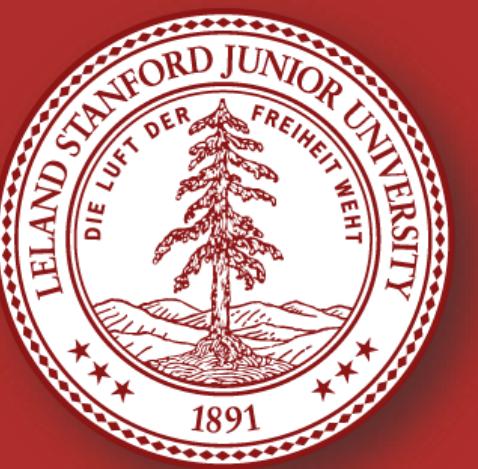


Behind the TV Shows: Top-Rated Series Characterization and Audience Rating Prediction

Yushu Chai – Yiwen Xu – Zihui Liu

CS 229 – Machine Learning

Professor Andrew Ng



Motivation

In this project, we are going to use the viewers' ratings on IMDb website as an indicator of audience's preferences of various TV series, try to discover features of well-received TV series and make recommendations.



Data

The TV series data with record on IMDb were provided by Andrej Krevl from Stanford SNAP Lab.

Name	Type
Genre	Vector of Booleans
Release Year	Numeric
# Seasons	Numeric
# Ratings	Numeric
# Critics	Numeric
# Ep./Season	Numeric
# Reviewers	Numeric
Run Time	Numeric
Aspect Ratio	Categorical
Color	Booleans

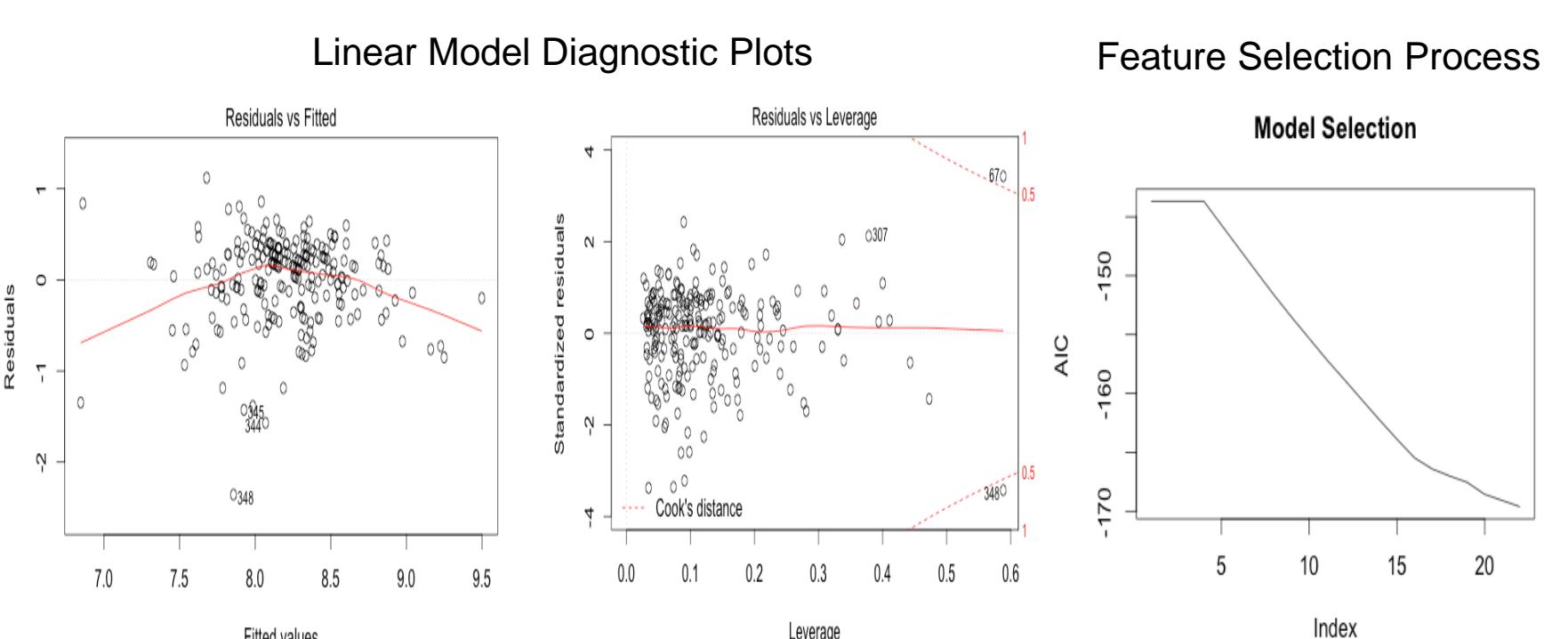
Models and Discussions

1. Basic Model: Multiple Linear Regression

$$\text{rating} = h(x) = \theta_0 + \theta_1 x_1 + \dots + \theta_n x_n = \sum_{i=1}^n \theta_i x_i,$$

The results demonstrate that the assumptions of linear regression on IID normally distributed error terms were not perfectly met. So we considered the following improvements.

Models and Discussions (Cont'd)



2. Improved Model I: Locally Weighted Linear Regression

$$\sum_{i=1}^m w_i (y_i - \sum_{j=1}^n \theta_j x_j)^2, \text{ where } w_i = \exp\left(-\frac{(x_i - x)^2}{2\tau^2}\right).$$

The bandwidth parameter τ is chosen to minimize the mean testing error.

3. Improved Model II: Linear Regression with Selected Features and Interaction Effects between Original Features

We used backward search starting with the full sets including all possible interaction features. For each iteration of the backward search, we did cross validation to evaluate the model. The selected model gave a lower error.

Comparisons Between Models

Linear regression models	Features included	Estimate of generalization error
Linear regression	All	0.3456
Locally weighted linear regression	All	0.3398
Linear regression with selected features and interaction effects between original features	Features selected by backward selection	0.3321

[1] Ng. Andrew, CS 229 Lecture notes.

[2] James, Gareth, et al. An introduction to statistical learning. New York: Springer, 2013. 373-413

[3] N.A. (2015, Oct. 15). The Internet Movie Database. Retrieved from <http://www.imdb.com/>

Acknowledgements & References

We want to thank Professor Ng for such an inspiring course and this project opportunity, and CS 229 TA's, in particular Youssef Ahres, for their help and support.

Classification

4. Classification

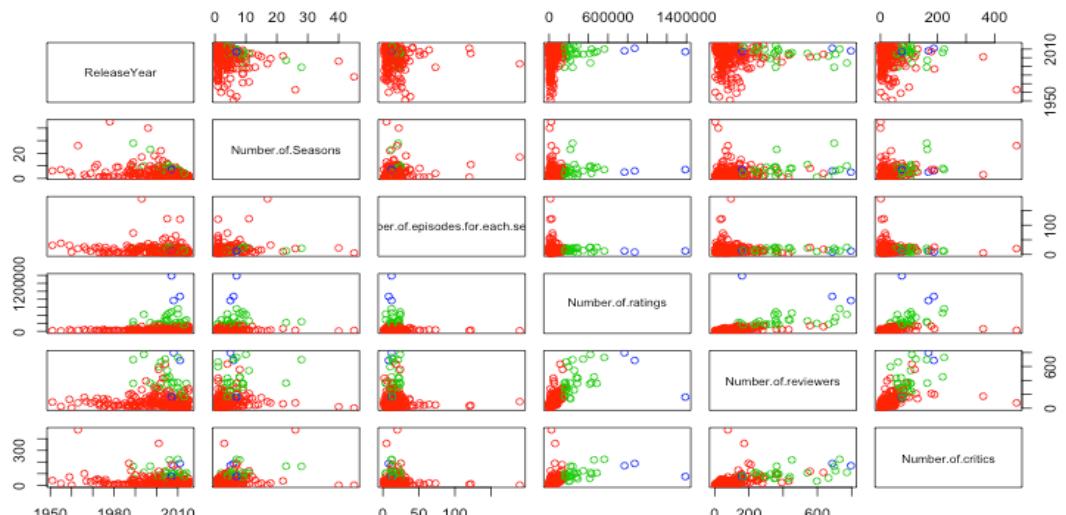
Based on the current dataset, we segment the ratings into several categories: Fair (rating below 8), Good (8-9), and Very Popular (9-10), to fit a classification model using L1-regularized logistic regression. When predicting the ratings of an upcoming TV show, this classification model can predict the most likely category for it.

Original Rating	Class	Counts	# in this group	Misclassification Rate
< 8	Fair	65	52	
8-9	Good	148	166	
9-10	Very Popular	12	5	0.08

Unsupervised Learning

5. Unsupervised Learning

- *K-means clustering*. $K = 3$ because we would like to find a more reasonable segment for the three classes above. But other values of K will also be considered.



• PCA

The plots are projections of the data onto the first three principle components, i.e. the scores for the first three principal components. The observations of different subgroups lie near each other in the low-dimensional space.

