Learning Latent Parameters without Human Response Patterns: Item Response Theory with Artificial Crowds

John P. Lalor [1,2*] Hao Wu [3] Hong Yu [1,4]

[1]University of Massachusetts [2]University of Notre Dame [3]Vanderbilt University [4]Bedford VA

*Email: john.lalor@nd.edu, Web: http://jplalor.github.io

Introduction

Evaluating new machine learning models in vision and language relies on aggregate metrics. Accuracy on a held-out test set is the metric by which newly proposed models are compared with the current literature. However, such methods do not consider the differences between specific elements of a data set. There is a need to model the difficulty of the data sets used in machine learning to guide progress in the field and to place the progress of new models into context. Item Response Theory (IRT) has been used for many years to jointly model latent parameters of data points such as difficulty and latent human ability [1]. Using IRT in machine learning can provide valuable insights for machine learning model interpretability and evaluation [6, 5]. The need for human inputs is a significant bottleneck to learning IRT models for machine learning data. We propose learning IRT models with variational inference (VI) methods [4] using machine-generated data, removing humans from the loop and using the outputs of neural networks (NNs) as IRT model inputs. We demonstrate the effectiveness of learning IRT models with VI using NN-generated data through quantitative and qualitative analyses for image classification and natural language inference tasks.

Item Response Theory

IRT models are designed to estimate latent ability parameters (θ) of subjects and latent item parameters such as difficulty of items (b). The probability that subject j will answer item i correctly is:

$$p(y_{ij} = 1|\theta_j, b_i) = \frac{1}{1 + e^{-(\theta_j - b_i)}} \quad (1)$$

The probability that subject j will answer item i incorrectly is $p(y_{ij} = 0|\theta_j, b_i) = 1 - p(y_{ij} = 1|\theta_j, b_i)$. The likelihood of a data set of RPs Y from J subjects to a set of I items is:

$$p(Y|\theta, b) = \prod_{j=1}^{J} \prod_{i=1}^{I} p(Y_{ij} = y_{ij}|\theta_j, b_i) \quad (2)$$

The item parameters are typically estimated by marginal maximum likelihood (MML) via an Expectation-Maximization (EM) algorithm [2], in which subject parameters are considered random effects $\theta_i \sim N(0, \sigma_\theta^2)$ and marginalized out. Once item parameters are learned, subjects' θ parameters are scored typically with maximum a posteriori (MAP) estimation.

IRT with Variational Inference

Bayesian methods in IRT assume that the individual θ and b parameters in Eq. (2) both follow Gaussian prior distributions and make inference through the resultant joint posterior distribution $\pi(\theta, b|Y)$. As this posterior is usually intractable, VI approximates it by the variational distribution:

$$q(\theta, b) = \prod_{j=1}^{J} \pi_j^\theta(\theta_j) \prod_{i=1}^{I} \pi_i^b(b_i) \quad (3)$$

Where $\pi_j^\theta()$ and $\pi_i^b()$ denotes different Gaussian densities for different parameters whose means and variances are determined by minimizing the KL-Divergence between $q(\theta, b)$ and $\pi(\theta, b|Y)$.

The choice of priors in Bayesian IRT can vary. Prior work has shown that vague and hierarchical priors are both effective [8]. We experiment with both in this work. A vague prior assumes $\theta_j \sim N(0, 1)$ and $b_i \sim N(0, 10^3)$, where the large variance indicates a lack of information on the difficulty parameters. A hierarchical Bayesian model assumes

$$\theta_j \mid m_\theta, u_\theta \sim N(m_\theta, u_\theta^{-1})$$
$$b_i \mid m_b, u_b \sim N(m_b, u_b^{-1})$$
$$m_\theta, m_b \sim N(0, 10^6)$$
$$u_\theta, u_b \sim \Gamma(1, 1)$$

Data and Experiments

For our experiments we used two computer vision data sets, MNIST and CIFAR, as well as one natural language inference data set, SNLI. For each data set, we built a simple deep learning model in order to generate response patterns. For MNIST and CIFAR, we trained a CNN model [7], and for SNLI we trained an LSTM model [3].To generate response patterns we trained an ensemble of 1000 NNs to simulate an artificial crowd so that enough response patterns were obtained to fit the IRT models. For each model m_i, we sampled a subset of the training set, x^i_{train} and randomly added noise to the labels to vary the performance of the models. Training and test predictions were saved as the models' *response patterns*.

Algorithm 1 shows the steps taken to generate a single machine response patterns (RP). In order to generate a set of response patterns, Algorithm 1 is repeated some number of times (e.g. for our experiments, 1000 times). The *sample* function randomly samples a subset of X_{train} of size t. The *grade* function returns the sequence $Z_{mrp} = \{\mathbb{I}[\hat{y}_j = y_j]\}$

Algorithm 1 Machine response pattern generation

Input:
$X_{train}, Y_{train}, X_{test}, Y_{test}$,
NLP model M
Output:
Z_{mrp}: Response pattern for M after training set sampling and noise corruption

function generate_mrp($X_{train}, Y_{train}, X_{test}, Y_{test}, M$)
  $Z_{mrp} \leftarrow \emptyset$
  $Y^* \leftarrow set(Y_{train})$ // Set of possible labels
  $t \sim \mathcal{U}(0, length(X_{train}))$ // Sample a new training set size
  $X_{train}, Y_{train} \leftarrow sample(X_{train}, Y_{train}, t)$
  $\eta \sim \mathcal{U}(0, 0.5)$ // Sample a noise corruption level
  for $x_i, y_i \in X_{train}$ **do**
    $r \leftarrow \mathcal{U}(0, 1)$
    if $r \leq \eta$ **then**
      $y_i \sim Y^* \setminus y_i$ // Sample a new (incorrect) label
    else
      $y_i \leftarrow y_i$
    end if
  end for
  $M.fit(X_{train})$ // Train NLP model
  $\hat{Y}_{test} \leftarrow M.predict(X_{test})$
  $Z_{mrp} \leftarrow grade(\hat{Y}_{test}, Y_{test})$ // Grade output to generate model RP
  return Z_{mrp}
end function

References

[1] Frank B. Baker and Seock-Ho Kim. *Item Response Theory: Parameter Estimation Techniques, Second Edition*. CRC Press, July 2004.

[2] R Darrell Bock and Murray Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443--459, 1981.

[3] S Hochreiter and J Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735--1780, November 1997.

[4] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183--233, 1999.

[5] John P Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, 2018.

[6] John P. Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648--657. Association for Computational Linguistics, 2016.

[7] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[8] Prathiba Natesan, Ratna Nandakumar, Tom Minka, and Jonathan D Rubright. Bayesian prior choice in irt estimation using mcmc and variational bayes. *Frontiers in psychology*, 7:1422, 2016.

Results

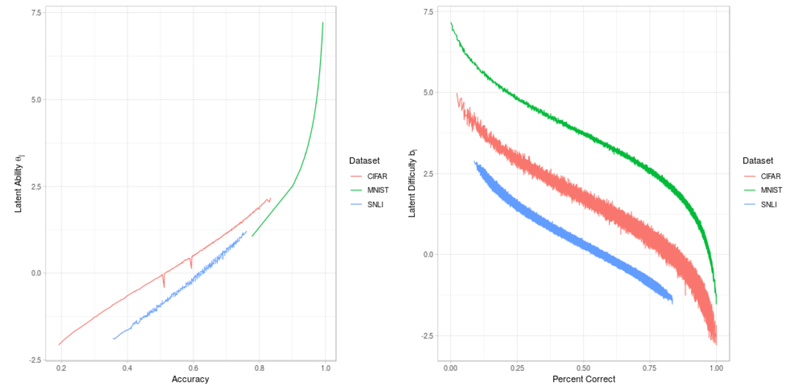

Figure: Plots comparing ability θ with accuracy (1a) and difficulty b with percent correct (1b).

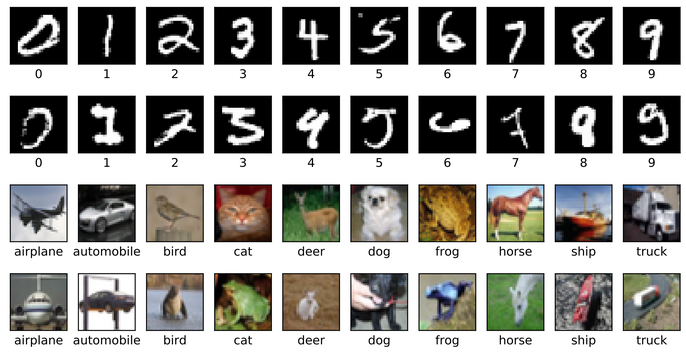

Figure: The easiest (first and third rows) and hardest (second and fourth rows) items in MNIST and CIFAR test sets.

Premise	Hypothesis	Label	Difficulty
Two men and a woman are inspecting the front tire of a bicycle.	There are a group of people near a bike.	Entailment	-3.675074
A girl in a newspaper hat with a bow is unwrapping an item.	The girl is going to find out what is under the wrapping paper.	Entailment	3.1443624
Two dogs playing in snow.	A cat sleeps on floor	Contradiction	-4.0138426
Man sweeping trash outside a large statue.	A man is on vacation.	Contradiction	3.7660716
People sitting in chairs with a row flags hanging over them.	A family reunion for Fourth of July	Neutral	-3.6031332
A group of dancers are performing.	The audience is silent.	Neutral	3.797989

Table: The easiest and hardest items for each class in the SNLI test set.