

TRI Profundo: uma aplicação de métodos de redes neurais profundas à Teoria de Resposta ao Item

Lucas de Moraes Bastos¹

¹ Universidade de Brasília; lucasdmorebastos@gmail.com

Resumo

Este trabalho introduz um novo método de estimativa no contexto de TRI, baseado em uma representação via Redes Neurais Artificiais (RNA) do modelo conhecido como ML3. Essa técnica foi aplicada ao Exame Nacional do Ensino Médio (Enem), resultando em estimativas com maior acurácia do que métodos tradicionais de estimativa, como o utilizado oficialmente para a correção da prova. Dessa forma, foi demonstrado como a aplicação do otimizador ADAM, frequentemente adotado para o treinamento de modelos de Inteligência Artificial, no contexto de uma avaliação educacional de nível nacional possibilita incorporar a capacidade preditiva de redes neurais profundas à interpretabilidade da TRI.

Palavras-chave: Teoria de Resposta ao Item (TRI); Redes Neurais Artificiais (RNA); aprendizado de máquina; inteligência artificial; TRI Profunda

1. Introdução

O Exame Nacional do Ensino Médio (Enem) é um teste multidisciplinar administrado anualmente, desde 1998, pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). É a maior avaliação educacional do Brasil, de grande relevância, uma vez que é a principal porta de entrada para a maioria das universidades públicas do país e proporciona uma medição da qualidade e efetividade da educação pública brasileira. Dessa forma, é crucial que a credibilidade e acurácia da mensuração do nível de conhecimento dos alunos seja garantida.

O modelo atualmente utilizado pelo Inep é o tradicional Modelo Logístico de 3 Parâmetros, ML3 (BIRNBAUM, 1968), que possui pressuposições como independência local e de uma distribuição geradora dos dados (gaussiana no caso do Enem), as quais, quando não satisfeitas, podem levar a resultados insatisfatórios, comprometendo a acurácia das estimativas,

principalmente no que concerne à variável latente de habilidade dos alunos, que corresponde diretamente às notas dos alunos.

Nos últimos anos, as redes neurais profundas ressurgiram em popularidade, sendo aplicadas nas mais diversas áreas com resultados considerados estado da arte, de física de partículas à produção de textos. Métodos recentes da literatura combinam o poder preditivo das redes neurais profundas à interpretabilidade e aplicabilidade da TRI, culminando em uma nova área que vem sendo chamada de TRI Profunda, primeiramente na tarefa de Rastreamento do Conhecimento - *Knowledge Tracing* (YEUNG, 2019), e, mais recentemente, em aplicações na Teoria do Teste (TSUTSUMI et al., 2021).

2. Objetivos

- Investigar o processo de estimação oficial do Enem e replicá-lo em linguagem aberta;
- Introduzir uma nova estratégia de otimização dos parâmetros do modelo TRI ML3 usando o algoritmo ADAM e uma representação na forma de uma rede neural artificial;
- Comparar o desempenho preditivo da abordagem proposta em relação às outras formas de estimação usualmente adotadas.

3. Metodologia

Para a tarefa de investigação e replicação do processo de estimação conduzido pelo Inep, o pacote *mirt*, do R, foi utilizado, por sua robustez e flexibilidade, permitindo a escolha dos parâmetros e métodos equivalentes ou mais próximos possíveis da estimação oficial. Os microdados do Enem 2022 contemplam uma tabela com 3.389.832 observações e 76 colunas, contendo as respostas dos alunos, e outra tabela com os códigos de referência do item e do caderno de prova. O pré-processamento desses dados consistiu em organizar as respostas dos alunos (uma *string* com as letras em sequência das alternativas escolhidas como respostas) em variáveis binárias para cada caderno de prova. Essas respostas são utilizadas para a estimação pelo método da máxima verossimilhança marginal, utilizando processos como quadratura Gaussiana e o algoritmo EM.

Para a construção do modelo de TRI como RNA, o pacote *keras*, do R, foi utilizado. As respostas dos alunos a cada item e o vetor de identificação de cada aluno são transformados por camadas de *embedding* iniciais, com uma relação direta do aluno à sua respectiva habilidade e do item aos seus três parâmetros, funcionando como uma tabela de referência eficiente.

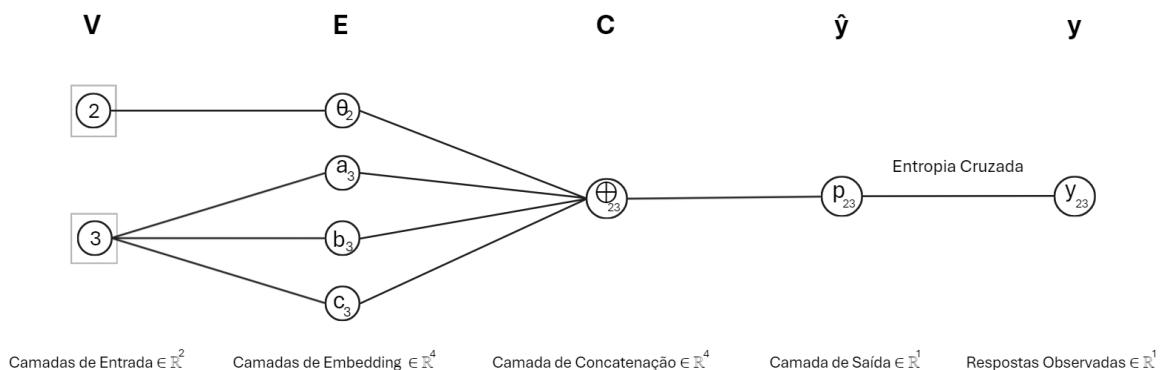


Figura 1: Modelo ML3 como RNA. Essa visualização destaca o caminho na rede para o segundo aluno ($v_s = 2$) e terceiro item ($v_q = 3$).

Com N alunos e I itens, as entradas (*inputs*) são vetores um para um:

$$\begin{aligned}\mathbf{V} &= \{v_s, v_q\}, \\ v_s &= i; i = \{1, 2, \dots, N\}, \\ v_q &= j; j = \{1, 2, \dots, I\},\end{aligned}$$

ligada por *embedding* a:

$$\begin{aligned}\mathbf{E} &= \{e_s, e_q\}, \\ e_s &= \theta_i, \\ e_q &= \{a_j, b_j, c_j\}.\end{aligned}$$

A camada de concatenação $\oplus = \{\theta_i, a_j, b_j, c_j\}$ é necessária para que se possa utilizar a função de ativação *softmax* customizada para a saída (*output*) \hat{y} :

$$\begin{aligned}\hat{y}_{ij} &= \sigma(c_j) + (1 - \sigma(c_j))\text{softmax}[a_j(\theta_i - b_j)] \\ &= \sigma(c_j) + (1 - \sigma(c_j)) \frac{1}{1 + \exp[a_j(\theta_i - b_j)]},\end{aligned}$$

Em que $\sigma(\cdot)$ é a função sigmoide, de forma a garantir que c_j esteja no intervalo [0,1].

A função de perda é a entropia cruzada, que corresponde diretamente à máxima verossimilhança marginal utilizada na estimação oficial:

$$l(\hat{y}_{ij}, y_{ij}) = -[y_{ij} \log(\hat{y}_{ij}) + (1 - y_{ij}) \log(1 - \hat{y}_{ij})],$$

que é otimizada, pelo algoritmo de aprendizado de máquina Adam.

4. Resultados e Discussão

Nesta seção, serão apresentados os resultados comparados entre três métodos ou processos de estimação para a nota do Enem: a replicação no R do método oficial feita com o *mirt*; a

estimação oficial do Inep conduzida no BILOG; e, por fim, a estimação do modelo TRI ML3 usando o algoritmo ADAM, ou ML3 como RNA.

Pode-se observar na Figura 2, pelo gráfico à esquerda, que a replicação no R do método oficial de estimação das notas dos alunos, feito com todas as informações abertas ao público, aproximou-se da estimação oficial do Inep conduzida no BILOG, com coeficiente de correlação de Pearson de 98,95%.

Pelo gráfico à direita, é possível visualizar como a estimação da RNA do ML3 (curva em azul) aproxima-se melhor da probabilidade empírica de acerto do que os métodos tradicionais, com o *mirt* e o BILOG seguindo curvas (amarela e preta, respectivamente) bem próximas.

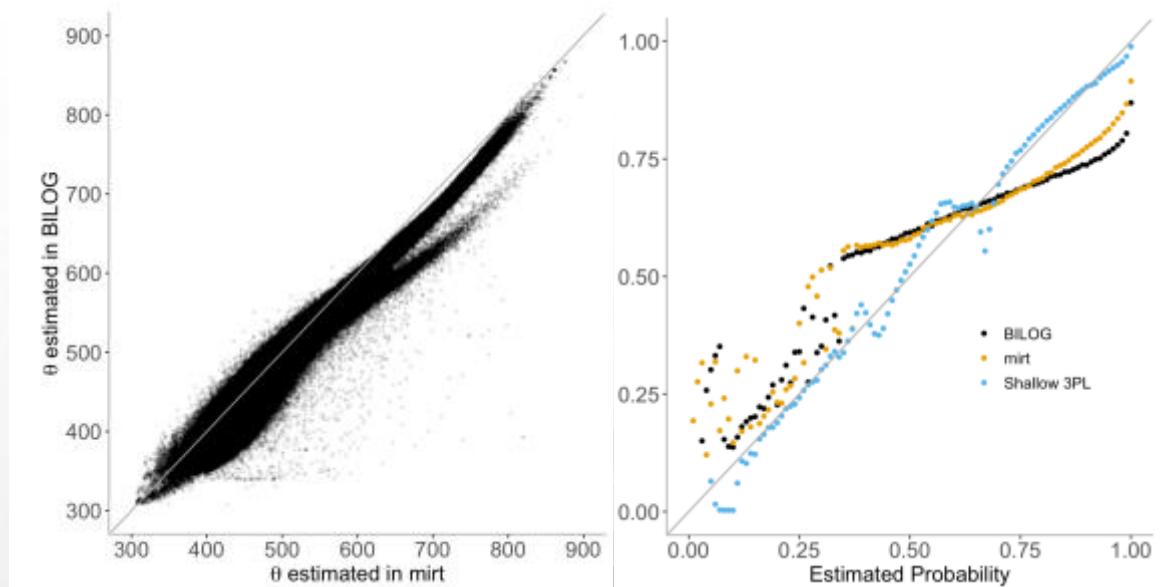


Figura 2: À esquerda: gráfico de dispersão entre a estimação conduzida pelo Inep (no *software* proprietário BILOG) e sua replicação no *mirt* para o parâmetro de habilidade do aluno θ . À direita: gráfico de dispersão entre probabilidades estimadas e probabilidades empíricas, obtidas pela média dos agrupamentos em intervalos até a segunda casa decimal dos \hat{y}_{ij} e y_{ij} .

Pela Tabela 1, pode-se observar que os valores estimados para o parâmetro de habilidade do aluno, θ , entre os três métodos, apresentaram correlações altas. Todavia, os valores estimados dos parâmetros dos itens não se aproximaram entre os três métodos de estimação, exceto em relação ao parâmetro de dificuldade do item, b .

	ρ_a	ρ_b	ρ_c	ρ_θ
BILOG vs. <i>mirt</i>	67,12%	89,99%	53,89%	98,95%
BILOG vs. ML3 como RNA	26,06%	89,92%	62,07%	92,45%
<i>mirt</i> vs. ML3 como RNA	17,94%	91,49%	46,72%	94,19%

Tabela 1: Correlações entre os valores estimados pelo Inep, pelo *mirt* e pela rede neural para os parâmetros dos itens (a , b e c) e o parâmetro dos alunos (θ).

Pela Tabela 2, as métricas de performance entre os métodos e a Teoria Clássica do Teste (TCT) podem ser comparadas, destacando-se a menor perda pelo ML3 como RNA e menor Erro Quadrático Médio (EQM).

Método	Perda (<i>loss</i>)	EQM
BILOG	59,79%	18,74%
<i>mirt</i>	57,85%	18,71%
ML3 como RNAProfun	53,25%	17,62%
TCT	58,08%	19,71%

Tabela 2: Performance de cada método aplicado à estimação das notas do Enem, em que TCT corresponde à Teoria Clássica do Teste, sendo apenas a soma das respostas corretas, na qual a probabilidade estimada de acerto corresponde ao percentual de acerto observado na amostra.

5. Conclusões

Este trabalho contribui à complexa tarefa de avaliar o traço latente de conhecimento ou habilidade de uma dada população, processo de grande relevância educacional e socioeconômica.

Os métodos utilizados pelo Inep na condução do processo de estimação das notas do Enem foram estudados e replicados em linguagem aberta, utilizando o *mirt*, da maneira mais próxima possível do processo oficial, incorporando o algoritmo EM e as aproximações numéricas utilizadas. Este primeiro objetivo conduziu a resultados satisfatórios, com alta correlação obtida entre as notas estimadas em linguagem aberta e as notas estimadas pelo Inep.

Com o *keras*, algoritmos de aprendizagem de máquina e técnicas de treinamento de redes neurais artificiais foram utilizadas na estimação por máxima verossimilhança. Isso confluí em uma nova abordagem de cálculo das notas do Enem e em uma nova forma de estimação dos

parâmetros de um modelo ML3, capaz de uma estimativa mais precisa e calibrada sem assumir nenhuma distribuição geradora de dados, superando as pressuposições do modelo tradicional. Para trabalhos futuros, podem ser realizados estudos de simulação, que permitam a análise da performance do novo modelo em cenários diversos. Entretanto, o próximo passo de maior relevância seria estimar as notas do Enem utilizando o novo modelo com os parâmetros dos itens fixos, com os valores obtidos pelos pré-testes realizados anteriormente à prova aplicada nacionalmente, de forma que seja possível propor este novo método para o cálculo das notas.

Referências

- BIRNBAUM, A. Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. Lord, F.M. and Novick, M.R., Eds., Statistical Theories of Mental Test Scores, Addison-Wesley, Reading, 397-472, 1968.
- BOCK, R. D. and MISLEVY. Marginal maximum likelihood estimation of item parameters. An application of an EM algorithm. *Psychometrika*, 46, 443-459, 1981.
- GOODFELLOW, Ian, BENGIO, Yoshua, and COURVILLE, Aaron. Deep Learning. Disponível em: <http://www.deeplearningbook.org>. MIT Press. 2016. Acesso em: 28 de ago. de 2024.
- INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA /MINISTÉRIO DA EDUCAÇÃO. Exame Nacional do Ensino Médio (Enem): fundamentação teórico-metodológica. 2005. Disponível em: https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/enem_exame_nacional_do_ensino_medio_fundamentacao_teorico_metodologica.pdf. Acesso em: 28 de ago. de 2024.
- STROUD, A. H. and SECREST, Don. Gaussian Quadrature Formulas. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, 47, 138-139, 1967.
- TSUTSUMI E.; KINOSHITA, R. and UENO, M. Deep Item Response Theory as a Novel Test Theory Based on Deep Learning. *Electronics*, 10, 1020-1020, 2021.
- YEUNG, C. Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory. 2019. Disponível em: <https://doi.org/10.48550/arXiv.1904.11738>. Acesso em: 28 de ago. de 2024.