from Section 4.3 are made for each method.

(IV) 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?

(V) The report includes everything required from Section 5.1.

(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 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.

6.3 Grading

Report: Your report has to include all parts listed in section 5.1. If we see that your report does not meet these criteria, we have to fail your report and you will have to re-do the project in the next course iteration.
For all other groups, no official grading of the report is done at this stage. Every group adhering to the criteria has to revise with the peer review feedback.

For the peer-review: Every student has to pass the peer review individually. In order to pass, your peer review should cover the listed points with comments in the rubric. You are not limited by the number of comments you can give. A guideline is to write 1/4 to 1/2 page of comments in order to pass. A peer-review without or with too little comments, leads to a "fail" of this part.

Note: Even if you think that a point is well done, then comment on it and explain why you think so.

If you do not pass the peer review, you will be assigned a second peer review after the third submission. Meaning that you have to do twice the work. Failing this new peer-review leads to extra tasks, e.g. summarizing (multiple) machine learning research papers or finally to a "fail" grade for the complete project.

7 Second submission

7.1 Submission details

Report After peer-review *all groups* have to 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 so the second submission is not anonymous any more.
- Remove the line numbering from the template.

Code Furthermore, you upload to the same submission field a zip file containing all your (readable and commented) code such that it can be run without problems. Your results and figures should be reproducible with this code.

Other uploads In addition you have to upload:

- A contribution statement (as a separate document, see below).
- A prediction for the test set (as a separate file, see below).
- Write a few sentences about how you have updated your report based on the peer-review. These sentences are submitted in the comment field on Studium when you submit your updated project report.

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

7.2 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. This should be uploaded as a separate single document. One document should be uploaded per group.

If we see, that some students did not contribute to the project, we remove them from the group and they receive the grade fail. So make sure that you check the uploaded contribution statement in order to avoid conflicts between group members.

You need to include the contribution statement in order to pass. An empty contribution statement is not valid.

7.3 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 performance of each group on the test set is evaluated and published.

You need to include the model prediction in the correct format in order to pass. If you do not upload this, we cannot evaluate whether your model works.

7.4 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 an extra technical contribution beyond the minimum requirements (such as one more model family on top). *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.

8 Third submission: graded by teachers

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.

Good luck! ♣