Informative Prediction based on Ordinal Questionnaire Data

Tsuyoshi Idé Amit Dhurandhar
IBM Research, T. J. Watson Research Center
{tide,adhuran}@us.ibm.com

***Abstract*—Supporting human decision making is a major goal of data mining. The more decision making is critical, the more interpretability is required in the predictive model. This paper proposes a new framework to build a fully interpretable predictive model for questionnaire data, while maintaining high prediction accuracy with regards to the final outcome. Such a model has applications in project risk assessment, in health care, in sentiment analysis and presumably in any real world application that relies on questionnaire data for informative and accurate prediction.**

Our framework is inspired by models in Item Response Theory (IRT), which were originally developed in psychometrics with applications to standardized tests such as SAT. We first extend these models, which are essentially unsupervised, to the supervised setting. We then derive a distance metric from the trained model to define the informativeness of individual question items. On real-world questionnaire data obtained from information technology projects, we demonstrate the power of this approach in terms of interpretability as well as predictability.

To the best of our knowledge, this is the first work that leverages the IRT framework to provide informative and accurate prediction on ordinal questionnaire data.

***Index Terms*—psychometrics, questionnaire data, item response theory, metric learning**

I. Introduction

Supporting human decision-making is one of the most important goals of data mining. In recommender systems for example, certain actions are recommended. Depending on the domain these actions could vary from being buying decisions [1] for shoppers to being important business decisions that are recommended to executives or managers based on historical data. Irrespective of the domain the final recommended action presented by itself is rarely sufficient to convince the decision maker of its "plausibility". Ordinarily, additional supporting evidence needs to be provided in support of the recommendation. Hence, a lower likelihood recommendation from a learning model may be a better choice if it can be clearly justified.

The plausibility, or more precisely *interpretability*, is in fact a critical success factor in many business applications. For example, imagine that you are a manager of a company and you are making decisions of lay-offs based on a scorecard for individual employees, which includes a number of qualitative questions such as "Has he/she made good enough contributions to teamwork?" You have a database of the historical records of best practices, which contains a collection of pairs $(\boldsymbol{x}, y)$, where $\boldsymbol{x}$ is a filled scorecard as represented by a binary or graded vector (see Fig. 1) and y is the binary indicator to represent termination ($y = 1$) or not ($y = 0$). Although the problem can be viewed formally as simple binary classification to predict y given $\boldsymbol{x}$, the nature of the problem is glaringly different in at least two aspects.

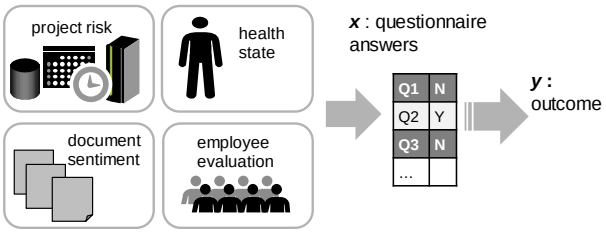


Fig. 1. Questionnaire-based diagnosis is ubiquitous. In many real applications, black-box predictive models are not practical. Full interpretability is often required at all instance-, dimension-, and ordinal grade-levels (see the text).

First, the input data are typically ordinal. In general it is not valid to naively use standard probabilistic assumptions such as the Gaussian-distributed noise for ordinal variables. Second, the model must have a high degree of interpretability. For the year-end assessment meeting, you as a manager will want to be very clear on the rationale of the suggested outcome from at least three perspectives:

1) Comparison to other employees: What is the difference between lay-off and no lay-off groups?
2) Comparison between different questions in the scorecard: What kind of weighting is used for individual questions? How can we justify the weighting?
3) Comparison between different question choices: Some questions may be easily achieved and others may not. How can we quantify the heterogeneity?

In other words, we need to ensure at least three different interpretabilities: instance-wise, dimension-wise, and ordinal-grade-wise interpretabilities. As long as a predictive model is used to support critical decision-making, the model must be fully interpretable in this sense. This is especially true in applications such as healthcare, project audit, and company reputation analysis, as illustrated in Fig. 1.

The goal of this paper is to introduce a new framework to build a fully interpretable predictive model for questionnaire data. Our method is inspired by the item response theory (IRT) [2], which was originally developed in psy-

chometrics with applications to standardized academic tests such as SAT [3]. As explained later, IRT provides a natural way to ensure dimension-wise (*i.e.* between individual question items) and grade-wise (*i.e.* between individual question choices of each question item) interpretabilities. However, the original IRT has two limitations when applying to our setting as in Fig. 1. First, the original IRT framework is unsupervised and does not incorporate the outcome variable y. Second, related to the first point, there is no direct method to evaluate the *informativeness of each question items* in terms of predictability of the outcome variable.

To address these limitations, we extend the original IRT to the supervised setting and incorporate it into a framework of distance metric learning [4]. Metric learning plays an important role to ensure full interpretability from two perspectives. First, as shown later, the learned Riemannian metric directly serves as the informativeness score, which is non-negative and bounded. Second, it provides us with a metric space, where different instances (*i.e.* scorecards in the employee evaluation example) are quantitatively compared through Riemannian distance, leading to instance-wise interpretability when combined with the k-nearest neighbor (k-NN) classifier. Note that the original ordinal data $\boldsymbol{x}$ is not defined in a metric space, and thus the distance between different instances is not well-defined.

To the best of our knowledge, this is the first work to

- propose a supervised extension of IRT, and
- propose an IRT-based metric learning framework for questionnaire data.

II. Related Work

There are four categories of previous work that is relevant to the present study.

The *first* category is obviously standard classification methods. As mentioned in Introduction, our task in Fig. 1 mirrors the task of binary classification. However, standard binary classifiers are not very useful in terms of analyzing the quality of the questions. For instance, support vector machines (SVMs) [5] or regularized logistic regression (LR) [5] may be accurate in predicting the outcome variable y, but the information they provide about the questions is in the form of unbounded signed weights, which can be difficult to interpret. On the other hand, in decision trees [5] it can be challenging to evaluate the importance of a variable as it might occur at different levels in different parts of the tree. Ensemble methods [5] may help compute variable importance, but they end up with losing instance-level information. We thus want a systematic way of evaluating the quality of the questions that is more informative and easier to digest, while maintaining predictability.

The *second* category is about modeling of human cognition. As Fig. 1 illustrates, we are interested in modeling the generative process of questionnaire answers. It amounts to modeling the decision-making process of humans, which is one of the typical examples of dynamics of complex systems. To model complex systems, deep learning has become a more and more practical tool in recent years, and dramatic successes in image and speech recognition [6], [7] are well-known. Also, if a fair amount of text data is given, sentiment analysis [8] for text documents provides a powerful method to understand the human cognition. Although we share a part of research motivation of modeling complex dynamics of human decision making, we pursue a completely opposite direction from those approaches that are mostly black-box: we request that our model should achieve interpretability at all different levels of instance, question item, and answer choices within each question. While some recent work addresses personal cognitive process in decision making [9], [10], which may be relevant to questionnaire analysis, our work differs in that we are interested in handling questionnaire data as the primary data source.

In psychometrics, on the other hand, quantitatively modeling human cognition bias has been an important topic for years. For a useful review, the reader may refer to Baker and Kim [11]. In the machine learning community, Lan *et al.* [12] recently extended the original IRT to incorporate factor analysis in an unsupervised setting. As explained in a later section, the original motivation of IRT was to quantitatively estimate the ability of examinees and the difficulty of individual question items in academic tests. If we are allowed to rephrase the ability as, *e.g.*, the medical risk in the case of diagnosis questionnaire, this is exactly what we want. Unlike the traditional setting of academic tests, however, we assume additional data of the final outcome such as occurrence of serious side effects, project failure, or termination of employment. In the context of the SAT test, in addition to the SAT scores themselves, we were as if we had information that the individual examinees had succeeded in their life later on. Using the information on the final outcome, we should be able to evaluate the true informativeness of the test items. To the best of the authors' knowledge, little attention has been paid to such a problem setting in psychometrics and data mining.

The *third* category is the study on ordinal data. Modeling ordinal data has been one of the major research topics in statistics and statistical data mining. Well-known examples include ordinal regression [13] and learning to rank [14]. Rank-constrained nonlinear discriminant analysis [15] is another recent instance. These assume that the response variable is ordinal. In our case, however, we are interested in handling ordinal predictor variables instead. From this perspective, the most relevant work will be Koren and Sill [16], [17], which addresses the ordinal nature of human rating in collaborative filtering, although their problem clearly differs from ours.

The *fourth* category is metric learning. Since the advent of the seminal paper of Xing [4], metric learning has been one of the most active areas in the data mining communities [18], [19], [20], [21]. For a recent review, the reader may refer to Bellet et al. [22]. By definition

of the task, metric learning (often implicitly) assumes that the samples distribute in a *metric space*, just like dots placed on a piece of paper, whose coordinates and the distance are well-defined and ready to be calculated using *e.g.* the Euclidean distance. However, it is clear that a special attention is required when handling ordinal variables since the ordinal scale distinguishes only relative goodness or badness. For example, an ordinal variable may ask about the goodness of personal relationship with your boss, and another ordinal variable may be the level of satisfaction to your family life. It is clear that relative comparison between two different ordinal variables is not trivial at all [23]. In spite of the popularity of metric learning research, only limited attention has been paid to metric learning for ordinal variables.

Recently, Ouyang and Gray proposed a rank-constrained approach to kernel learning. Also, Terada and Luxburg [24] proposed a method for order-preserving embedding. These works are somewhat relevant to ours, but their setting differs from ours in that they assume that the ordinal relationship between the instances is given; in our case, what is given is the final rating of projects, which is by no means sufficient to define the total order. Another relevant piece of work is ground metric learning [25], which handles the ordinal nature of the variables by considering histograms. However, their problem setting differs from us since we need to make a prediction for a single project, rather than for a histogram as a collection of projects.

Our framework for informative prediction attempts to achieve practical interpretability and predictability by combining a psychometric model with metric learning. To bridge the two, we use a particular form of probability density of neighborhood component analysis [21]. Instead of solving semi-definite programming as large-margin nearest neighbors [18], we take the path of Kostinger et al. [20], which first proposed an "optimization-free" method to metric learning and achieved a state-of-the art performance in the image classification task. As explained in a later section, we introduce an information-theoretic view to their approach.

III. Problem Definition and Motivation

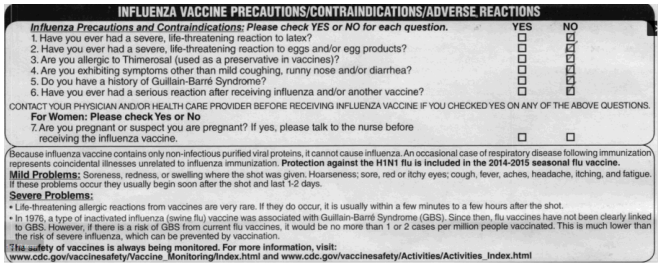
We first formally describe our problem setting. We then provide real world examples of where we encounter this setting thus showcasing its wide presence.

A. Problem Statement

Imagine we have a questionnaire containing M question items and N subjects (patients, projects, employees, etc.) take the questionnaire to answer the questions. Our training data set can be formally represented as

$$\mathcal{D} = \{(\boldsymbol{x}^{(n)}, y^{(n)}) \mid n = 1, 2, \ldots, N\}, \tag{1}$$

where $\boldsymbol{x}^{(n)}$ is an M-dimensional vector representing the questionnaire answers of the n-th subject, and $y^{(n)}$ is the class label for the n-th subject. Our goal is to build a fully interpretable model to predict y given a new $\boldsymbol{x}$, and to evaluate the informativeness of the individual question items, through which a qualitative feel to the user in terms of the predictability of the final outcome is provided.


INFLUENZA VACCINE PRECAUTIONS/CONTRAINDICATIONS/ADVERSE REACTIONS

Influenza Precautions and Contraindications: Please check YES or NO for each question.

	YES	NO
1. Have you ever had a severe, life-threatening reaction to latex?	☐	☑
2. Have you ever had a severe, life-threatening reaction to eggs and/or egg products?	☐	☑
3. Are you allergic to Thimerosal (used as a preservative in vaccines)?	☐	☑
4. Are you exhibiting symptoms other than mild coughing, runny nose and/or diarrhea?	☐	☑
5. Do you have a history of Guillain-Barré Syndrome?	☐	☑
6. Have you ever had a serious reaction after receiving influenza and/or another vaccine?	☐	☑

CONTACT YOUR PHYSICIAN AND/OR HEALTH CARE PROVIDER BEFORE RECEIVING INFLUENZA VACCINE IF YOU CHECKED YES ON ANY OF THE ABOVE QUESTIONS.

For Women: Please check Yes or No

7. Are you pregnant or suspect you are pregnant? If yes, please talk to the nurse before receiving the influenza vaccine.	☐	☐

Because influenza vaccine contains only non-infectious purified viral proteins, it cannot cause influenza. An occasional case of respiratory disease following immunization represents coincidental illnesses unrelated to influenza immunization. **Protection against the H1N1 flu is included in the 2014-2015 seasonal flu vaccine.**

Mild Problems: Soreness, redness, or swelling where the shot was given. Hoarseness; sore, red or itchy eyes; cough, fever, aches, headache, itching, and fatigue. If these problems occur they usually begin soon after the shot and last 1-2 days.

Severe Problems:

- Life-threatening allergic reactions from vaccines are very rare. If they do occur, it is usually within a few minutes to a few hours after the shot.
- In 1976, a type of inactivated influenza (swine flu) vaccine was associated with Guillain-Barré Syndrome (GBS). Since then, flu vaccines have not been clearly linked to GBS. However, if there is a risk of GBS from current flu vaccines, it would be no more than 1 or 2 cases per million people vaccinated. This is much lower than the risk of severe influenza, which can be prevented by vaccination.

The safety of vaccines is always being monitored. For more information, visit:
www.cdc.gov/vaccinesafety/Vaccine_Monitoring/Index.html and www.cdc.gov/vaccinesafety/Activities/Activities_Index.html

Fig. 2. Above is a brief snapshot of a 2014 flu shot vaccination questionnaire.

Here we say that a model is *fully interpretable* if a predictive model allows

- quantitative comparison between subjects in terms of their importance,
- quantitative comparison between question items in terms of their importance,
- quantitative comparison between answer choices in terms of probability of choosing each option,

while maintaining a comparable accuracy to other less interpretable methods.

B. Application Domains

We now instantiate the above problem definition to different domains.

Project Risk Assessment: In our motivating example, $\boldsymbol{x}^{(n)}$ is an M-dimensional vector representing the questionnaire answers, and $y^{(n)}$ is the health rating indicating the troubled or healthy status. Each of the dimensions of $\boldsymbol{x}^{(n)}$ takes an integer value in the predefined risk levels, while $y^{(n)}$ takes either of $+1$ or -1 (troubled or non-troubled). It is known that $\{\boldsymbol{x}^{(n)}\}$ has human bias which modulates the true risk levels of projects in some nonlinear fashion.

Health Care: Before administering any treatment doctors require patients to fill up (yes/no) questionnaires indicative of their condition. An example, flu shot questionnaire from last year is depicted in figure 2. Here the $\boldsymbol{x}^{(n)}$ are binary yes/no questions and $y^{(n)}$ indicates if the treatment was successful or not, *i.e.*, in our example if the person got flu or not.

Sentiment Analysis: In document sentiment analysis, pre-selected important keywords and/or expressions, which may be related to positive or negative sentiment, are often used to characterize a document. This can be thought of as questionnaire data, where the $\boldsymbol{x}^{(n)}$ indicate inclusion or non-inclusion of the keywords with different levels of the strength of sentiment. For example, if a word expressing

a negative sentiment appears, you could assign a value -1. If an expression expressing a positive feeling appears, you could assign a value $+1$. The $y^{(n)}$ here indicates if the overall document has a positive or negative connotation.

Employee Evaluations: See Introduction.

IV. Outcome-Aware Item Response Model

This section introduces a probabilistic framework towards informative prediction to meet the requirements explained in Subsection III-A.

A. Probabilistic Model for Answer Choice

To allow flexible modeling, we introduce a latent variable θ to represent the internal state of the system. The latent variable θ can be the ability of an examinee in the case of academic tests, the true failure tendency of an IT development project, the health state of a patient, etc., depending on applications. To be specific, let us take project risk management as a running example hereafter. In this case, the questionnaire questions are about project risks and are assumed to be bi-level [1], where the probability answering as at-risk should be a monotonically increasing function of θ. For the i-th question, one of the simplest models for such probability will be

$$P(\theta, a_i, b_i, c_i) \equiv c_i + \frac{1 - c_i}{1 + \mathrm{e}^{-a_i(\theta - b_i)}}, \tag{2}$$

whose functional form is given in Fig. 3. In psychometrics, this model is called three-parameter item response model [11], and the S-shaped curve in the figure is often called the item characteristic curve (ICC). Also, the parameters a_i, b_i, c_i are called the discrimination, difficulty, and guessing parameters, respectively. Literally, a_i represents the discriminablity or sensitivity to get, *e.g.*, the at-risk option and b_i represents the difficulty of the i-th question. The parameter c_i represents the probability of picking the at-risk option even when $\theta \to -\infty$, and thus corresponds to selection just by guess. Comparing ICCs across different question items, we can obtain vivid information about the individual question items.

By stacking M ICCs, we have the probability of an answer pattern $\boldsymbol{x}$, given θ and model parameters $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$, as

$$p(\boldsymbol{x}|\theta, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \prod_{i=1}^{M} P(\theta, a_i, b_i, c_i)^{\delta(x_i,1)} [1 - P(\theta, a_i, b_i, c_i)]^{\delta(x_i,0)} \tag{3}$$

where δ represents Kronecker's delta.

B. Prior Distribution to Latent State Variable θ

Although the original IRT assumes that the latent state variable θ follows a single prior, in the present setting, where the outcome variable y is available, it makes sense to consider a prior distribution conditioned on y:

$$f(\theta|y) = \begin{cases} \frac{\gamma}{\sqrt{2\pi}} \exp\left(-\frac{\gamma}{2}\theta^2\right) & \text{for} \quad y = -1, \\ \frac{\gamma}{\sqrt{2\pi}} \exp\left(-\frac{\gamma}{2}(\theta - \omega)^2\right) & \text{for} \quad y = +1, \end{cases} \tag{4}$$

where γ and ω are hyper-parameters to be learned from the training data. This is a natural extension of the original IRT, which assumes that all of the subjects form a single cluster around the value of zero.

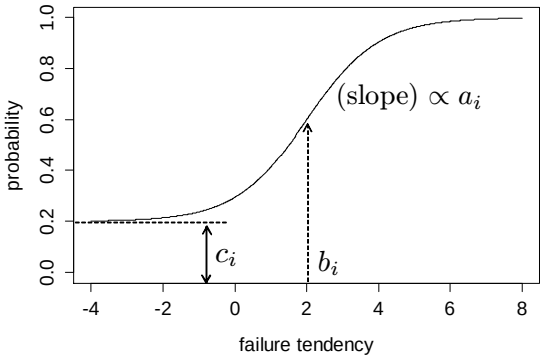


Fig. 3. Item Characteristic Curve for the example of project risk management.

C. Maximum Likelihood for Model Parameters, $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$

Now the log likelihood function of the model is written as follows:

$$L(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|\mathcal{D}) = \sum_{n=1}^{N} \ln \left[\pi(y^{(n)}) p(\boldsymbol{x}^{(n)}|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, y^{(n)})\right] \tag{5}$$

$$p(\boldsymbol{x}^{(n)}|\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, y^{(n)}) \equiv \int_{-\infty}^{\infty} \mathrm{d}\theta^{(n)}\ p(\boldsymbol{x}^{(n)}|\theta^{(n)}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})\ f(\theta^{(n)}|y^{(n)}) \tag{6}$$

where $\mathcal{D}$ symbolically represents the dependency on the training data, and $y^{(n)}$ is the variable representing the n-th project health indicator. The distribution $\pi(y^{(n)})$ is the prior distribution for $y^{(n)}$, which is assumed to be the same as the ratio of each of the labels to N.

In Fig. 4, we summarize the probabilistic model using the plate notation of probabilistic graphical models.

The model parameters $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$ can be found by maximizing the likelihood:

$$(\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*) = \arg\max_{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}} L(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|\mathcal{D}) \quad \text{subject to} \quad 0 \le c_i \le 1 \quad (i = 1, \ldots, M) \tag{7}$$

One well-known technical challenge in IRT is that the particular model of $P(\theta, a_i, b_i, c_i)$ does not have a conjugate prior, and thus the integration in (6) cannot be performed analytically. Fortunately, however, there is a very useful mathematical trick to perform integration. Specifically, using the orthogonality of Hermite polynomials, the

[1]For extensions for multi-level, see Sec. V-D.

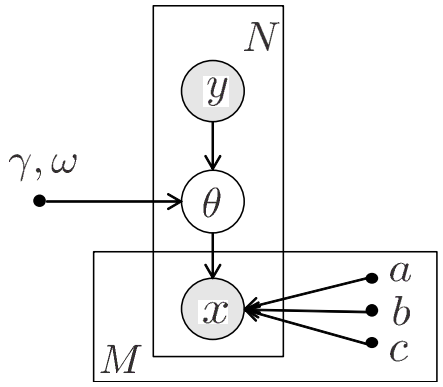


Fig. 4. Graphical model of outcome-aware IRT model.

following approximation holds [26]:

$$\int_{-\infty}^{\infty} \mathrm{d}\theta \; f(\theta|y) \; p(\boldsymbol{x}|\theta, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) \approx \sum_{i=1}^{N_h} w_i \; p\left(\boldsymbol{x} \left| \sqrt{\frac{2}{\gamma}}\theta_i + \omega\delta(y,1), \; \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\right.\right), \quad (8)$$

where practically good enough approximation is obtained by taking $N_h \approx 20$. The coefficients $\{w_i\}$ are defined by

$$w_i \equiv \frac{2^{N_h-1} N_h!}{N_h^2 [H_{N_h-1}(\theta_i)]^2},$$

and the position of break points $\{\theta_i\}$ is determined by the roots of the Hermite polynomial $H_{N_h}(\theta)$, which are tabulated [26]. The approximation (8) means that the integration is readily performed by performing summation over about 20 terms for arbitrary values of $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$. Thus the use of gradient method for solving the optimization problem (7) should not be a problem.

The box constraints on the guessing parameter $\{c_i\}$ are easily handled by the method of barrier function:

$$\tilde{L}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|\mathcal{D}) \equiv L(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|\mathcal{D}) + \mu_1 \sum_{i=1}^{M} \ln c_i + \mu_2 \sum_{i=1}^{M} \ln(1-c_i) \quad (9)$$

We simply solve the unconstrained optimization problem using the gradient method combined with line search for the step width [27]. Typically, the solution is not very sensitive to the choice of the coefficients μ_1 and μ_2. We set $\mu \equiv \mu_1 = \mu_2$, and determine the value by cross validation along with the hyper-parameter ω.

Algorithm 1 describes the major steps to learn the model parameters. We call it outcome-aware IRT (oIRT). The computational complexity is the same as the original IRT, which is NM^2 [11].

For the hyper-parameters γ, ω, one practical approach is as follows. We first assign initial values such as $(\gamma, \omega) = (1, 1)$, and perform optimization for $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$. Then we maximize the log-likelihood with respect to γ, ω given the $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$. Otherwise, for a given $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$, the hyper-parameter can be tuned to maximize another performance criteria such as the prediction accuracies.

Algorithm 1 Outcome-aware IRT algorithm.

Input: Training data $\mathcal{D}$. Hyper-parameters μ, ω. Initial values of the model parameters $\boldsymbol{a}^0, \boldsymbol{b}^0, \boldsymbol{c}^0$.
Output: The maximizer $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$.
repeat
  Compute the gradient of $\tilde{L}(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}|\mathcal{D})$ with respect to $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$.
  Determine the best step size η using the line search.
  Update the parameters as $\boldsymbol{a} \leftarrow \boldsymbol{a} - \eta \frac{\partial \tilde{L}}{\partial \boldsymbol{a}}$ and analogously $\boldsymbol{b}, \boldsymbol{c}$.
until Convergence
Return $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$.

V. Evaluating the Informativeness of Items

As discussed in the previous section, once $\{\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\}$ are determined, ICCs allow comparing different question items and different answer choices. The next question is how to leverage the (dis)similarity between subjects for further interpretability. For that purpose, in this section, we combine the oIRT with k-NN framework. To ensure a practical accuracy, we develop a new metric learning method and show that a learned Riemanian metric serves as the informativeness score.

A. Making Prediction

To maintain the interpretability on the predicted result, we use the k-NN method to predict y. The k-NN algorithm first finds k nearest neighbors from the training data. To capture complex dependencies in the space defined by ordinal variables, we use the Riemannian distance

$$d_{\mathsf{A}}^2(\boldsymbol{x}, \boldsymbol{x}') \equiv (\boldsymbol{x} - \boldsymbol{x}')^\top \mathsf{A} (\boldsymbol{x} - \boldsymbol{x}') \quad (10)$$

for the distance measure. The Riemannian metric A is optimally determined from $\mathcal{D}$ as explained later.

Then we compute the local probability distribution for y as:

$$p(y|\boldsymbol{x}, k) = \frac{1}{k} \sum_{n \in \mathcal{N}(\boldsymbol{x})} \delta(y, y^{(n)}) \quad (11)$$

where $\mathcal{N}(\boldsymbol{x})$ is the set of k nearest neighbor of $\boldsymbol{x}$. Based on If

$$\ln \frac{p(y=+1 \mid \boldsymbol{x}, k)}{p(y=-1 \mid \boldsymbol{x}, k)} \quad (12)$$

exceeds a certain threshold, the sample $\boldsymbol{x}$ is predicted as $y = +1$. The threshold is determined typically by leave-one-out (LOO) cross validation (CV).

Note that the use of k-NN is strongly motivated by practical applications. For example, in healthcare questionnaire analysis, finding similar subjects or patience is a part of doctors' daily routines. In project risk management, lessons

and learned from historical records is also an important part of the quality assurance process.

B. Deriving Distance Metric from oIRT

Our next step is to learn the metric tensor A from the oIRT model and interpret it as the informativeness. To establish the relationship between the metric and oIRT, we start with the probability distribution of neighborhood component analysis (NCA) [21] as

$$p_{\rm NCA}(\boldsymbol{x} \mid \boldsymbol{x}') = \frac{1}{Z_{\mathsf{A}}} \exp\left[-d^2_{\mathsf{A}}(\boldsymbol{x}, \boldsymbol{x}')\right],$$

where Z_{A} is the normalization constant. This distribution is defined in the neighborhood of $\boldsymbol{x}'$ and thus, in general, A depends on $\boldsymbol{x}'$, although we omitted the dependency from the notation for simplicity. Due to the finite range, Z_{A} can be a complex function of A unlike the Gaussian distribution.

The metric A governs local geometric structure in the $\boldsymbol{x}$-space. To associate it with oIRT, which is a probabilistic model, we consider the following equation

$$\left\langle \ln \frac{p(\boldsymbol{x} \mid \boldsymbol{\phi}^*, y=+1)}{p(\boldsymbol{x} \mid \boldsymbol{\phi}^*, y=-1)} \right\rangle = \langle -\ln p_{\rm NCA}(\boldsymbol{x} \mid \boldsymbol{x}') \rangle, \quad (13)$$

where $\boldsymbol{\phi}^*$ is a shorthand notation of $(\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)$, and $\langle \cdot \rangle$ represents the average with the empirical distribution over $\boldsymbol{x}$. The left hand side represents the Kullback-Leibler (KL) divergence between the distributions of different labels of the outcome variable, which measures the distance in the information-theoretical sense [28]. The right hand side is the entropy. The above equation means that the distance metric A is determined so that A compensates the KL divergence with the information $p_{\rm NCA}$ holds. The above equation can be called the entropy equation.

By inserting the definition of $p_{\rm NCA}$, we have

$$\ln Z_{\mathsf{A}} + {\rm Tr}(\mathsf{A}\Sigma_{\boldsymbol{x}'}) = \left\langle \ln \frac{p(\boldsymbol{x} \mid \boldsymbol{\phi}^*, y=+1)}{p(\boldsymbol{x} \mid \boldsymbol{\phi}^*, y=-1)} \right\rangle, \quad (14)$$

where $\Sigma_{\boldsymbol{x}'}$ is the local covariance matrix defined by

$$\Sigma_{\boldsymbol{x}'} \equiv \frac{1}{|\mathcal{N}(\boldsymbol{x}')|} \sum_{n \in \mathcal{N}(\boldsymbol{x}')} (\boldsymbol{x}^{(n)} - \boldsymbol{x}')(\boldsymbol{x}^{(n)} - \boldsymbol{x}')^\top.$$

The set of nearest neighbors of $\boldsymbol{x}'$ is denoted by $\mathcal{N}(\boldsymbol{x}')$, and its size is denoted by $|\cdot|$.

Although it is not straightforward to solve Eq. (14), we can derive an analytic approximated solution. First, to compute the right hand side, we exploit the conditional independence between different items in Eq. (3). Specifically, if we approximate the prior Eq. (4) as Dirac's delta functions, i.e., $\gamma \to \infty$, we have

$$\left\langle \ln \frac{p(\boldsymbol{x} \mid \boldsymbol{\phi}^*, y=+1)}{p(\boldsymbol{x} \mid \boldsymbol{\phi}^*, y=-1)} \right\rangle \approx \left\langle \ln \frac{p(\boldsymbol{x} \mid \theta_{+1}, \boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)}{p(\boldsymbol{x} \mid \theta_{-1}, \boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*)}, \right\rangle \quad (15)$$

where $\theta_{+1} = \omega$ and $\theta_{-1} = 0$. We assume that the hyperparameter ω is optimized in the way described in Sec. IV-C.

Considering the fact that, from the form of Eq. (3), the right hand side is represented as the summation over M terms, we put an additional constraint that A is diagonal. We readily see that the left hand side of Eq. (14) can be written

$$\sum_{i=1}^{M} \left\{ \ln Z_{A_{i,i}} + A_{i,i}[\Sigma_{\boldsymbol{x}'}]_{i,i} \right\}, \quad (16)$$

where

$$\ln Z_{A_{i,i}} \equiv \ln \int_{\boldsymbol{x} \in \mathcal{N}(\boldsymbol{x}')} {\rm d}x_i \ \exp\left[-A_{i,i}(x_i - x_i')^2\right].$$

This term changes much more slowly with respect to $\mathsf{A}_{i,i}$ than the second term of Eq. (16). By dropping this unimportant term and putting Eqs. (13), (15), and (16) together, we have

$$A_{i,j} = \frac{\delta_{i,j}}{[\Sigma_{\boldsymbol{x}'}]_{i,i}} \left\langle \ln \frac{p(x_i|\theta_{+1}, a_i^*, b_i^*, c_i^*)}{p(x_i|\theta_{-1}, a_i^*, b_i^*, c_i^*)}, \right\rangle \quad (17)$$

where $\delta_{i,j}$ Kronecker delta, and we defined

$$p(x_i|\theta, a_i^*, b_i^*, c_i^*) \equiv P(\theta, a_i^*, b_i^*, c_i^*)^{\delta(x_i,1)} \times [1 - P(\theta, a_i^*, b_i^*, c_i^*)]^{\delta(x_i,0)}. \quad (18)$$

To get the global metric, it makes sense to further approximate Eq. (17) by replacing the local variance $[\Sigma_{\boldsymbol{x}'}]_{i,i}$ with the global variance (i.e. based on all of the samples) of the i-th variable, σ_i^2, resulting in

$$A_{i,j} = \frac{\delta_{i,j}}{\sigma_i^2} \left\langle \ln \frac{p(x_i|\theta_{+1}, a_i^*, b_i^*, c_i^*)}{p(x_i|\theta_{-1}, a_i^*, b_i^*, c_i^*)} \right\rangle. \quad (19)$$

This is the equation that bridges oIRT and the Riemanian metric.

C. Informativeness Scores

We have derived the distance metric in Eq. (19), which bridges the oIRT model with the k-NN prediction. The interpretation of this equation is Eq. (19) is clear. Since A corresponds to the inverse of the covariance matrix in the Gaussian distribution, it is natural that $\mathsf{A}_{i,i}$ is proportional to σ_i^{-2}. The factor $\langle \cdot \rangle$ is the log-likelihood ratio between the two distributions corresponding to the two different classes. The role of the likelihood ratio in metric learning was first suggested by Kostinger et al. [20]. Also, as our approach, some authors explicitly use the KL divergence to derive information-theoretic metric [29], [30]. The expression Eq. (19) is new in the sense that it uses the particular form of oIRT model.

To get further insights from the result Eq. (19), we prove the following proposition that hold generally for a distribution of x_i conditioned on y_i.

Proposition 1: Consider a decision rule of classification for an instance x

$$y = +1, \quad \text{if} \ \ a(x) > 0$$
$$y = -1, \quad \text{if} \ \ a(x) \leq 0,$$

where

$$a(x) = \ln \frac{p(x \mid y=+1)}{p(x \mid y=-1)}. \quad (20)$$

We call the class $y=+1$ the minor class, and $y=-1$ the major class. This criterion (20) is optimal in the sense that it maximizes the minor sample accuracy while keeping the major sample accuracy constant.

Proof: To prove this proposition, by the definition of the major and minor sample accuracy, the optimal decision criterion a^* can be formally written as

$$a^* = \arg\max_a \int \mathrm{d}x \; I\left[a(x) \geq \tau_\alpha\right] p(x|y=1),$$

where α is a given major sample accuracy and $I[\cdot]$ is the indicator function. The constant τ_α satisfies

$$\int \mathrm{d}x \; I\left[a(x) \geq \tau\right] \; p(x|y=0) = 1-\alpha$$

Using a Lagrange multiplier λ, this problem can be rephrased as the minimization of $\Psi[a|\lambda]$ with respect to a:

$$\Psi[a|\lambda] = \int \mathrm{d}x \; I\left[a(x) \geq \tau_\alpha\right] \left\{p(x|y=1) - \lambda p(x|y=0)\right\}$$

To maximize the integral, the indicator function $I[\cdot]$ must be 1 wherever $\{\cdot\} > 0$. The condition is readily given as

$$a(x) = \frac{p(x|y=1)}{p(x|y=0)}, \;\; \lambda = \tau_\alpha \quad (21)$$

If we re-define a new criterion by transforming it using the logarithm function as $a(x)$, this coincides with Eq. (20). ■

Proposition 1 show that Eq. (19) is the product between the scaling factor (σ_i^{-2}) and the optimal classification criterion. Although Eq. (19) is an approximated solution, this means that by selecting the metric by Eq. (19), we should be able to achieve the maximum benefit in terms of classification accuracy (Note that Proposition also ensures the optimality of the criteria Eq. (12) via Bayes' theorem). Thus it is reasonable to take $\mathsf{A}_{i,i}$ as the definition of the *informativeness* of the i-th question item.

One issue in Eq. (19) as the informativeness is that it does not have a clear upper bound. Considering the fact that the KL distance measures the distance between distributions, we could use a different divergence measure such as Kolmogorov-Smirnov (KS) goodness-of-fit statistic:

$$A_{i,i} = |p(x_i=1|y=+1,\mathcal{D}) - p(x_i=1|y=-1,\mathcal{D})| \quad (22)$$

after standardizing the data to have unit variance. A major advantage of this choice is that it has a clear interpretation of the difference between probability values, and thus, it is bounded within [0,1].

The major steps in our questionnaire-based prediction approach are given in algorithm 2.

Algorithm 2 Questionnaire-based informative prediction.

Input: Training data $\mathcal{D}$. Hyper-parameters μ, ω. Initial values of the model parameters $\boldsymbol{a}^0, \boldsymbol{b}^0, \boldsymbol{c}^0$. The number of NNs, k.

Output: Informativeness scores $\{s_1, ..., s_M\}$ and the predicted label y for a new entity $\boldsymbol{x}$.

Determine the solution of oIRT as $\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*$ using algorithm 1.

Learn the distance metric A based on Eq. (19) or (22) and return $s_i = \mathsf{A}_{i,i}$.

For a new $\boldsymbol{x}$, find k NNs in $\mathcal{D}$ based on the A.

Compute $p(y|\boldsymbol{x}, k)$ for both $y=+1$ and $y=-1$ using Eq. (11).

Use Eq. (12) to decide a predicted label y.

Return y.

TABLE I
SUMMARY OF SYNTHETIC DATA.

$\boldsymbol{x}$	$(y=+1)$	$(y=-1)$
(0,0)	8	9
(0,1)	6	16
(1,0)	20	20
(1,1)	16	16

D. Discussion

So far we have assumed that question items are bi-level. To extend the framework to multi-level questions, one approach is to leverage so-called 1-of-K notation for the binary $\{x_i\}$. This amounts to decomposing each K-level question into K bi-level questions. This can be viewed as an approximation that disregards the order of the levels. To take account of the level order, one may use multi-graded IRT models [11], which is left to future work.

VI. EXPERIMENT

This section presents results of experimental evaluation of our metric learning framework based on oIRT. We first use a synthetic data set for illustration. Then we show results based on two real project review questionnaire data.

A. Illustration using synthetic data

To explain why informativeness matters, we randomly generated a questionnaire data of $M=2, N=100$ as summarized in Table I, where the numbers of generated samples are described. In this 2-dimensional setting, we have only four choices in $\boldsymbol{x}$. For comparison with the oIRT, we trained the L_1-regularized logistic regression (LR) [31], whose central model is given by

$$\ln \frac{p(y=+1 \mid \boldsymbol{x})}{p(y=-1 \mid \boldsymbol{x})} = \boldsymbol{\alpha}^\top \boldsymbol{x} + \beta.$$

The parameters $\boldsymbol{\alpha} \equiv (\alpha_1, \alpha_2)^\top$ and β are learned via maximum likelihood under an L_1 constraint on $\boldsymbol{\alpha}$. The regularization constant was optimized using LOO CV.

Learned coefficients of LR are shown in Fig. 5. As shown in the figure, α_2 takes a relatively large negative value,

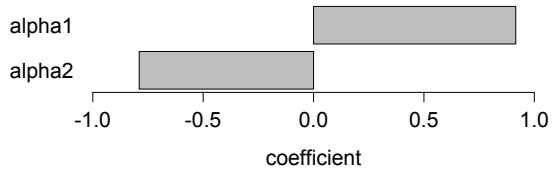


Fig. 5. Learned coefficients of regularized logistic regression for the synthetic data.

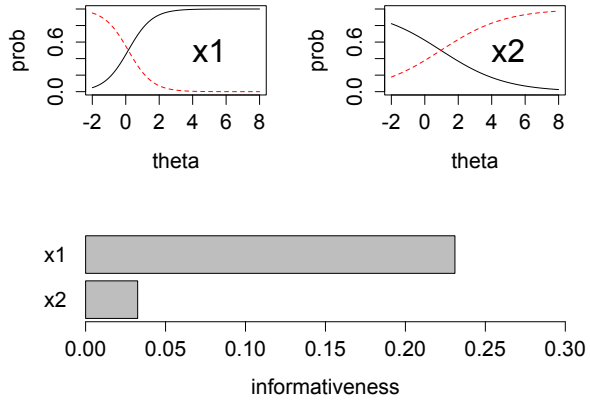


Fig. 6. Item characteristic curves and informativeness score for the synthetic data.

and almost the only conclusion we can draw would be something like "you cannot ignore either one".

Figure 6 shows results of oIRT, where we fixed $c_1 = c_2 = 0$ for simplicity. The informativeness score calculated by Eq.(19) clearly shows that x_1 is more important than x_2. This is confirmed by the ICC, where x_2 is less sensitive and even negatively depends on θ. If this is a diagnostic inquiry and a doctor is trying to infer the level of medical risk, the doctor may decide to use only the inquiry x_1 based on the ICCs to distinguish between low risk (small θ) and high risk (large θ) subjects. We see that decision-making becomes much easier with the aid of ICCs. Note that putting a stronger regularizer on LR and thus getting a sparser solution does not improve the situation because the LR coefficients still look like black-box metric that may take negative values.

B. Project Risk Assessment: Learning oIRT Model

Data set: We applied our method to real IT project assessment data called CRA (Contractual Risk Assessment) and PBA (Project Baseline Assessment). For both data sets, question items takes Y(at-risk) or N(no-risk), and each of filled questionnaires are associated with the label of project success ($y = -1$) or non-success ($y = +1$). For the data size, $(N, M) = (262, 22)$ and (1056,56) for CRA and PBA, respectively. The latter is perhaps the biggest project risk assessment data studied so far. For details of IT project risk management process and how the data is collected, see [32].

Informativeness: We calculated $\{\boldsymbol{a}^*, \boldsymbol{b}^*, \boldsymbol{c}^*, \}$ based on the CRA and PBA data. For the hyper-parameters ω and μ, we used the values of 1.85 and 1.0×10^{-6} for CRA, and 3.03 and 0.011 for PBA, respectively. These were determined by LOO CV to maximize the F-value defined later. For the initial values, we used $a_i = 1.0, b_i = 0.5, c_i = 10^{-5}$. To handle the imbalanced nature between troubled and healthy samples, we did bootstrap resampling for the non-success instances to obtain the same sample size in either class.

Figure 7 shows the oIRT parameters and the informativeness for CRA. We see that the 7th and 9th questions have major informativeness. Interestingly, these ones have negative discrimination parameters. This is due to the nature of risk management process. Since this risk assessment is done after completing all of risk mitigation actions, significant indication of risks that is readily visible to auditors cannot exist. Instead, some of them tend to be' 'unnaturally good", ending up with the negative $\{a_i\}$'s.

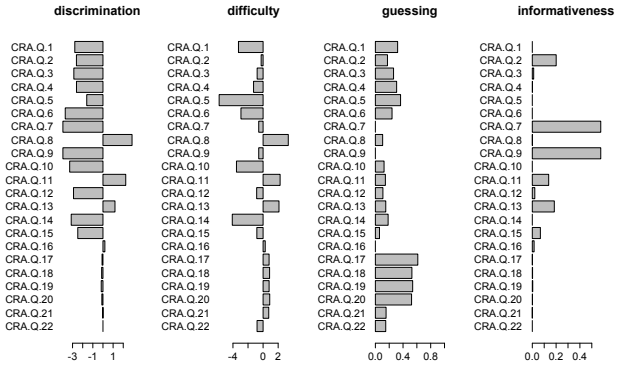


Fig. 7. IRT parameters learned for the CRA data.

ICCs: Figure 8 shows some examples ICCs. We drew $P(\theta, a_i, b_i, c_i)$ with the solid lines as well as $[1 - P(\theta, a_i, b_i, c_i)]$ with the dashed lines as before. We clearly see that the 10th question is hardly informative, being consistent to the negligible informativeness score in Fig. 7. Interestingly, this question is about future project plan after contract signing, which will be conducted by a different team from the one being reviewed. Thus negligible informativeness makes a lot of sense.

C. Project Risk Assessment: Predicting Project Failure

We compared prediction performance of our approach with alternatives. For performance criteria, we used

- r_1: Major sample prediction accuracy (accuracy in the healthy projects),
- r_2: Minor sample prediction accuracy (accuracy in the troubled projects),
- $f = 2r_1r_2/(r_1 + r_2)$: F-value. The harmonic mean between r_1 and r_2.

To compute these metrics, we used LOO CV. For example, to check hit or miss for the n-th sample in the training

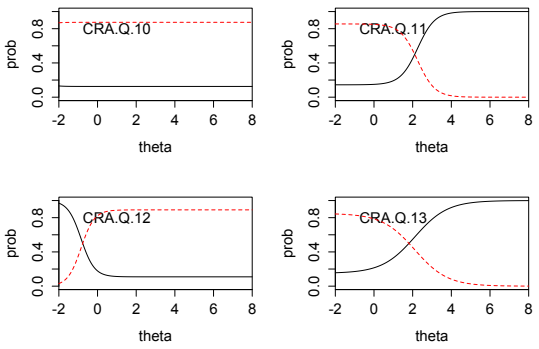


Fig. 8. Examples of ICCs for the CRA data.

data set $\mathcal{D}$, we held out the sample from $\mathcal{D}$, and learned the model from the remaining $N-1$ samples to make prediction for that sample. The number of NNs as well as the other hyper-parameters μ, ω were determined so as to maximize the F-value evaluated by LOO CV.

In addition to KL and KS-based metrics of Eqs. (19) and (22), we tested large-margin nearest neighbor (LMNN) [18], which is the standard baseline method in metric learning. In LMNN, the *full* Riemannian metric A is determined by minimizing the objective function

$$E(\mathsf{A}) = \frac{1-\mu}{N}\sum_{n=1}^{N}\sum_{j\in\mathcal{N}_n} d^2_{\mathsf{A}}(n,j) + \frac{\mu}{N}\sum_{n=1}^{N}\sum_{j\in\mathcal{N}_n}\sum_{l:\ y^{(l)}\neq y^{(n)}}\left[1+d^2_{\mathsf{A}}(n,j)-d^2_{\mathsf{A}}(n,l)\right]_+,$$

where $d^2_{\mathsf{A}}(n,j)$ is a shorthand notation of $d^2_{\mathsf{A}}(\boldsymbol{x}^{(n)},\boldsymbol{x}^{(j)})$, $\mathcal{N}_n$ represents the set of the nearest-neighbors of $\boldsymbol{x}^{(n)}$ chosen from the same label samples, *i.e.*, $y^{(j)} = y^{(n)}$, and $[h]_+ = \max\{0,h\}$ for $\forall h \in \mathcal{R}$. LMNN can be thought of as an improved version of NCA and is believed to be one of the best off-the-shelf metric learning methods [22], thanks mainly to the hinge loss function and its convex formulation [18].

We also tested L_1 regularized logistic regression (LR) [31] as a representative linear classifier, as well as the k-NN classifier with uniform weights (*i.e.* $A_{ii} = 1$).

Upon training the model, we did bootstrap resampling for the troubled projects to obtain the same sample size in either class. We optimized the classification threshold for Eq. (12) to achieve the best LOO CV F-value. The L_1 regularization constants of LR and the number of NN for k-NN of uniform weight was also determined using LOO CV.

Table II shows the comparison of the classification accuracies at the optimized parameter sets. We see that oIRT achieved better F-values than alternative approaches. The accuracy of KS is almost the same as KL. We see that the KS-based informativeness can be a reasonable choice in practice due to the good accuracy and its positive and bounded natures.

TABLE II
BEST CLASSIFICATION ACCURACIES.

	CRA			PBA		
	r_1	r_2	F	r_1	r_2	F
oIRT (KL)	0.814	0.733	0.771	0.612	0.701	0.653
oIRT (KS)	0.781	0.733	0.757	0.651	0.650	0.650
LMNN	0.757	0.733	0.745	0.550	0.693	0.614
LR	0.591	0.600	0.596	0.620	0.628	0.623
k-NN	0.648	0.533	0.585	0.585	0.647	0.566

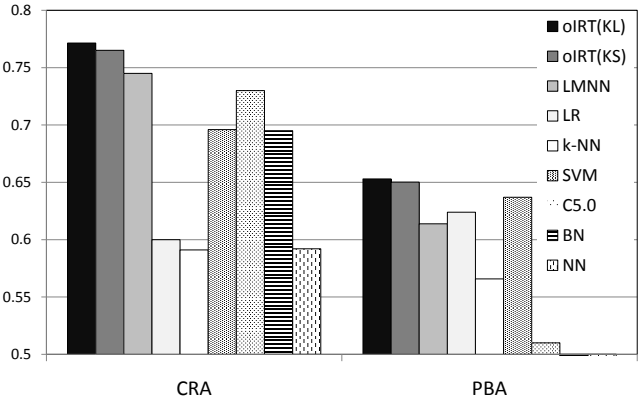


Fig. 9. Comparison of F-values (BN and NN are not visible for PBA).

It is very interesting to see that the simpler oIRT method using only the binary input and diagonal metric is competitive over *e.g.* LMNN, which optimizes the full Riemannian metric. This clearly suggests the importance of the nonlinear transformation by the logistic curve of IRT, and the risk of naively applying metric learning in non-metric spaces.

In Fig. 9, in addition to the alternatives shown in Table II, we further added commonly used classification techniques such as support vector machines (SVM), Neural Networks (NN), decision trees (C5.0) and Bayesian network (BN). All parameters for the models were chosen based on 10 fold cross validation. For SVM, we used RBF kernel. For BN, the structure was learned using tree augmented naive Bayes [33]. For NN, we used three hidden layers. All other parameters, and settings were fixed to the default ones of SPSS Modeler 15.0 including the numbers of nodes in NN. Again, for the both data set, oIRT achieves the best performance. LMNN, LR, and SVM were comparable and slightly worse than oIRT. For CRA, C5.0 achieved a good accuracy, but very bad for PBA. The accuracies for BN and NN were too low for PBA and the histograms are not visible in the graph. All these results clearly demonstrates that explicitly taking account of human cognition bias is critical in questionnaire data analysis. Our approach successfully captured the latent failure tendency with the aid of the psychoanalytical approach.

VII. Concluding remarks

We have addressed the task of informative prediction for questionnaire data. Our primary goal was to establish a method to quantitatively evaluate the informativeness of question items based on the predictability of the final outcome of individual samples.

To tackle the task, we introduced a new framework of outcome-aware item response theory (oIRT) by extending an existing theory in psychometrics. We have proposed two new ideas. One is to extend the prior distribution for the latent variable to include multiple states. The other is to define a Riemannian metric based on oIRT to improve the k-NN prediction. In spite of the simplicity of the algorithm, experiments using real questionnaire data showed that our method outperforms alternatives such as LMNN in terms of project failure prediction, while producing practical information on how individual question items work. Although the ordinal nature has not been seriously considered in many data mining tasks so far, this work demonstrates that adequately treating the ordinal nature is practically quite important.

References

[1] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, 2009.

[2] M. Wilson, *Constructing Measures*. Psychology Press, 2004.

[3] SAT, "Wikipedia; http://en.wikipedia.org/wiki/SAT."

[4] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Advances in neural information processing systems*, 2002, pp. 505–512.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* , 2nd ed. Springer, 2009.

[6] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 609–616.

[7] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[8] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, Jan. 2008.

[9] A. Borji and L. Itti, "Bayesian optimization explains human active search," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., 2013, pp. 55–63.

[10] T. Osogami and M. Otsuka, "Restricted boltzmann machines modeling human choice," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 73–81.

[11] F. B. Baker and S.-H. Kim, *Item Response Theory: Parameter Estimation Techniques* , 2nd ed. CRC Press, 2004.

[12] A. S. Lan, A. E. Waters, C. Studer, and R. G. Baraniuk, "Sparse factor analysis for learning and content analytics," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1959–2008, 2014.

[13] P. McCullagh, "Regression models for ordinal data," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 42, no. 2, pp. 109–142, 1980.

[14] O. Chapelle, Y. Chang, and T. Liu, Eds., *Proceedings of the Yahoo! Learning to Rank Challenge, held at ICML 2010, Haifa, Israel, June 25, 2010*, ser. JMLR Proceedings, vol. 14, 2011.

[15] B.-Y. Sun, J. Li, D. Wu, X.-M. Zhang, and W.-B. Li, "Kernel discriminant learning for ordinal regression," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 6, pp. 906–910, 2010.

[16] Y. Koren and J. Sill, "Collaborative filtering on ordinal user feedback," in *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, 2013, pp. 3022–3026.

[17] ——, "Ordrec: An ordinal model for predicting personalized item rating distributions," in *Proceedings of the Fifth ACM Conference on Recommender Systems*, ser. RecSys '11, 2011, pp. 117–124.

[18] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, 2009.

[19] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *Computer Vision, 2009 IEEE 12th International Conference on*, 2009, pp. 498–505.

[20] M. Kostinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2288–2295.

[21] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood component analysis," in *Advances in Neural Information Processing Systems, 17*, 2005, pp. 513–520.

[22] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," *ArXiv e-prints*, 2013.

[23] S. S. Stevens, "On the theory of scales of measurement," *Science*, vol. 103, no. 2684, pp. 677–680, 1946.

[24] Y. Terada and U. V. Luxburg, "Local ordinal embedding," in *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. JMLR Workshop and Conference Proceedings, 2014, pp. 847–855.

[25] M. Cuturi and D. Avis, "Ground metric learning," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 533–564, 2014.

[26] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, 3rd ed. New York, NY, USA: Cambridge University Press, 2007.

[27] W. Murray and M. H. Wright, "Line search procedures for the logarithmic barrier function," *SIAM Journal on Optimization*, vol. 4, no. 2, pp. 229–246., 1995.

[28] S. Amari and H. Nagaoka, *Methods of information geometry. Translation from the Japanese by Daishi Harada*, reprint of the 2000 edition ed. American Mathematical Society (AMS), 2008.

[29] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th International Conference on Machine Learning*, 2007, pp. 209–216.

[30] S. Wang and R. Jin, "An information geometry approach for distance metric learning," in *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS-09)*, vol. 5, 2009, pp. 591–598.

[31] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[32] K. Ratakonda, R. Williams, J. Bisceglia, R. Taylor, and J. Graham, "Identifying trouble patterns in complex it services engagements," *IBM Journal of Research and Development*, vol. 54, no. 2, pp. 5:1–5:9, March 2010.

[33] N. Friedman, D. Geiger, M. Goldszmidt, G. Provan, P. Langley, and P. Smyth, "Bayesian network classifiers," in *Machine Learning*, 1997, pp. 131–163.