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

CSSH Computational Social Sciences and Humanities | RWTH AACHEN UNIVERSITY

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

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)?

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

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!

a) $x^{LH} = (1, 0, 109)$

$x^{HL} = (1, 0, 113)$

b) Model $\hat{y} = 4.5 + 1.5 \cdot [\text{year} \geq 2015] - [\text{age rating} < 16] + 0.01 \cdot [\text{length(min)}]$

The King: $\hat{y} = 4.5 + 1.5 \cdot 1 - 1 \cdot 0 + 0.01 \cdot 140 = 4.5 + 1.5 + 1.4 = 7.4$

Devil: $\hat{y} = 4.5 + 1.5 \cdot 1 - 1 \cdot 0 + 0.01 \cdot 138 = 7.38 \rightarrow y = 7.2$ (True score 7.2)

Tenet: $\hat{y} = 4.5 + 1.5 \cdot 1 - 1 \cdot 1 + 0.01 \cdot 150 = 6.5 \rightarrow y = 7.8$

c) $R^2 = 1 - \frac{RSS}{TSS}$, $RSS = \sum_i (\hat{y}_i - y_i)^2 = (7.4 - 7.2)^2 + (7.38 - 7.2)^2 + (6.5 - 7.8)^2$

$= 1.7624$

$TSS = \sum_i (y_i - \bar{y})^2 = (7.2 - 7.4)^2 + (7.2 - 7.4)^2 + (7.8 - 7.4)^2 = 0.24$

$\bar{y} = \frac{1}{3}(7.2 + 7.2 + 7.8) = 7.4$

$R^2 = 1 - \frac{1.7624}{0.24} = -6.343$

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

d) 1. Feature: [Genre = Drama]

2. Feature: ["Twilight" in Movie Title]

e) (*) $y = \beta_0 + \beta_1 \cdot [year] + \beta x$

(**) $y = \beta_0' + \beta_1' \cdot [year - 2016] + \beta' x$

$= \beta_0' + \beta_1' \cdot [year] - 2016 \cdot \beta_1' + \beta' x$

$= \underbrace{\beta_0' - 2016 \cdot \beta_1'}_{\text{const.}} + \beta_1' [year] + \beta' x$

if we set $\beta_0 = \beta_0' - 2016 \cdot \beta_1'$, and $\beta_1 = \beta_1'$, then the models are identical

f) In regularized regression, the weights β are not as free to adapt two translated/or differently scaled feature values. These would actually have big impact on the absolute values of $||(\beta_0, \beta_1, \beta)||$ which are used as penalty, and [year - 2016] would yield very different penalties than [year].

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

CSSH Computational Social Sciences and Humanities | RWTH AACHEN UNIVERSITY

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.

ID	Age	Ethnicity	Bachelor's Degree	T	Y	Propensity
1	71	A	0	0	84	0.09
2	22	A	0	0	35	0.96
3	35	C	0	0	64	0.76
4	46	D	1	0	76	0.35
5	55	A	1	0	45	0.21
6	61	B	1	0	60	0.32
7	73	D	0	0	77	0.12
8	43	D	0	0	55	0.6
9	67	B	0	0	101	0.04
10	58	C	1	0	95	0.25
11	59	C	0	0	87	0.18
12	30	C	1	0	40	0.71
13	62	B	0	1	65	0.34
14	43	D	0	1	60	0.6
15	24	D	0	1	40	0.95
16	55	A	1	1	47	0.21
17	30	C	1	1	42	0.71
18	21	C	1	1	35	0.96

a) Compute the ATE without matching, i.e., we want to compare all treated with all untreated individuals

$$ATE = \frac{1}{N_T}\sum_{T_i=1} Y_i - \frac{1}{N_C}\sum_{T_i=0} Y_i = \frac{1}{6}(65+60+40+47+42+35)$$
$$- \frac{1}{12}(84+35+\dots+40) = -\frac{291}{12} \simeq -20$$

b) Apply exact propensity score matching to compute the ATE

exact matches

8 - 14
5 - 16
12 - 17
18 - 2

$$ATE = \frac{1}{4}(60+97+90+35) - \frac{1}{4}(55+45+40+35) = 3$$

↳ Now positive

c) Apply greedy PS matching and compute the ATE

matches
13-4
14-8
15-2
16-5
17-12
18-3

$$ATE = \frac{1}{6}(65+60+40+97+92+35) - \frac{1}{6}(76+55+35+45+60+64)$$

$$= -\frac{27}{6}$$ → now negative again

↳ if we consider the "bad" match 18-3, we can see, that this had huge impact as the y-values were 35/64

↳ since the propensity scores were far away from each other, it might have made sense to use a calliper.