Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

Exercise 7

Social Data Science

1 Regression for Prediction

Consider the following list of movies that star Robert Pattinson:

Title	Year	IMDB score	Age rating	Length (min)	Genre
The Lighthouse	2019	7.5	16	109	Drama
High Life	2018	5.8	16	113	Adventure
Damsel	2018	5.5	12	113	Adventure
Good Time	2017	7.4	12	101	Drama
Life	2015	6.1	0	111	Biography
Queen of the Desert	2015	5.7	0	128	Biography
Twilight: Breaking Dawn pt. 2	2012	5.5	12	115	Drama
Twilight: Breaking Dawn pt. 1	2011	4.9	12	117	Adventure
Remember Me	2010	7.1	12	113	Drama
Twilight: New Moon	2009	4.7	12	130	Adventure
Twilight	2008	5.2	12	122	Drama
Harry Potter and the Goblet of Fire	2005	7.7	12	157	Adventure

Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

a) Assume you want to fit the following regression model:

$$\text{IMDB score} \simeq \beta_0 + \beta_1 \cdot [\text{Year} \geq 2015] + \beta_2 \cdot [\text{age rating} < 16] + \beta_3 \cdot [\text{length (min)}]$$

What would be the feature vector representation for the first two movies?

b) After training the regressor you obtain $\beta = (4.5, 1.5, -1, 0.01)$. What would you predict would be the IMDB scores of the more recent films *The King* (released 2019, age rating 16 years, 140 minutes, Biography), *The Devil All the Time* (released 2020, age rating 16 years, 138 minutes, Drama), and *Tenet* (released 2020, age rating 12 years, 150 minutes, Drama)?

c) Compute the R^2 statistic of your predictions on the small test set of movies from b), assuming that *The King* and *The Devil All the Time* have an IMDb score of 7.2, and *Tenet* has an IMDb score of 7.8.

d) Name two other features that would benefit the regression, and how you would encode them!

e) Assume you change your mind on transforming the year into a binary feature, and want to regress on the value of the year itself. Assuming an ordinary least squares regression model with no regularizer, show that using the feature [year] is equivalent to using the feature [year-2016].

f) Briefly explain why the two feature representations from d) would not be equivalent when training a regression model with a regularizer!

Prof. Dr. Markus Strohmaier
Dr. Ivan Smirnov
Tobias Schumacher

2 Matching

Consider the following (fictional) dataset, based on which the effect of a job training program (T) on the yearly income in 1000 Euros (Y) is to be evaluated. Aside from these treatment and outcome variable, there are three columns of sociodemographic data, and an additional column of propensity scores that have been obtained from a logistic regressor on the covariates.

Age	Ethnicity	Bachelor's Degree	T	Y	Propensity
71	A	0	0	84	0.09
22	A	0	0	35	0.96
35	C	0	0	64	0.76
46	D	1	0	76	0.35
55	A	1	0	45	0.21
61	B	1	0	60	0.32
73	D	0	0	77	0.12
43	D	0	0	55	0.6
67	B	0	0	101	0.04
58	C	1	0	95	0.25
59	C	0	0	87	0.18
30	C	1	0	40	0.71
62	B	0	1	65	0.34
43	D	0	1	60	0.6
24	D	0	1	40	0.95
55	A	1	1	47	0.21
30	C	1	1	42	0.71
21	C	1	1	35	0.96

a) Compute the average treatment effect considering the full data, i.e. the differences in mean incomes between the people have have been treated, and those who have not been treated

b) Apply exact matching on the data and compute the average treatment effects on the matched data!

c) Apply greedy propensity score matching on the data, pairing treated individuals with control group individuals from top to bottom. What is the average treatment effect in this matching?