Interpretable representation of latent variables in matrix factorizations

Bachelor's Thesis - Presentation

Lam Ngo

Faculty of Computer Science, University of Vienna

September 30, 2021

Introduction

Figure: Why does Trump appear when one googles "idiot" ? [1]

[1] https://www.youtube.com/watch?v=o6zfp6lRw2E&t=33s&ab_channel=PBSNewsHour

Overview

1. Why is interpretability important (in deep learning models)?
2. Latent Factor Models and Interpretability
   - Autoencoder
   - Data
   - Interpretable Lens Variable Model
3. Experiments
4. Conclusion

Literature

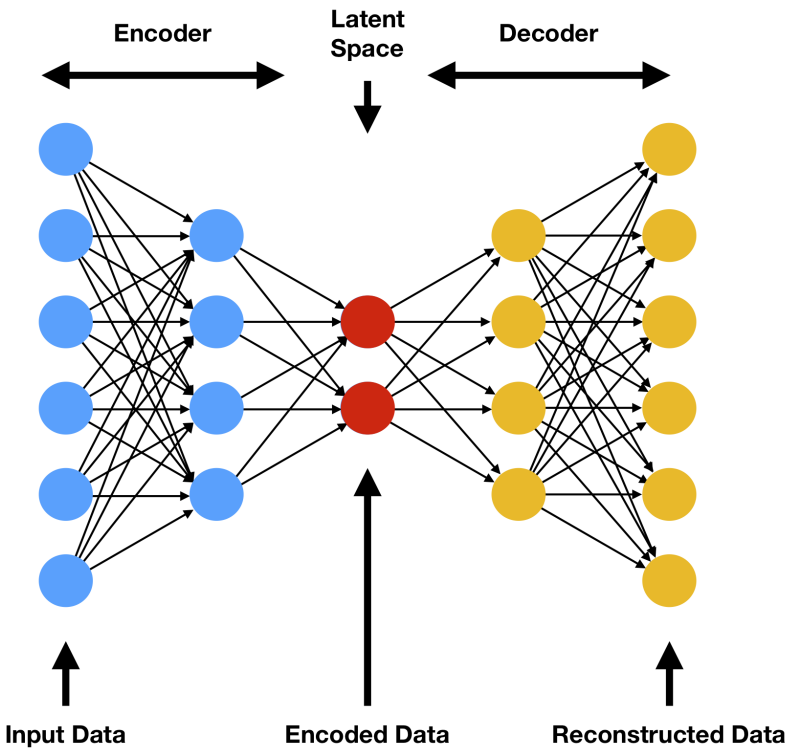
- Interpretability: relationship to something humanly understandable established[2]
- Deep learning models come close to the accuracy of human perception but are black box models [3]
- GDPR: data subject has right to obtain meaningful information about the logic involved in automated decision making [4]

[2] Adel, Ghahramani, and Weller, "Discovering Interpretable Representations for Both Deep Generative and Discriminative Models" (Oct. 2018).

[3] Chakraborty et al., "Interpretability of deep learning models: A survey of results" (2017).

[4] *2018 reform of EU data protection rules* (May 25, 2018).

Problem - opaque latent factor models



Autoencoder

- $\arg\min_{A,B} \quad \mathbb{E}[\Delta(x, B \circ A(x))]$
- $A : \mathbb{R}^n \mapsto \mathbb{R}^p$ and $B : \mathbb{R}^p \mapsto \mathbb{R}^n$
- Goal: make latent space interpretable

Figure: Architecture of an Autoencoder[5]

[5] https://www.compthree.com/blog/autoencoder/

Data

1. Latent factor model can predict if student answers questions correctly
2. Data is from Eedi, an online education platform [6]
3. Students answered diagnostic mathematical questions
4. Approx. 15m student-question pairs, 27k questions and 120k students
5. Meta data available

[6] *Diagnostic Questions - The NeurIPS 2020 Education Challenge* (2020).

Data - Student/Question pairs

	QuestionId	UserId	AnswerId	IsCorrect	CorrectAnswer	AnswerValue
0	16997	65967	12453206	0	4	2
1	16531	62121	15686710	1	1	1
2	15911	50013	13598796	0	3	1
3	1701	104909	10511925	0	4	3
4	22896	21748	941747	0	1	4

Data - Question Meta Data

	QuestionId	SubjectId
0	13090	[3, 32, 71, 77, 141, 185, 186, 214]
1	1855	[3, 71, 75, 86, 178]
2	10423	[3, 32, 38, 239]
3	2290	[3, 32, 33, 144]
4	12785	[3, 32, 33, 144]

Data - Topic Meta Data

	SubjectId	Name	ParentId	Level
0	3	Maths	NaN	0
1	32	Number	3.0	1
2	33	BIDMAS	144.0	3
3	34	Upper and Lower Bounds	141.0	3
4	35	Calculator Use	32.0	2

2-Step Procedure for post-hoc Interpretability

1. Train the matrix factorization model
   - $\hat{c}_{ij} = s_i^T q_j + b_i + b_j$
   - c_{ij} = question j answered correctly by student i
   - s_i and q_j are latent factors
   - b_i and b_j are biases
   - $\min_{u^*,q^*} \sum_{i,j \in \mathcal{A}} (c_{ij} - s_i^T q_j - b_i - b_j)^2 + \lambda(||s_i||^2 + ||q_j||^2 + b_i^2 + b_j^2)$
2. Train a relationship between latent space and side information
   - $\hat{c}_{ij} = g(s_i) \times (topics_j * difficulty_j)$
   - g is a linear layer
   - $difficulty_j$ is the share of correct answers for question j of all students

Matrix Factorization Training

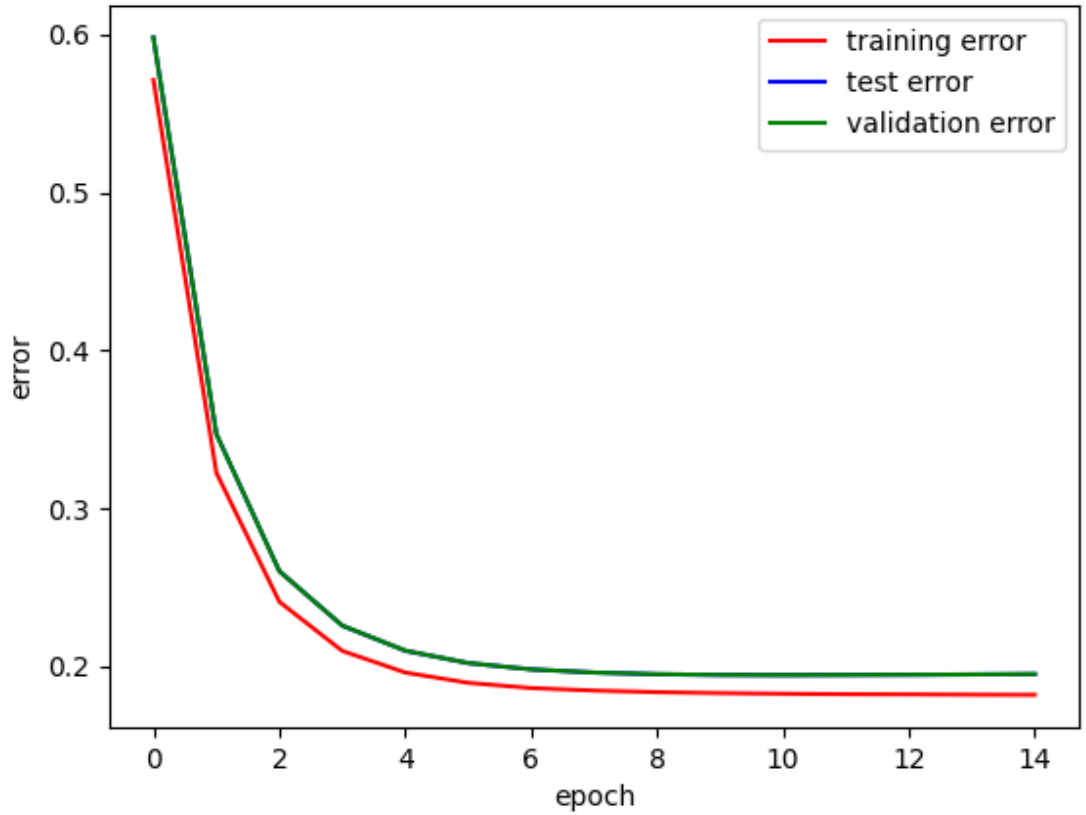


Figure: Error convergence

Lensmodel Training

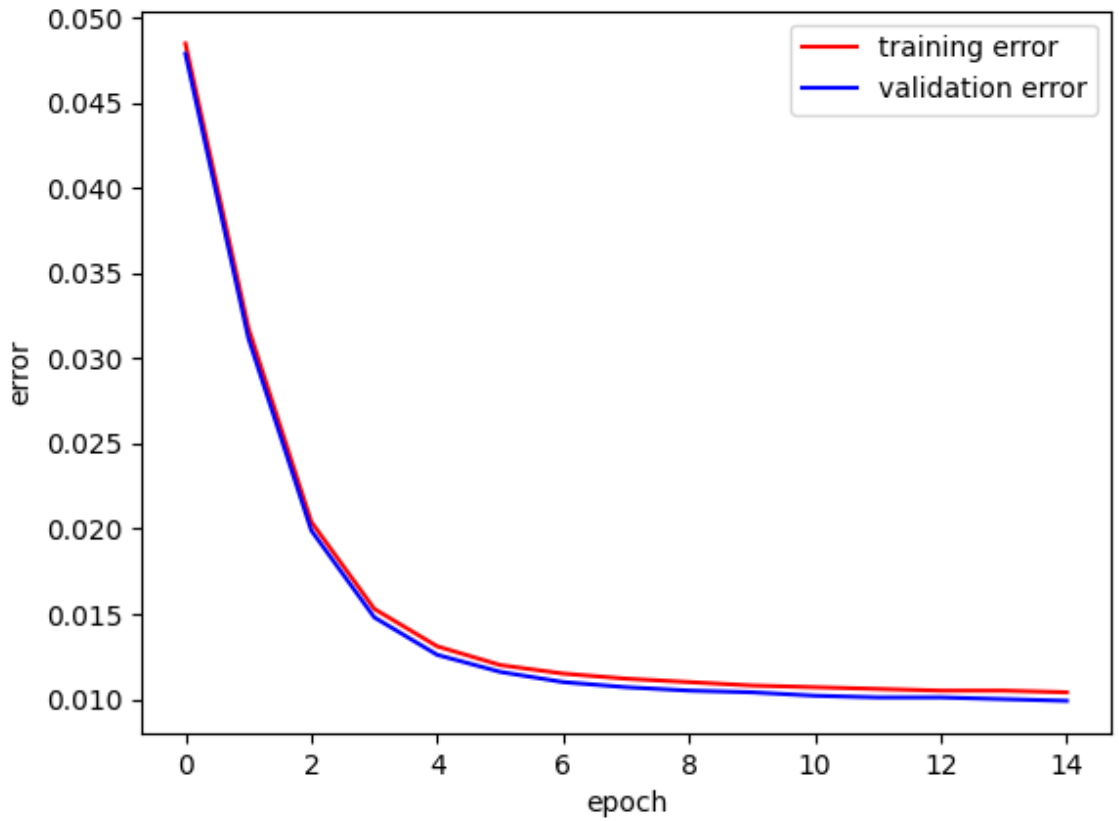


Figure: Error convergence

Student Knowledge Visualization

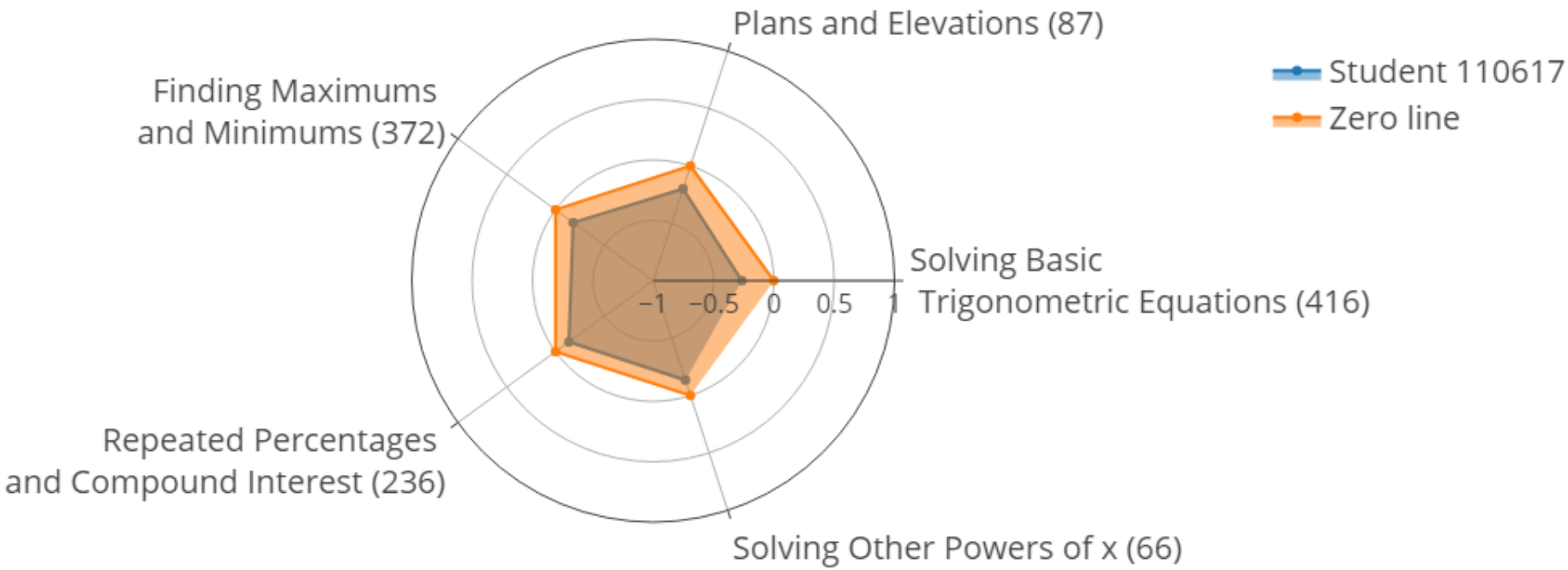


Figure: Lowest knowledge values for Student 110617

Student Knowledge Visualization

Subject	Knowledge value	Mean(IsCorrect)	Mean(difficulty) wrong answers	#questions
416	-0.26	0.00	0.37	1
87	-0.20	0.66	0.53	9
372	-0.18	0.00	0.46	1
236	-0.13	1	-	2
66	-0.13	0.5	0.51	4

Table: Lowest knowledge values for Student 110617

Student Knowledge Visualization

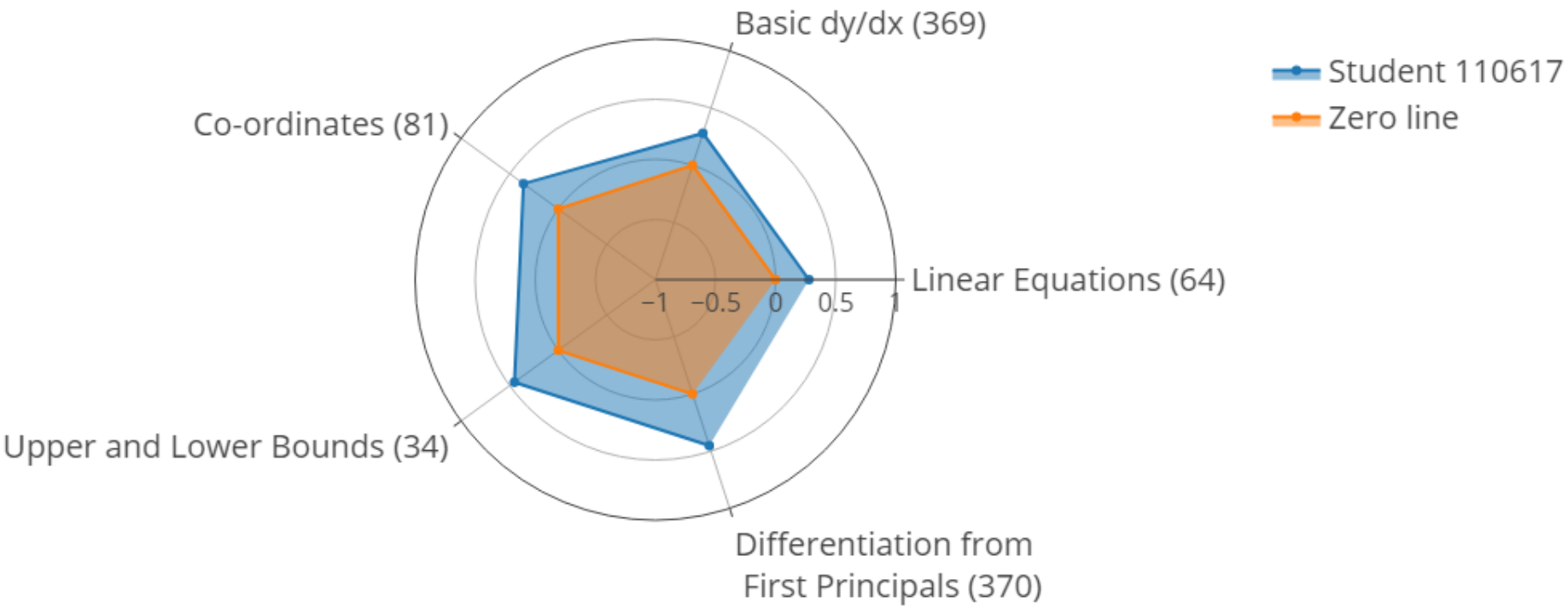


Figure: Highest knowledge values for Student 110617

Student Knowledge Visualization

Subject	Knowledge value	Mean(IsCorrect)	Mean(difficulty) wrong answers	#questions
64	0.28	0.75	0.51	8
369	0.28	1	-	2
81	0.36	0.76	0.48	17
34	0.45	0.77	0.47	56
370	0.45	1	-	1

Table: Highest knowledge values for Student 110617

Knowledge Correlations

Subject1	Subject2	Correlation
Fractions, Decimals and Percentage Equivalence	Fractions	0.88
Writing and Simplifying Expressions	Perimeter and Area	0.86
Inequalities on Number Lines	Solving Equations	0.85
Indices, Powers and Roots	Volume of Prisms	0.84
Formula	Data Representation	0.83

Table: Highest Correlations between Subjects

Summary

1. Main contribution is a simple way to obtain interpretability in matrix factorization
2. Major limitation: no full matrix imputation which would lead to more precise difficulty measure[7]
3. Possible extensions: advanced matrix factorization methods, other side information

[7] Wang et al., "Educational Question Mining At Scale: Prediction, Analysis and Personalization" (2021).

2018 reform of EU data protection rules. https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf. European Commission. May 25, 2018. (Visited on 09/17/2021).

Adel, Tameem, Zoubin Ghahramani, and Adrian Weller. "Discovering Interpretable Representations for Both Deep Generative and Discriminative Models". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 50–59. URL: http://proceedings.mlr.press/v80/adel18a.html.

Chakraborty, Supriyo et al. "Interpretability of deep learning models: A survey of results". In: *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation*

(SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). 2017, pp. 1–6. DOI: 10.1109/UIC-ATC.2017.8397411.

Diagnostic Questions - The NeurIPS 2020 Education Challenge. https://competitions.codalab.org/competitions/25449. 2020. (Visited on 09/17/2021).

Wang, Zichao et al. "Educational Question Mining At Scale: Prediction, Analysis and Personalization". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 17. 2021, pp. 15669–15677.