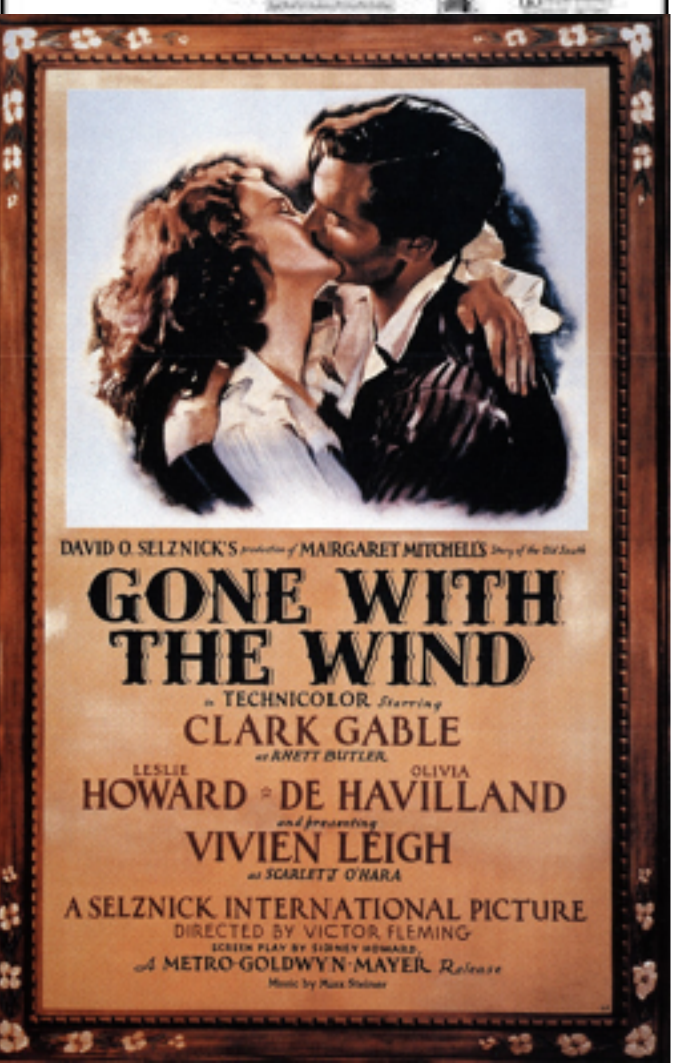
Applied Natural Language Processing

Info 256

Lecture 6: Text regression (Feb 7, 2019)

David Bamman, UC Berkeley



Regression

A mapping *h* from input data x (drawn from instance space $\mathcal{X}$) to a point y in $\mathbb{R}$

x	"Star Wars is not a great movie in the sense that it describes the human condition. It is simply a fun picture that will appeal to those who enjoy Buck Rogers-style adventures. What places it a sizable cut above the routine is its spectacular visual effects, the best since 2001: A Space Odyssey." (Siskel, 1977)
y	$1.32B

Regression problems

task	x	y
predicting box office revenue	movie reviews	opening box office
predicting real estate sales prices	real estate description	sales price
predicting stock movements	all tweets	price of $GOOG

$399000 Stunning skyline views like something from a postcard are yours with this large 2 bed, 2 bath loft in Dearborn Tower! Detailed hrdwd floors throughout the unit compliment an open kitchen and spacious living-room and dining-room /w walk-in closet, steam shower and marble entry. Parking available.

$13000 4 bedroom, 2 bath 2 story frame home. Property features a large kitchen, living-room and a full basement. This is a Fannie Mae Homepath property.

$65000 Great short sale opportunity... Brick 2 flat with 3 bdrm each unit. 4 or more cars parking. Easy to show.

Foster et al. (2013), “Featurizing Text: Converting Text into Predictors for Regression Analysis”

Regression

Supervised learning

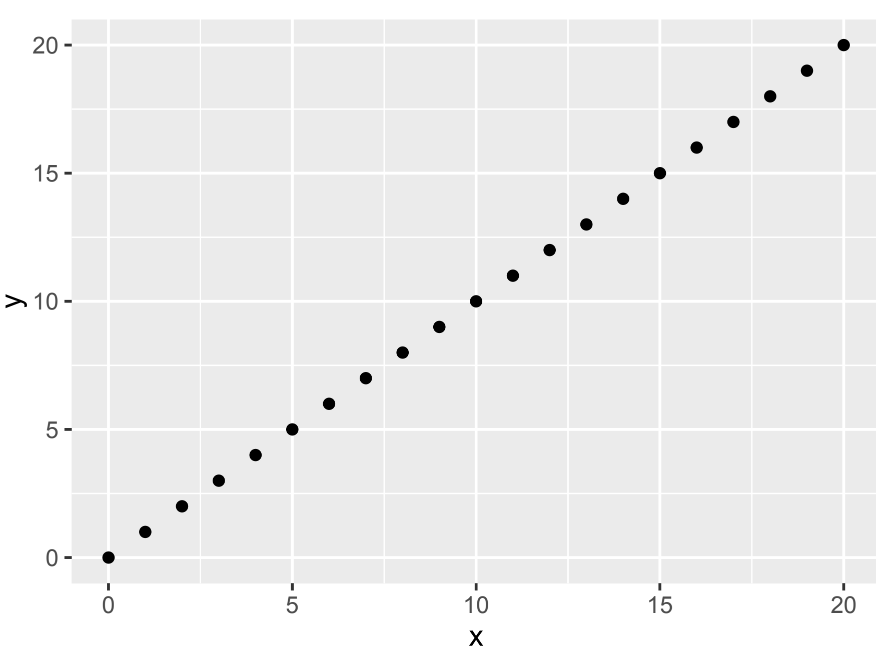
Given training data in the form of <x, y> pairs, learn $\hat{h}(x)$

Regression


Deep learning
Decision trees
Probabilistic graphical models
Random forests
Linear regression
Networks
Support vector machines (regression)
Neural networks

Linear regression

$$\hat{y} = \sum_{i=1}^{F} x_i \beta_i$$



$\beta \in \mathbb{R}^F$

(F-dimensional vector of real numbers)

x = feature vector

Feature	Value
the	0
and	0
action	1
love	1
animation	0
audiences	1
not	0
fruit	0
BIAS	1

β = coefficients

Feature	β
the	0.01
and	0.03
action	15.3
love	3.1
animation	13.2
audiences	3.4
not	-3.0
fruit	-0.8
BIAS	16.4

Linear regression

$$y = \sum_{i=1}^{F} x_i \beta_i + \varepsilon$$

true value y

$$\hat{y} = \sum_{i=1}^{F} x_i \beta_i$$

prediction ŷ

$$\varepsilon = y - \hat{y}$$

ε is the difference between the prediction and true value

Evaluation

Goodness of fit (to training data)

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

sum of square errors

total sum of squares

For most models, R^2 ranges from 0 (no fit) to 1 (perfect fit)

Experiment design

	training	development	testing
size	80%	10%	10%
purpose	training models	model selection	evaluation; never look at it until the very end

Metrics

- Measure difference between the prediction ŷ and the true y

Mean squared error (MSE)

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

Mean absolute error (MAE)

$$\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

81.7% of total MAE →

y	ŷ	MAE	MSE
1	2	1	1
1	1.1	0.1	0.01
1	100	99	9801
1	5	4	16
1	-5	6	36
1	10	9	81
1	3	2	4
1	0.9	0.1	0.01
1	1	0	0
		121.2	9939.02

← 98.6% of total MSE

MSE error penalizes outliers more than MAE

How do we get good values for β?

β = coefficients

Feature	β
follow clinton	-3.1
follow trump	6.8
“benghazi”	1.4
negative sentiment + “benghazi”	3.2
“illegal immigrants”	8.7
“republican” in profile	7.9
“democrat” in profile	-3.0
self-reported location = Berkeley	-1.7

Least squares

$$\beta = \min_{\beta} \sum_{i=1}^{N} \varepsilon^2$$

we want to minimize the errors we make

$$\beta = \min_{\beta} \sum_{i=1}^{N} (y - \hat{y})^2$$

$$\beta = \min_{\beta} \sum_{i=1}^{N} \left(y - \sum_{j=1}^{F} x_j \beta_j \right)^2$$

Least squares

$$\beta = \min_{\beta} \sum_{i=1}^{N} \left(y - \sum_{j=1}^{F} x_j \beta_j \right)^2$$

- We can solve this in two ways:
  - Closed form (normal equations)
  - Iteratively (gradient descent)

Many features that show up rarely may likely only appear (by chance) with one label

More generally, may appear so few times that the noise of randomness dominates

β = coefficients

Feature	β
follow clinton	-3.1
follow trump + follow NFL + follow bieber	7299302
“benghazi”	1.4
negative sentiment + “benghazi”	3.2
“illegal immigrants”	8.7
“republican” in profile	7.9
“democrat” in profile	-3.0
self-reported location = Berkeley	-1.7

Ridge regression

$$\beta = \min_{\beta} \underbrace{\sum_{i=1}^{N} (y - \hat{y})^2}_{\text{error}} \underbrace{+\eta \sum_{i=1}^{F} \beta_i^2}_{\text{coefficient size}}$$

We want both of these to be small!

This corresponds to a prior belief that β should be 0

Ridge regression

$$\beta = \min_{\beta} \underbrace{\sum_{i=1}^{N} (y - \hat{y})^2}_{\text{error}} \underbrace{+\eta \sum_{i=1}^{F} \beta_i^2}_{\text{coefficient size}}$$

A.K.A.

L2 regularization
Penalized least squares

low L2

Matt Gerald	$295,619,605
Peter Mensah	$294,475,429
Lewis Abernathy	$188,093,808
Sam Worthington	$186,193,754
CCH Pounder	$184,946,303
…	…
Steve Bacic	-$65,334,914
Jim Ward	-$66,096,435
Karley Scott Collins	-$66,612,154
Dee Bradley Baker	-$73,571,884
Animals	-$110,349,541

BIAS: $5,913,648

med L2

Computer Animation	$68,629,803
Hugo Weaving	$39,769,171
John Ratzenberger	$36,342,438
Tom Cruise	$36,137,757
Tom Hanks	$34,757,574
…	…
Western	-$13,223,795
World cinema	-$13,278,965
Crime Thriller	-$14,138,326
Anime	-$14,750,932
Indie	-$21,081,924

BIAS: $13,394,465

high L2

Adventure	$6,349,781
Action	$5,512,359
Fantasy	$5,079,546
Family Film	$4,024,701
Thriller	$3,479,196
…	…
Western	-$752,683
Black-and-white	-$1,389,215
World cinema	-$1,534,435
Drama	-$2,432,272
Indie	-$3,040,457

BIAS: $45,044,525

Interpretation

$$\hat{y} = x_0\beta_0 + x_1\beta_1$$

$$x_0\beta_0 + (x_1 + 1)\beta_1$$

Let's increase the value of x_1 by 1 (e.g., from 0 → 1)

$$x_0\beta_0 + x_1\beta_1 + \beta_1$$

$$= \hat{y} + \beta_1$$

β represents the degree to which y changes with a 1-unit increase in x

Regularization

- Regularization applies to linear models that are used for both regression and classification.

Feature selection

- We could threshold features by minimum count but that also throws away information
- We can take a probabilistic approach and encode a prior belief that all β should be 0 unless we have strong evidence otherwise

L2 regularization

$$\ell(\beta) = \underbrace{\sum_{i=1}^{N} \log P(y_i \mid x_i, \beta)}_{\text{we want this to be high}} \quad - \quad \underbrace{\eta \sum_{j=1}^{F} \beta_j^2}_{\text{but we want this to be small}}$$

- We can do this by changing the function we're trying to optimize by adding a penalty for having values of β that are high
- This is equivalent to saying that each β element is drawn from a Normal distribution centered on 0.
- η controls how much of a penalty to pay for coefficients that are far from 0 (optimize on development data)

no L2 regularization		some L2 regularization		high L2 regularization	
33.83	Won Bin	2.17	Eddie Murphy	0.41	Family Film
29.91	Alexander Beyer	1.98	Tom Cruise	0.41	Thriller
24.78	Bloopers	1.70	Tyler Perry	0.36	Fantasy
23.01	Daniel Brühl	1.70	Michael Douglas	0.32	Action
22.11	Ha Jeong-woo	1.66	Robert Redford	0.25	Buddy film
20.49	Supernatural	1.66	Julia Roberts	0.24	Adventure
18.91	Kristine DeBell	1.64	Dance	0.20	Comp Animation
18.61	Eddie Murphy	1.63	Schwarzenegger	0.19	Animation
18.33	Cher	1.63	Lee Tergesen	0.18	Science Fiction
18.18	Michael Douglas	1.62	Cher	0.18	Bruce Willis

L1 regularization

$$\ell(\beta) = \underbrace{\sum_{i=1}^{N} \log P(y_i \mid x_i, \beta)}_{\text{we want this to be high}} \quad - \quad \underbrace{\eta \sum_{j=1}^{F} |\beta_j|}_{\text{but we want this to be small}}$$

- L1 regularization encourages coefficients to be exactly 0.
- η again controls how much of a penalty to pay for coefficients that are far from 0 (optimize on development data)

Activity

- Explore regularization in linear regression
- How does changing the regularization strength affect:
  - Accuracy
  - Rank of important features
  - Number of non-zero features (for L1)