Exponential Family Embeddings: Application to Economics

Francisco J. R. Ruiz

With Maja Rudolph, Stephan Mandt, Susan Athey, David M. Blei

December 12th, 2016



Overview

- 1. Exponential Family Embeddings
 - Generalization of word embeddings to other types of data
- 2. Item Embeddings
 - Application to supermarket data

▶ Word embeddings¹ are a powerful approach for analyzing language

- Words are placed in a low-dimensional latent space
- Distances capture semantic similarity
- Capture local word-to-word interactions
- One approach: Model target word given context

¹Bengio et al. (2006); Mikolov et al. (2013); Mnih & Kavukcuoglu (2013); Levy & Goldberg (2014); Pennington et al. (2014); Vilnis & McCallum (2015)

▶ Word embeddings¹ are a powerful approach for analyzing language

- Words are placed in a low-dimensional latent space
- Distances capture semantic similarity
- Capture local word-to-word interactions
- One approach: Model target word given context

"yesterday, the dog and the cat were chasing each other for a while"

p(cat | context)

¹Bengio et al. (2006); Mikolov et al. (2013); Mnih & Kavukcuoglu (2013); Levy & Goldberg (2014); Pennington et al. (2014); Vilnis & McCallum (2015)

is the effects in four unners many in another sources and the ROMEO | less, too can the shorts in foot from bars. This over a stop an make of long T. Make wells for which more another ACMED [line, the CT SERVINIT Away with the approximation of t curct Which are not set of a stand or has the transmission of a stand of the set of In order of lotters with the weath to be and an a SOMMa the proof FIRST SEMMANT with of you all What are the short to be and they are weather to be at fail thing FIRST SEMMANT with of you all whether the short is and they are weather to be at fail thing FIRST SEMMANT with of you all at the short on fail to be and they are weather to be at what you Ah has my minterest which of you all ourselves. Support a done, and they are weather to be at what you Ah has my minterest which of you all in one or are and they are the ourse in the last what you Ah has my minterest which of you all in one or are and they are the ourse in the last you will be a last of the out to be deal. So there a value of the out And the set of the second set of the way between the server will have a be-two Support a hard with some will have a be-two supports and when will have a be-two supports and when will have a be-two supports and will have a be-two supports and two supports to be the support of the matter here, be-two supports and two supports and have a be-two supports and two supports and the support of the support the support of the support of the support of the support of the support the support of the support of the support of the support of the support the support of the support of the support of the support of the support the support of the support of the support of the support of the support the support of the support of the support of the support of the support the support of the support of the support of the support of the support the support of the support of the support of the support of the support the support of the support of the support of the support of the support the support of the support the support of the support the support of the (and the the study) dones have built conclusion to its (find the state in the state in the state is a st Suffer we don't work out it. Suffer we don't work out it. It is now that a soft work to prove the shall go to Mr Im-the new that a soft work to be shall go to Mr Im-the new that a soft work to be shall go to Mr Im-the new that a soft work to be shall now to be start to be s And the second s A JULIET appears above at a window But, soft W were there, they in her head? The briefs Will Here Hull Here, were they in her head? ROM Where there, they in her head? The briefs Will Here Hull H And a my foe's debt. BENVOLIO And Sime we not move, those to idance? NURSE I know ner. IULIET Go ask. To dance? NURSE I know ner. IULIET Go ask. The part of the former that size for which love ground for and the started to the ground the hearest not, he stirred not, he moved the started to the ground the hearest not, he stirred not, he moved the started to the ground the hearest not, he stirred not, he moved the started to the ground the hearest not, he stirred not, he moved the started to the started to the stirred not, he stirred not, he moved the started to the started to the started to the bright low. The former, Romeo! wherefore are tho as thee Take all myself. ROM A the take



[Image from Paul Ginsparg]

Generalize this idea to other types of data

- Tools:
 - Exponential families²
 - Generalized linear models³

²Brown (1986) ³McCullagh and Nelder (1989)

Generalize this idea to other types of data

- Tools:
 - Exponential families²
 - Generalized linear models³

Goals:

- Capture data-to-data interactions
- Improve the predictions of matrix factorization
- Obtain *features* that are useful for downstream tasks

²Brown (1986)

³McCullagh and Nelder (1989)

Applications:

- Market baskets (item embeddings)
- Recommender systems (movie embeddings)
- Neuroscience (neuron embeddings)
- Networks (network embeddings)
- Bird watching (bird embeddings)

▶ ...





Observations x_i

DOMAIN	INDEX	VALUE
Language	position in text i	word indicator
Neuroscience	neuron and time (n, t)	activity level
Shopping	item and basket (m, b)	number purchased
Recommendations	item and user (m, u)	rating

Model Description



- Two latent vectors per data index (embedding, context)
- Model each data point conditioned on its context
- The latent variables interact in the conditional

Model Description



Three ingredients: context, conditional exponential family, embedding structure

Context

Each data point *i* has a *context* C_i, a set of indices of other data points.





• Model the conditional of x_i given its context C_i .

Context

 \bigcirc

- Model the conditional of x_i given its context C_i .
- ► Examples:

DOMAIN	DATA POINT	CONTEXT
Language Neuroscience Shopping Recommendations	word neuron activity purchased item movie rated	surrounding words surrounding neurons other items in basket other movies rated

Exponential Family

Exponential family distribution:

$$x \sim \text{EXPFAM}(\eta, t(x)) = h(x) \exp\{\eta^T t(x) - a(\eta)\}$$

 Examples: Gaussian for reals, Poisson for counts, categorical for categorical, Bernoulli for binary, ...

Exponential Family

Use exponential families for the conditional of each data point,

$$p(x_i | \mathbf{x}_{C_i}) = \text{EXPFAM}(\eta_i(\mathbf{x}_{C_i}), t(x_i))$$

The natural parameter combines the embedding and context vectors,

$$\eta_i(\mathbf{x}_{\mathcal{C}_i}) = f_i\left(\rho[i]^\top \sum_{j \in \mathcal{C}_i} \alpha[j] x_j\right)$$

• $f_i(\cdot)$: Link function (identity, log, ...)

Embedding Structure



> The embedding structure determines how parameters are shared

• Example: $\rho[i] = \rho[j]$ from i = (Oreos, b) and j = (Oreos, b')

Pseudolikelihood

- We model each datapoint conditioned on the others
- Combine these terms in a "pseudolikelihood"

$$\mathcal{L}(
ho, lpha) = \sum_i \left(\eta_i^ op t(x_i) - \pmb{a}(\eta_i)
ight) + \mathcal{L}^{(ext{reg})}$$

▶ The objective resembles a bank of GLMs

Pseudolikelihood

- Fit with stochastic optimization (exponential families simplify the gradients)
- ► The gradient is

$$abla_{
ho}\mathcal{L}(
ho,lpha) = \sum_{i} \left(t(x_i) - \mathbb{E}\left[t(x_i)\right]
ight)
abla_{
ho} \eta_i +
abla_{
ho} \mathcal{L}^{(\mathrm{reg})}$$

Stochastic gradients give justification to negative sampling⁴

⁴Mikolov et al. (2013)

Empirical Study: Neural Activity of Zebrafish



- Data: Calcium expression levels of 10K neurons across time
- Context: Nearby neurons (KNN)
- EXPFAM: Gaussian
- Structure: Latent vectors shared across time

Empirical Study: Neural Activity of Zebrafish

► Gaussian exponential family embeddings outperform Gaussian MF

	single neur	on held out	25% of neur	ons held out
Model	K = 10	K = 100	K = 10	K = 100
FA	0.261 ± 0.004	0.251 ± 0.004	0.261 ± 0.004	0.252 ± 0.004
G-EMB	0.226 ± 0.003	$\textbf{0.222} \pm \textbf{0.003}$	$\textbf{0.233} \pm \textbf{0.003}$	$\textbf{0.230} \pm \textbf{0.003}$
NG-EMB	0.238 ± 0.004	0.233 ± 0.003	$\textbf{0.258} \pm \textbf{0.004}$	0.244 ± 0.004

Empirical Study: Market Basket Analysis



- > Data: Purchase counts of items in shopping trips at grocery store
- Context: Other items in basket
- EXPFAM: Poisson
- Structure: Latent vectors shared across baskets

Empirical Study: Market Basket Analysis



Datasets:

- Safeway at category level (6.8M purchases, 635K trips, 478 items)
- Safeway at item level (5.6M purchases, 620K trips, 6K items)
- IRI data⁵ (700K purchases, 200K trips, 8K items)

⁵Bronnenberg et al. (2008)

Empirical Study: Market Basket Analysis

- Poisson exponential family embeddings outperform Poisson MF
- Downweighting the zeros helps⁶

Model	K = 20	K = 100
P-EMB	-7.497 ± 0.007	-7.199 ± 0.008
P-EMB (dw)	-7.110 ± 0.007	$-\textbf{6.950} \pm \textbf{0.007}$
AP-EMB	-7.868 ± 0.005	-8.414 ± 0.003
HPF^{7}	-7.740 ± 0.008	-7.626 ± 0.007
Poisson PCA ⁸	-8.314 ± 0.009	-11.01 ± 0.01

 6 Hu et al. (2008); Liang et al. (2016) 7 Gopalan et al. (2015) 8 Collins et al. (2001)



2D projection of the context vectors (Safeway at group level)

	 shortening 	powc - fle	lered sugar condensed milk extracts pur
		granulated sugar	baking ingredients brown sugar
 pie crust 			evaporated milk
• cream		∘ pie filli	ng
			∘ salt
special	y/miscellaneous o	frozen pastry dough deli items	

infant formula

disposable diapers disposable pants baby accessories

baby/youth wipes

· infant toiletries

childrens/infants analgesics



2D projection of the context vectors (Safeway at item level)





Market Basket Analysis: Queries

Queries for most similar item:

Maruchan chicken ramen	Yoplait strawberry yogurt
Maruchan creamy chicken ramen	Yoplait apricot mango yogurt
Maruchan oriental flavor ramen	Yoplait strawberry orange smoothie
Maruchan roast chicken ramen	Yoplait strawberry banana yogurt

Mountain Dew soda	Dean Foods 1 $\%$ milk
Mountain Dew orange soda	Dean Foods 2 % milk
Mountain Dew lemon lime soda	Dean Foods whole milk
Pepsi classic soda	Dean Foods chocolate milk

Identify complements and substitutes

Define the exchangeability metric

$$E_{ij} = -\sum_{k \neq i,j} D_{\mathrm{KL}}(p_{k \mid j} \mid\mid p_{k \mid j})$$

and the **co-purchase metric** between items (i, j)

$$C_{ij} = \sigma_{ji} \log \left(\frac{\sigma_{ji}}{1 - \sigma_{ji}} \right), \qquad \sigma_{ij} \triangleq p(x_i \neq 0 \mid x_j)$$

(We use the symmetrized versions)

▶ High co-purchase score *C_{ij}* (category level):

ITEM 1	ITEM 2	SCORE (RANK)
organic vegetables	organic fruit	6.18 (01)
vegetables (< 10 oz)	beets (\geq 10 oz)	5.64 (02)
baby food	disposable diapers	3.43 (32)
stuffing	cranberries	3.30 (36)
gravy	stuffing	3.23 (37)
pie filling	evaporated milk	3.09 (42)
deli cheese	deli crackers	2.87 (55)
dry pasta/noodles	tomato pasta/sauce/puree	2.73 (63)
mayonnaise	mustard	2.61 (69)
cake mixes	frosting	2.49 (78)

• High score $E_{ij} - C_{ij}$ (category level):

item 1	ITEM 2	SCORE (RANK)
bouquets	roses	0.20 (01)
frozen pizza 1	frozen pizza 2	0.18 (02)
bottled water 1	bottled water 2	-0.07 (03)
carbonated soft drinks 1	carbonated soft drinks 2	-0.12 (04)
orange juice 1	orange juice 2	-0.37 (05)
bathroom tissue 1	bathroom tissue 2	-0.58 (06)
bananas 1	bananas 2	-0.61(07)
salads-convenience 1	salads-convenience 2	-0.63 (08)
potatoes 1	potatoes 2	-0.66 (09)
bouquets	blooming	-1.18(10)

▶ High co-purchase score *C_{ij}* (item level):

item 1	ITEM 2	SCORE (RANK)
ygrt peach ff	ygrt m×d berry ff	19.83 (0001)
s&w beans garbanzo	s&w beans red kidney	14.42 (0002)
whiskas cat fd beef	whiskas cat food tuna/chicken	8.45 (0149)
parsnips loose	rutabagas	8.32 (0157)
celery hearts organic	apples fuji organic	4.36 (0995)
85p In gr beef patties 15p fat	sesame buns	4.35 (1005)
kiwi imported	mangos small	3.22 (1959)
colby jack shredded	taco bell taco seasoning mix	2.89 (2472)
star magazine	in touch magazine	2.87 (2497)
seasoning mix fajita	mission tortilla corn super sz	2.87 (2500)

• High score $E_{ij} - C_{ij}$ (item level):

ITEM 1	ITEM 2	SCORE (RANK)
coffee drip grande	coffee drip venti	-0.33 (001)
sandwich signature reg	sandwich signature lrg	-1.17 (020)
market bouquet	alstromeria/rose bouquet	-2.89 (186)
sushi shoreline combo	sushi full moon combo	-3.76 (282)
semifreddis bread baguette	crusty sweet baguette	-7.65 (566)
orbit gum peppermint	orbit gum spearmint	-7.96 (595)
snickers candy bar	3 musketeers candy bar	-7.97 (598)
cheer Indry det color guard	all Indry det liquid fresh rain	-7.99 (602)
coors light beer btl	coors light beer can	-8.12 (621)
greek salad signature	neptune salad signature	-8.15 (630)

Conclusions

- Word embeddings have become a staple in NLP
 We distilled its essential elements; generalized to other data
- Compared to classical factorization, good performance in many data (movie ratings, neural activity, scientific reading, shopping baskets)

Conclusions

- Word embeddings have become a staple in NLP
 We distilled its essential elements; generalized to other data
- Compared to classical factorization, good performance in many data (movie ratings, neural activity, scientific reading, shopping baskets)

Future/ongoing work:

- How can we capture higher-order structure in the embeddings?
- Why downweight the zeros?
- How to model user (or document) heterogeneity?
- How can we include price and other complexities?
- How can we study causal effects?

Thank you for your attention!

