ELEN 4903: Machine Learning

Columbia University, Spring 2016

Homework 5: Due May 2, 2016 by 11:59pm

Yuntong Wang
yw2768

Problem 1 (Markov chains) – 50 points

- Let w_t be the 1x759 state vector at step t. Set w_0 to the uniform distribution. Therefore, w_t is the distribution on the state after t steps given that the starting state at time 0 is uniformly distributed.
- Use w_t to rank the teams by sorting in decreasing value according to this vector. List the top 25 teams and their corresponding values in w_t for t = 10, 100, 1000, 2500.

Top 25 team ranks when t = 10
No. 1 , Team: 435 Mount Union , Score : 0.0164
No. 2 , Team: 608 St Thomas , Score : 0.0129
No. 3 , Team: 440 NW Missouri St , Score : 0.0115
No. 4 , Team: 6 Alabama , Score : 0.0113
No. 5 , Team: 363 Marian IN , Score : 0.0086
No. 6 , Team: 343 Linfield , Score : 0.0085
No. 7 , Team: 673 UW-Whitewater , Score : 0.0078
No. 8 , Team: 133 Clemson , Score : 0.0076
No. 9 , Team: 696 Wabash , Score : 0.0073
No. 10 , Team: 465 North Dakota St , Score : 0.0066
No. 11 , Team: 719 West Georgia , Score : 0.0066
No. 12 , Team: 432 Morningside , Score : 0.0065
No. 13 , Team: 668 UW-Oshkosh , Score : 0.0065
No. 14 , Team: 564 Shepherd , Score : 0.0063
No. 15 , Team: 736 Wheaton , Score : 0.0062
No. 16 , Team: 598 St Francis IN , Score : 0.0061
No. 17 , Team: 715 Wesley , Score : 0.0060
No. 18 , Team: 643 Thomas More , Score : 0.0059
No. 19 , Team: 302 Johns Hopkins , Score : 0.0058
No. 20 , Team: 370 Mary Hardin-Baylor , Score : 0.0058
No. 21 , Team: 489 Ohio State , Score : 0.0057
No. 22 , Team: 571 Slippery Rock , Score : 0.0055
No. 23 , Team: 20 Amherst , Score : 0.0053
No. 24 , Team: 611 Stanford , Score : 0.0052
No. 25 , Team: 247 Grand Valley St , Score : 0.0052

Top 25 team ranks when t = 100

No. 1 , Team: 435 Mount Union , Score : 0.0583

No. 2 , Team: 6 Alabama , Score : 0.0320

No. 3 , Team: 440 NW Missouri St , Score : 0.0309

No. 4 , Team: 608 St Thomas , Score : 0.0243

No. 5 , Team: 133 Clemson , Score : 0.0169

No. 6 , Team: 409 Mississippi , Score : 0.0134

No. 7 , Team: 489 Ohio State , Score : 0.0128

No. 8 , Team: 492 Oklahoma , Score : 0.0112

No. 9 , Team: 363 Marian IN , Score : 0.0112

No. 10 , Team: 611 Stanford , Score : 0.0105

No. 11 , Team: 390 Michigan St , Score : 0.0103

No. 12 , Team: 274 Houston , Score : 0.0100

No. 13 , Team: 598 St Francis IN , Score : 0.0087

No. 14 , Team: 343 Linfield , Score : 0.0086

No. 15 , Team: 389 Michigan , Score : 0.0083

No. 16 , Team: 292 Iowa , Score : 0.0079

No. 17 , Team: 624 TCU , Score : 0.0079

No. 18 , Team: 715 Wesley , Score : 0.0077

No. 19 , Team: 202 Emporia St , Score : 0.0076

No. 20 , Team: 215 Florida , Score : 0.0076

No. 21 , Team: 321 LSU , Score : 0.0075

No. 22 , Team: 29 Arkansas , Score : 0.0073

No. 23 , Team: 482 Notre Dame , Score : 0.0072

No. 24 , Team: 673 UW-Whitewater , Score : 0.0069

No. 25 , Team: 219 Florida St , Score : 0.0068

Top 25 team ranks when t = 1000

No. 1 , Team: 6 Alabama , Score : 0.0609

No. 2 , Team: 133 Clemson , Score : 0.0319

No. 3 , Team: 409 Mississippi , Score : 0.0254

No. 4 , Team: 489 Ohio State , Score : 0.0242

No. 5 , Team: 492 Oklahoma , Score : 0.0209

No. 6 , Team: 611 Stanford , Score : 0.0196

No. 7 , Team: 390 Michigan St , Score : 0.0193

No. 8 , Team: 274 Houston , Score : 0.0190

No. 9 , Team: 389 Michigan , Score : 0.0156

No. 10 , Team: 624 TCU , Score : 0.0146

No. 11 , Team: 292 Iowa , Score : 0.0146

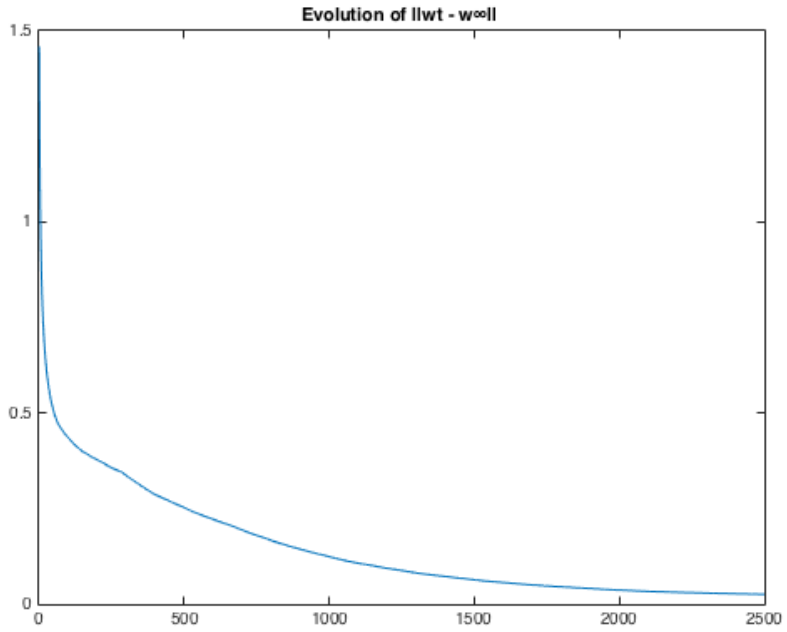
No. 12 , Team: 215 Florida , Score : 0.0144

No. 13 , Team: 321 LSU , Score : 0.0143
No. 14 , Team: 29 Arkansas , Score : 0.0137
No. 15 , Team: 482 Notre Dame , Score : 0.0135
No. 16 , Team: 440 NW Missouri St , Score : 0.0129
No. 17 , Team: 219 Florida St , Score : 0.0128
No. 18 , Team: 629 Tennessee , Score : 0.0126
No. 19 , Team: 435 Mount Union , Score : 0.0125
No. 20 , Team: 53 Baylor , Score : 0.0112
No. 21 , Team: 679 Utah , Score : 0.0110
No. 22 , Team: 442 Navy , Score : 0.0108
No. 23 , Team: 494 Oklahoma St , Score : 0.0107
No. 24 , Team: 476 Northwestern , Score : 0.0107
No. 25 , Team: 458 North Carolina , Score : 0.0101

Top 25 team ranks when t = 2500
No. 1 , Team: 6 Alabama , Score : 0.0667
No. 2 , Team: 133 Clemson , Score : 0.0349
No. 3 , Team: 409 Mississippi , Score : 0.0278
No. 4 , Team: 489 Ohio State , Score : 0.0266
No. 5 , Team: 492 Oklahoma , Score : 0.0228
No. 6 , Team: 611 Stanford , Score : 0.0214
No. 7 , Team: 390 Michigan St , Score : 0.0212
No. 8 , Team: 274 Houston , Score : 0.0208
No. 9 , Team: 389 Michigan , Score : 0.0171
No. 10 , Team: 292 Iowa , Score : 0.0160
No. 11 , Team: 624 TCU , Score : 0.0160
No. 12 , Team: 215 Florida , Score : 0.0158
No. 13 , Team: 321 LSU , Score : 0.0156
No. 14 , Team: 29 Arkansas , Score : 0.0150
No. 15 , Team: 482 Notre Dame , Score : 0.0148
No. 16 , Team: 219 Florida St , Score : 0.0140
No. 17 , Team: 629 Tennessee , Score : 0.0138
No. 18 , Team: 53 Baylor , Score : 0.0122
No. 19 , Team: 679 Utah , Score : 0.0120
No. 20 , Team: 442 Navy , Score : 0.0118
No. 21 , Team: 494 Oklahoma St , Score : 0.0117
No. 22 , Team: 476 Northwestern , Score : 0.0117
No. 23 , Team: 458 North Carolina , Score : 0.0111
No. 24 , Team: 751 Wisconsin , Score : 0.0107
No. 25 , Team: 498 Oregon , Score : 0.0107

- Plot $\|w_t - w_\infty\|$ as a function of t for t = 1 ,…, 2500. What is the value of $\|w_{2500} - w_\infty\|$?

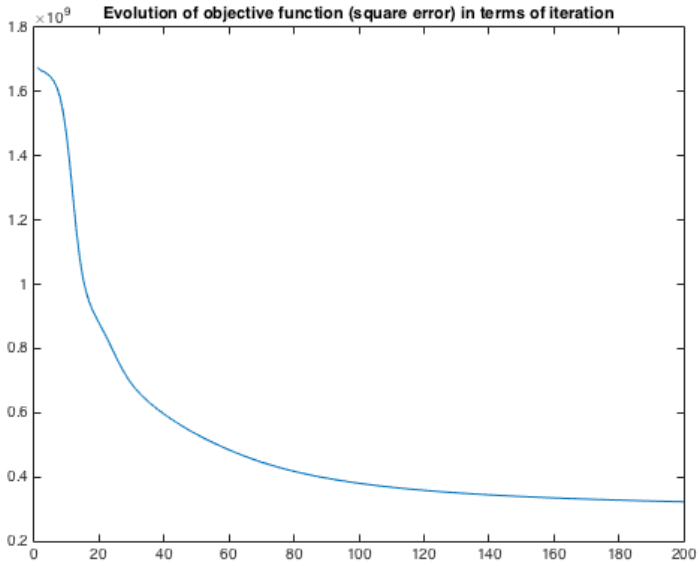
  The value of $\|w_{2500} - w_\infty\|$ is 0.0256.



Problem 2 (Nonnegative matrix factorization) – 70 points

Part 1: The data to be used for Part 1 consists of 1000 images of faces, each originally 32x32, but vectorized to length 1024. The data matrix is therefore 1024x1000.

- Implement and run the NMF algorithm on this data using the Euclidean penalty. Set the rank of the factorization to 25 and run for 200 iterations.
- Plot the objective as a function of iteration.



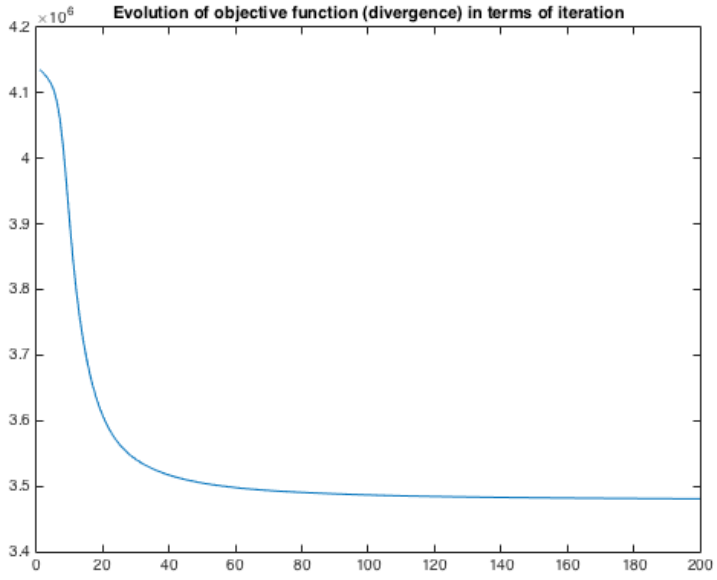
- Pick 10 columns from W and show them as 32x32 images. For each vector you select from W, find the column of H that places the highest weight on this vector and show the corresponding column of X as a 32x32 image.

Figure 1 Image with highest weight in the result (row 1 and row 3) ,their corresponding columns of X (row 2 and row 4)

Part 2: The data to be used for Part 2 consists of 8447 documents from The New York Times. (See below for how to process the data.) The vocabulary size is 3012 words. You will need to use

this data to constitute the matrix X, where X_{ij} is the number of times word i appears in document j. Therefore, X is 3012x8447 and most values in X will equal zero.

- Implement and run the NMF algorithm on this data using the divergence penalty. Set the rank to 25 and run for 200 iterations. This corresponds to learning 25 topics.
- Plot the objective as a function of iteration.



- After running the algorithm, normalize the columns of W so they sum to one. Pick 10 columns of W. For each column you select show the 10 words having the largest weight according to that vector and show the weight. The ith row of W corresponds to the ith word in the "dictionary" provided with the data.

For topic (column) 1, the most dominant words are:

	Word	Probability
No 1.	art	0.0194
No 2.	artist	0.0131
No 3.	museum	0.0106
No 4.	collection	0.0103
No 5.	photograph	0.0094
No 6.	exhibition	0.0092
No 7.	painting	0.0087
No 8.	information	0.0082
No 9.	open	0.0076
No 10.	design	0.0075

For topic (column) 2, the most dominant words are:

Word	Probability

No 1. group 0.0151
No 2. state 0.0125
No 3. public 0.0111
No 4. issue 0.0107
No 5. community 0.0098
No 6. law 0.0095
No 7. require 0.0087
No 8. member 0.0084
No 9. system 0.0083
No 10. program 0.0082

For topic (column) 3, the most dominant words are:

	Word	Probability
No 1.	history	0.0069
No 2.	fact	0.0067
No 3.	view	0.0064
No 4.	world	0.0060
No 5.	great	0.0056
No 6.	question	0.0054
No 7.	point	0.0054
No 8.	sense	0.0052
No 9.	american	0.0052
No 10.	thing	0.0049

For topic (column) 4, the most dominant words are:

	Word	Probability
No 1.	drug	0.0131
No 2.	study	0.0129
No 3.	health	0.0117
No 4.	doctor	0.0112
No 5.	patient	0.0099
No 6.	cause	0.0097
No 7.	medical	0.0093
No 8.	treatment	0.0090
No 9.	report	0.0089
No 10.	research	0.0085

For topic (column) 5, the most dominant words are:

	Word	Probability
No 1.	building	0.0231
No 2.	city	0.0180

No 3. build 0.0144
No 4. house 0.0143
No 5. area 0.0122
No 6. resident 0.0104
No 7. space 0.0104
No 8. project 0.0098
No 9. owner 0.0096
No 10. home 0.0091

For topic (column) 6, the most dominant words are:

	Word	Probability
No 1.	editor	0.0259
No 2.	book	0.0222
No 3.	life	0.0217
No 4.	family	0.0203
No 5.	write	0.0186
No 6.	child	0.0157
No 7.	article	0.0139
No 8.	woman	0.0135
No 9.	page	0.0130
No 10.	writer	0.0129

For topic (column) 7, the most dominant words are:

	Word	Probability
No 1.	executive	0.0481
No 2.	president	0.0439
No 3.	company	0.0292
No 4.	chief	0.0270
No 5.	director	0.0223
No 6.	vice	0.0208
No 7.	business	0.0194
No 8.	chairman	0.0168
No 9.	name	0.0153
No 10.	advertising	0.0119

For topic (column) 8, the most dominant words are:

	Word	Probability
No 1.	food	0.0116
No 2.	serve	0.0076
No 3.	red	0.0072
No 4.	restaurant	0.0072

No 5. white 0.0070
No 6. small 0.0068
No 7. add 0.0067
No 8. taste 0.0066
No 9. water 0.0065
No 10. green 0.0062

For topic (column) 9, the most dominant words are:

	Word	Probability
No 1.	game	0.0224
No 2.	second	0.0176
No 3.	score	0.0174
No 4.	play	0.0142
No 5.	hit	0.0135
No 6.	point	0.0133
No 7.	victory	0.0125
No 8.	third	0.0121
No 9.	lose	0.0108
No 10.	ball	0.0105

For topic (column) 10, the most dominant words are:

	Word	Probability
No 1.	school	0.0628
No 2.	student	0.0440
No 3.	father	0.0384
No 4.	mrs	0.0324
No 5.	son	0.0302
No 6.	graduate	0.0298
No 7.	daughter	0.0291
No 8.	mother	0.0254
No 9.	parent	0.0207
No 10.	teacher	0.0191