Do (wo)men talk too much in films?

Anonymous Author(s)
Affiliation
Address
email

Abstract

1 Introduction

2 Methods

We have chosen to focus on approaches using logistic regression, k-NN and LDA/QDA to classify the lead actor's gender.

In order to make the methods as comparable as possible, we have used a common set of transformations of the input variables for all tested methods.

2.1 Input transformations

In the given dataset, there are columns for the total number of words spoken as well as the number of words spoken by the lead, the co-lead etc. This could present a problem since if we compare a movie where the lead says 10 out of 100 total words and another movie where the lead says 100 out of 1000 words, most models would think that the lead speaks more in the second movie and miss the fact that the *proportion* of words spoken by the lead is the same. For that reason we have transformed several input variables to express a proportion instead of absolute numbers. We also believe it might be important to have a dummy variable indicating if the lead or the co-lead is oldest. All transformations are given in Table 2.1.

Original column	New column	Transformation
Number of words lead	Proportion of words lead	$\frac{\text{Number of words lead}}{\text{Total words}}$
N/A	Proportion of words co-lead	$\frac{\text{Number of words lead - Difference in words lead and co-lead}}{\text{Total words}}$
Difference in words lead and co-lead	Ratio words co-lead lead	$\frac{\text{Proportion of words co-lead}}{\text{Proportion of words lead}}$
Number words female	Proportion of words female	$\frac{\text{Number words female}}{\text{Total words - Number of words lead}}$
Number of female actors	Proportion of female actors	$\frac{\text{Number of female actors}}{\text{Number of female actors + Number of male actos}}$
N/A	Older lead	$\begin{cases} 1, \text{Age lead} > \text{Age Co-Lead} \\ 0, \text{else} \end{cases}$

Table 1: Transformations of input variables.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Note that when determining 'Proportion of words female', this should only measure the words spoken by non-lead female actors so we have to subtract the lead's contribution to the total number of words.

The column 'Number of male actors' was dropped since all necessary information in this column is contained in 'Proportion of female actors'.

In order to improve regularization and k-NN, all numerical input variables (after transformation) where centered and scaled by their standard deviation. This results in a dataset where every numerical column contains almost all data in the interval $[-3, 3]$ (in the limit, $\approx 99.7\%$ of the data should be in this interval), with higher density closer to 0.

2.2 Logistic Regression

Logistic regression is a *general linear model* (GLM), i.e. the relationship between the data $X \in \mathcal{X} \subseteq \mathbb{R}^p$ and the outcome Y is on the form

$$E(Y|X) = g^{-1}(X \cdot \beta) \tag{1}$$

where $\beta \in \mathbb{R}^p$ and g is the link function. In the case of logistic regression, $Y|X \sim Ber(p)$ and the canonical link function is the logit link $g(x) = \log\left(\frac{x}{1-x}\right)$ with $g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$. Since $Y|X \sim Ber(p)$, we get $E(Y|X) = p = g^{-1}(X \cdot \beta)$. In other words, $P(Y = 1|X = x) = g^{-1}(x \cdot \beta)$, which we can use to predict Y given data x.

To do the regression, we find $\hat{\beta} \in \arg\min_\beta \sum_{i=1}^n (y_i - \hat{y}(x_i; \beta))^2$ where $\hat{y}(x; \beta) = g^{-1}(x \cdot \beta)$. This minimizes the mean squared error (MSE) loss function. A potential problem with this approach is that there are no restrictions on the components of β and that can lead to overfitting, especially if n is not much larger than p. To address that issue, one can introduce regularization.

In general, regularization is done by adding a penalizing term to the loss function that restricts β in some way. If $L(\beta; x_i, y_i)$ is the loss function before regularization, we instead consider the new loss function $L(\beta; x_i, y_i) + \lambda R(\beta)$ and find $\hat{\beta}_{reg} \in \arg\min_\beta \left(L(\beta; x_i, y_i) + \lambda R(\beta)\right)$. R is some penalizing function and λ is a hyper-parameter that can be tuned. The two most common forms of regularization is LASSO and Ridge regression.

LASSO regression uses L_1-regularization, meaning that $R_{LASSO}(\beta) = ||\beta||_1 = \sum_{i=1}^p |\beta_i|$ while Ridge regression uses L_2-regularization, $R_{Ridge}(\beta) = ||\beta||_2^2 = \sum_{i=1}^p \beta_i^2$.

When attempting to classify

2.3 k-Nearest Neighbors

2.4 LDA and QDA

3 Results

3.1 Logistic Regression

3.1.1 k-Nearest Neighbors

3.2 LDA and QDA

4 Conclusions

5 Feature Importance