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.

2.1 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.

2.2 k-Nearest Neighbors

2.3 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

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