Generalized Low Rank Models

Anqi Fu

Machine Learning Scientist, H2O.ai

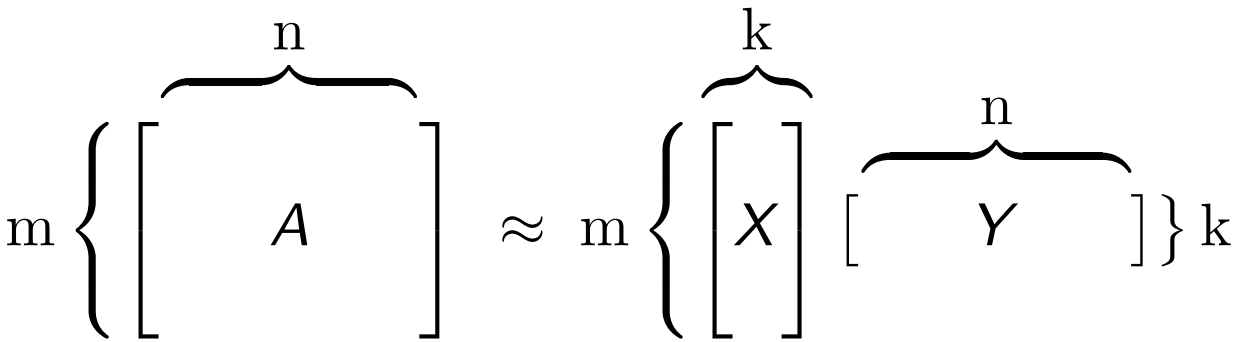
anqi@h2o.ai

Based on work by Madeleine Udell, Corinne Horn, Reza Zadeh and Stephen Boyd

November 6, 2015

What is a Low Rank Model?

- **Given:** Data table A with m rows and n columns
- **Find:** Compressed representation as numeric tables X and Y, where # cols in X = # rows in Y = small user-specified $k \ll \max(m, n)$
- # cols in Y is d = (total dimension of embedded features in A) $\geq n$



- Row of Y = archetypal feature created from columns of A
- Row of X = row of A in reduced feature space
- Can approximately reconstruct A from product XY

Why use Low Rank Models?

- Reduce storage space, e.g. 10 GB compressed to 100 MB
- Increase prediction speed, e.g. 10x speed-up with no accuracy loss
- Identify and visualize important features
- Impute missing data entries

Example 1: Visualizing Walking Stances

time	forehead (x)	forehead (y)	$\cdots$	right toe (y)	right toe (z)
t_1	1.4	2.7	$\cdots$	-0.5	-0.1
t_2	2.7	3.5	$\cdots$	1.3	0.9
t_3	3.3	-.9	$\cdots$	4.2	1.8
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- A contains 151 rows (observations over time) by 124 columns (location of body parts)
- Build a low rank model X, Y with rank $k = 10$
- Rows of Y are principal stances person takes while walking
- Rows of X decompose each bodily position into combination of principal stances

Example 2: Compressing Zip Codes

repeat violator	ZCTA	$\cdots$	violations	penalty	EE's affected
N/A	70525	$\cdots$	9	8100	0
R	75189	$\cdots$	6	935	5
RW	95621	$\cdots$	4	1155	3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- **Train:** U.S. Wage & Hour Division (WHD) compliance actions contains 208,806 rows (cases) and 252 columns (violation info)
- **Response:** Was firm a repeat and/or willful violator?
- **Predictors:** Zip code tabulation area (ZCTA), number of violations, civil penalties, employees (EE's) affected, etc
- Naive approach replaces ZCTA with indicator variables, which 1) is slow, 2) overfits, 3) cannot transfer knowledge between similar ZCTAs

Example 2: Compressing Zip Codes

ZCTA	associate degree	bachelor's degree	$\cdots$	welsh	west indian
01001	1584	1953	$\cdots$	34	57
01002	510	3098	$\cdots$	332	181
01003	27	49	$\cdots$	40	134
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- American Community Survey (ACS) data contains 32,989 rows (unique ZCTAs) by 150 columns (population info)
- Build a low rank model X, Y with rank $k = 10$ and regularization to sparsify features
- Rows of Y are demographic archetypes
- Rows of X map ZCTAs into combination of demographic archetypes

Example 2: Compressing Zip Codes

$$Train = \begin{array}{c} \begin{matrix} y & ZCTA & \cdots \end{matrix} \\ \begin{bmatrix} \text{N/A} & 70525 & \cdots \\ \vdots & \vdots & \ldots \\ \text{R} & 01002 & \cdots \end{bmatrix} \end{array} \quad X = \begin{array}{c} \begin{matrix} ZCTA & archetypes \end{matrix} \\ \begin{matrix} 01001 \\ 01002 \\ \vdots \\ 70525 \end{matrix} \begin{bmatrix} — x_1 — \\ — x_2 — \\ \vdots \\ — x_p — \end{bmatrix} \end{array}$$

- Replace ZCTA col of training data with low rank model (X) of ACS
- Predict if firm will be a repeat violator using modified training data

repeat violator	archetypes	$\cdots$	violations	penalty	EE's affected
N/A	$— x_p —$	$\cdots$	9	8100	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R	$— x_2 —$	$\cdots$	4	225	3

References

- M. Udell, S. Boyd, et al (2014), Generalized Low Rank Models
- Example 1: Visualizing Walking Stance
  - Walking Gait Data
  - Walking Gait Data with Missing Values
- Example 2: Compressing Zip Codes
  - Wage and Hour Division Data
  - American Community Survey Data