IBM **Research**

Informative Prediction based on Ordinal Questionnaire Data

T. Idé (Ide-san) and Amit Dhurandhar
IBM T. J. Watson Research Center

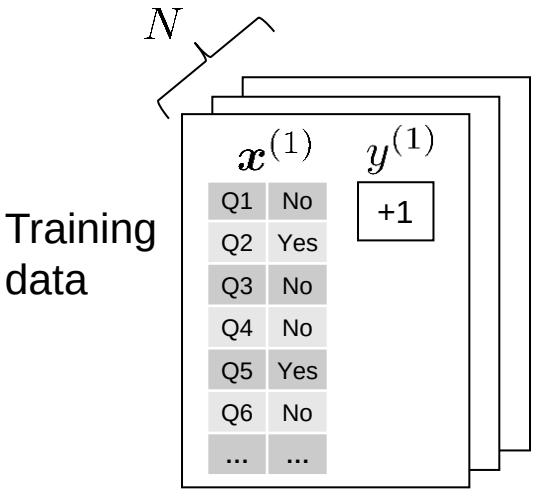
2015 International Conference on Data Mining (ICDM 2015)

Contents

- Problem setting
- Item response theory
- Maximum-a-posteriori framework for supervised IRT
- Metric learning from supervised IRT
- Experiments

General problem setting: Binary classification on *questionnaire* data

- Input: Questionnaire answer for *M* questions, $\boldsymbol{x} \in \{0,1\}^M$
- Output: class label, $y \in \{-1,+1\}$
- Data set: $(\boldsymbol{x}^{(n)}, y^{(n)})\ n = 1, \ldots, N$



$\boldsymbol{x} \in \{0,1\}^M$

M questions
- 1: at-risk
- 0: no risk

Prediction model

Q1	Yes
Q2	No
Q3	No
Q4	No
Q5	No
Q6	Yes
...	...

⇨ **?**

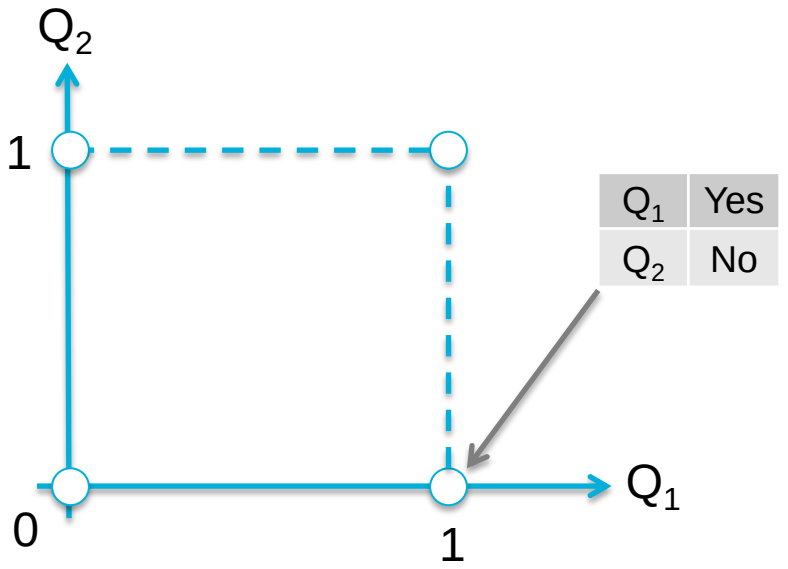
$y \in \{-1,+1\}$

Questionnaire data = ordinal data
We need special considerations

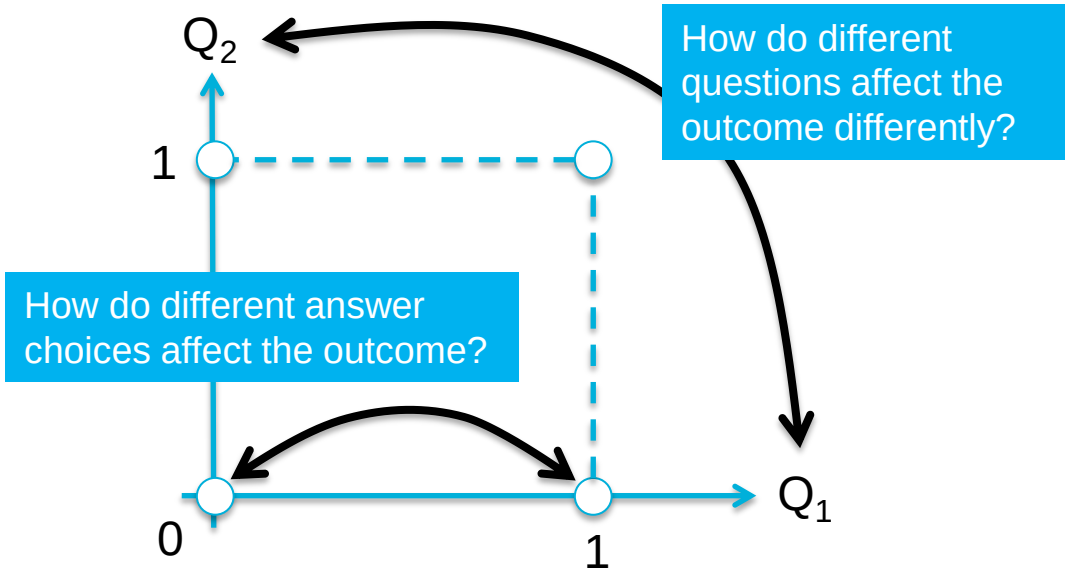
Questionnaire data = ordinal data
We need special considerations

No guarantee that the naïve notion of distance (e.g. Euclidean) holds for ordinal data

To define proper metric space



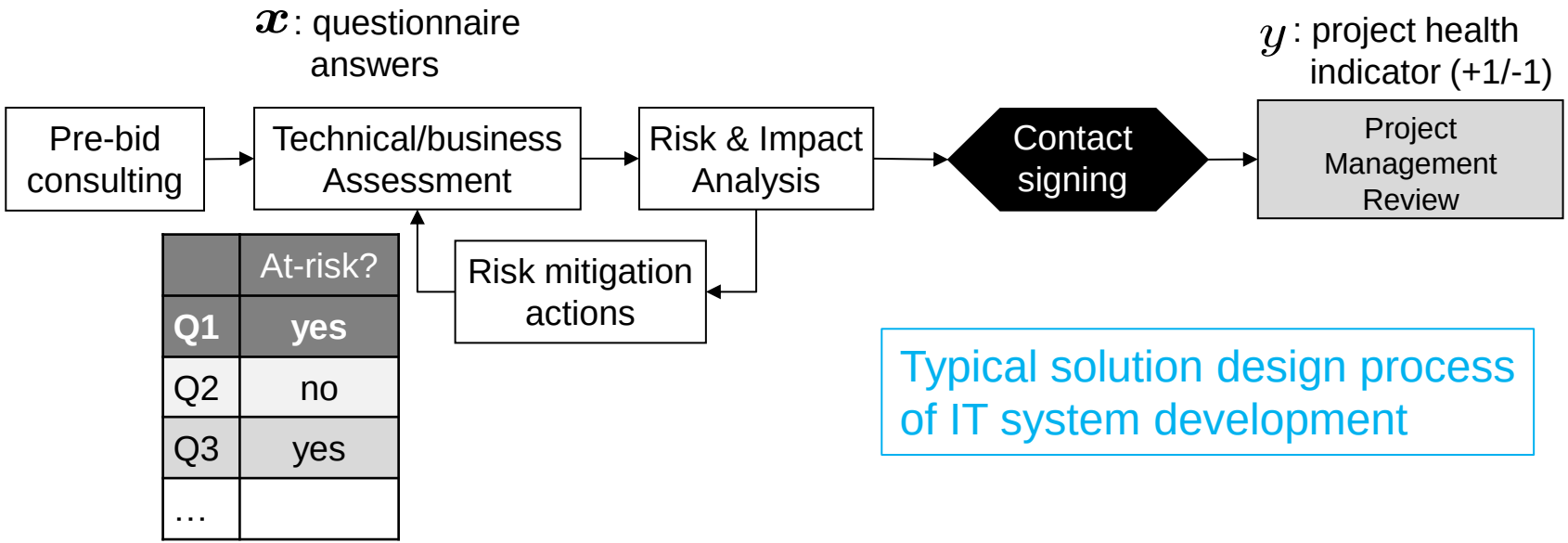
Questionnaire data = ordinal data
We need special considerations



To ensure full interpretability

Motivating real problem:
Predict how much likely a project is going to fail

- Input data, x : questionnaire answers by reviewer
- Outcome value, y : failure or success (after contract signing)

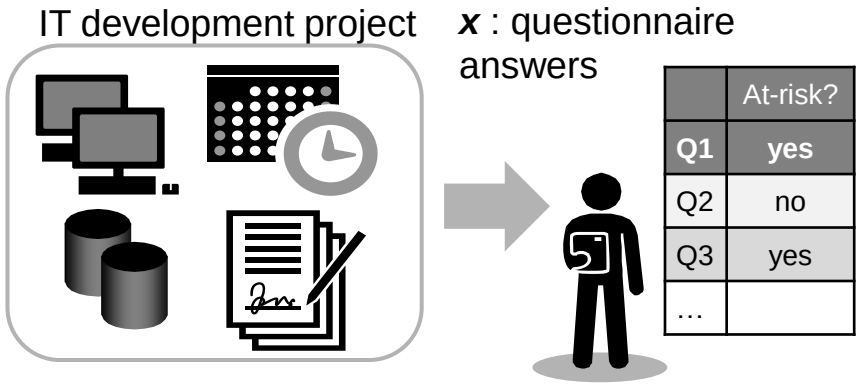


Typical solution design process of IT system development

(For ref.) What the questionnaire looks like

Major topics covered

- Communication issues with the client
- Well-definedness of the project scope
- Issues related to subcontractors and internal teams
- Project management issues
- etc.

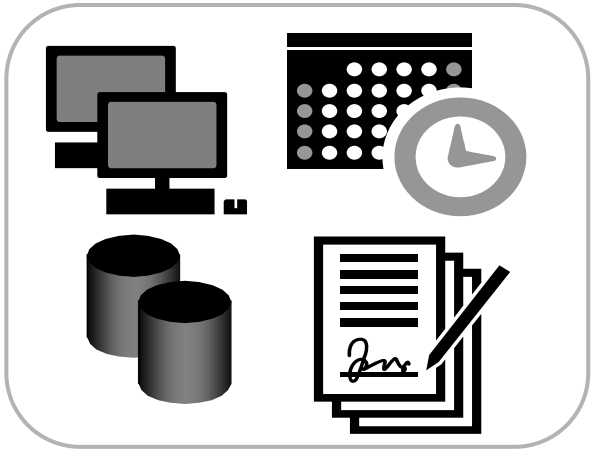


Problem setting summary

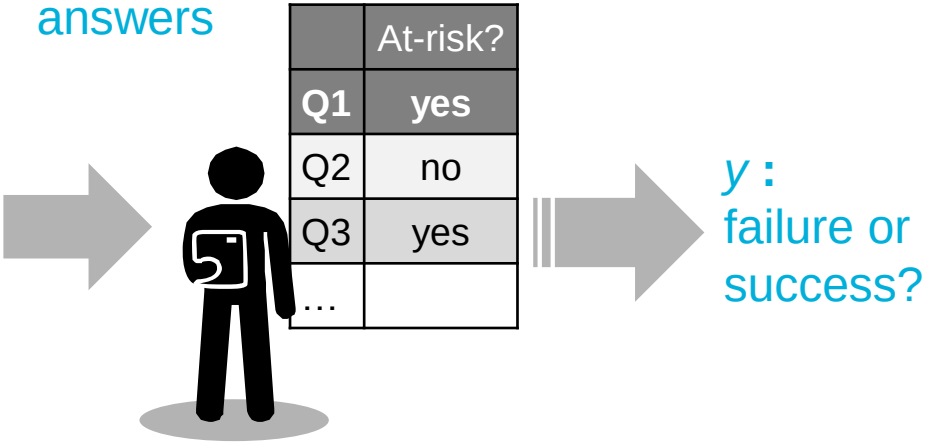
Wish to develop prediction model that is informative in terms of:

- Sample-sample difference
- Question-question difference
- Yes-no difference

IT system development project



x : questionnaire answers



y : failure or success?

Contents

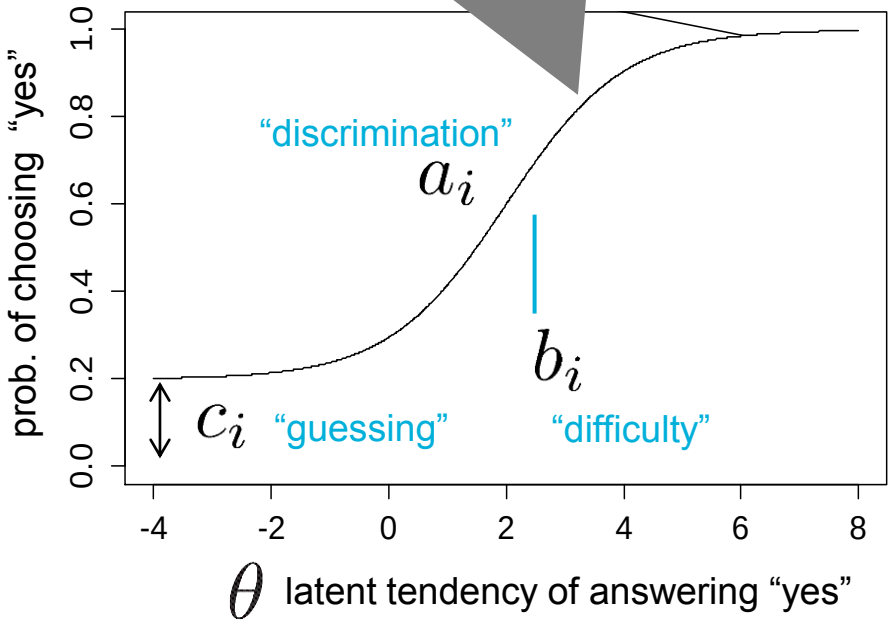
- Problem setting
- Item response theory
- Maximum-a-posteriori framework for supervised IRT
- Metric learning from supervised IRT
- Experiments

For natural interpretability, we employ the item response theory (IRT) in psychometrics

- Overly optimistic for smaller risks
- Overly pessimistic for larger risks
- Sometimes use a guess

Prob. of answering as “yes” for the *i*-th question

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

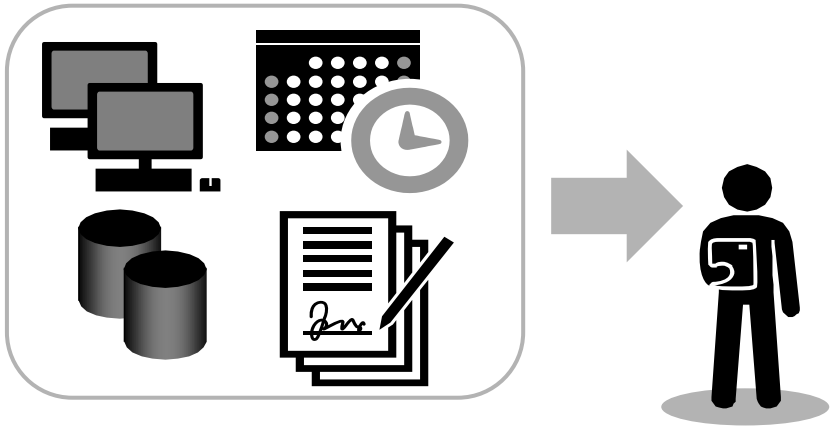


ICC naturally corrects cognitive biases of humans to uncover the true trait

latent failure tendency $\theta \rightarrow x$ questionnaire answer

Should be more faithful to the truth

May be biased



Generative model for a questionnaire answer x

Prob. of answering as “yes” for the *i*-th question

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

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

Contents

- Problem setting
- Item response theory
- Maximum-a-posteriori framework for supervised IRT
- Metric learning from supervised IRT
- Experiments

We extend the IRT to the supervised setting

Use the same ICC to take account of cognitive bias

Extend the original IRT in the **supervised** learning setting

- IRT is the standard method to analyze academic tests (e.g. SAT)
- IRT is unsupervised. We are developing a supervised version of IRT by including the outcome variable, y

We extend the IRT to the supervised setting by introducing a prior distribution conditional on *y*

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

Capture natural assumption that troubled projects should have higher failure tendency

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

Model parameters *a*,*b*,*c* are determined by maximizing log marginalized likelihood

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

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

$$(\boldsymbol{a}^*,\boldsymbol{b}^*,\boldsymbol{c}^*) = \arg\max_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}} L(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}|\mathcal{D})$$

$$\text{subject to} \quad 0 \le c_i \le 1 \quad (i = 1,\ldots,M)$$

Use numerical integration technique (Gauss-Hermite quadrature) to maximize marginalized likelihood

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

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

$$(\boldsymbol{a}^*,\boldsymbol{b}^*,\boldsymbol{c}^*) = \arg\max_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}} L(\boldsymbol{a},\boldsymbol{b},\boldsymbol{c}|\mathcal{D})$$

$$\text{subject to} \quad 0 \le c_i \le 1 \quad (i = 1,\ldots,M)$$

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

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.

See paper for details

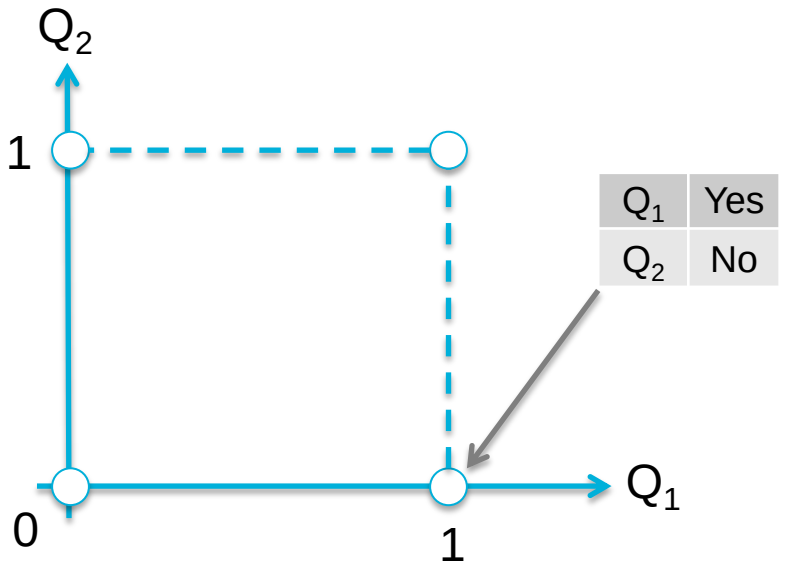
Contents

- Problem setting
- Item response theory
- Maximum-a-posteriori framework for supervised IRT
- Metric learning from supervised IRT
- Experiments

Questionnaire data = ordinal data
We need special considerations

No guarantee that the naïve notion of distance (e.g. Euclidean) holds for ordinal data

To define proper metric space



Once a metric space is properly defined, we can simply use k-NN for prediction

New question answer $\boldsymbol{x}$

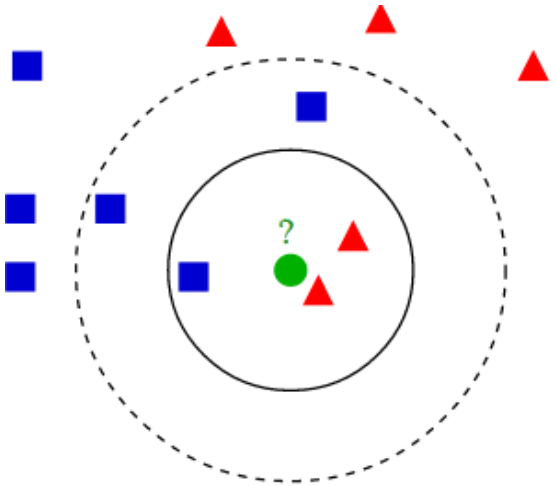
prediction phase

estimated binary label $\hat{y}$

Simply use k-NN classification based on the distance metric

$$d_{\mathrm{A}}(\boldsymbol{x}, \boldsymbol{x}^{(n)}) = (\boldsymbol{x} - \boldsymbol{x}^{(n)})^{\top} \mathrm{A} (\boldsymbol{x} - \boldsymbol{x}^{(n)})$$

How can we find *A* from the supervised IRT model?



Proposing entropy equation for learning the Riemannian metric *A*

- Intuition: “The deviation of *A* from the isotropic case is driven by the difference between the two classes (y=+1 or -1)”
  - Use the KL distance as a measure of difference
  - Use the p.d.f. of the neighborhood component analysis (NCA, [Goldberger 05])

$$p(\boldsymbol{x}|\boldsymbol{x}') \propto \exp\left[-d_{\mathrm{A}}(\boldsymbol{x},\boldsymbol{x}')^2\right]$$
$$d_{\mathrm{A}}(\boldsymbol{x},\boldsymbol{x}^{(n)}) = (\boldsymbol{x}-\boldsymbol{x}^{(n)})^{\top}\mathrm{A}(\boldsymbol{x}-\boldsymbol{x}^{(n)})$$

- The entropy equation to determine *A* (at ***x***’)

$$\left\langle \ln \frac{p(\cdot|\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},y=+1)}{p(\cdot|\boldsymbol{a},\boldsymbol{b},\boldsymbol{c},y=-1)} \right\rangle = \left\langle \ln \frac{p_{\mathrm{isotropic}}}{p_{\mathrm{A}}(\cdot|\boldsymbol{x}')} \right\rangle$$

Approximated solution of the entropy equation

$$A_{i,j} = \frac{\delta_{i,j}}{\sigma_i^2} \left\langle \ln \frac{p(x_i|\theta = \omega, a_i^*, b_i^*, c_i^*)}{p(x_i|\theta = 0, a_i^*, b_i^*, c_i^*)} \right\rangle$$

Diagonal element can be used as the **informativeness** of each question

- a_i^*, b_i^*, c_i^*: MAP solutions
- σ_i^2: Standard deviation of the i-th question
- $\langle \cdot \rangle$: Average over the empirical distribution

See paper for details

Summary of the model

Historical record

training phase

$$(\boldsymbol{x}^{(n)}, y^{(n)})$$
$$n = 1, \ldots, N$$

Generative model of ***x*** : IRT

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

Prior for supervised extension

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

MAP estimation

$$\hat{\boldsymbol{a}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{c}}$$

Metric learning to give the distance metric, ***A***, for *k*-NN

Contents

- Problem setting
- Item response theory
- Maximum-a-posteriori framework for supervised IRT
- Metric learning from supervised IRT
- Experiments

Toy example: binary classification for bi-variate binary inputs

- Compared with regularized logistic regression
- Took the diagonal metric as informativeness of each variable
- Proposed method gives much richer and more informative results

TABLE I
SUMMARY OF SYNTHETIC DATA.

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

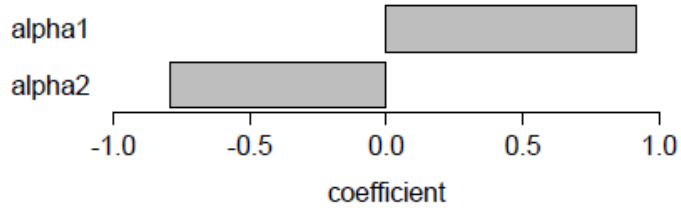


Fig. 5. Learned coefficients of regularized logistic regression for the synthetic data.

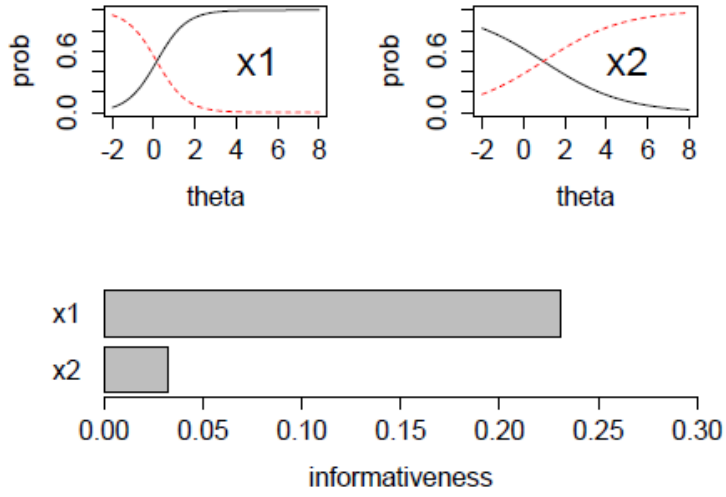
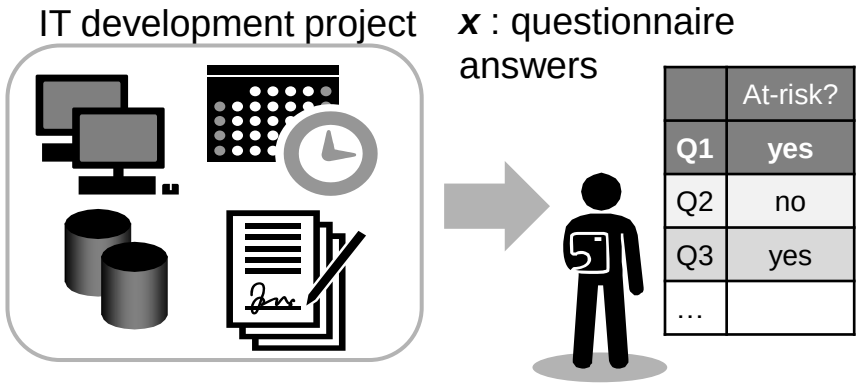


Fig. 6. Item characteristic curves and informativeness score for the synthetic data.

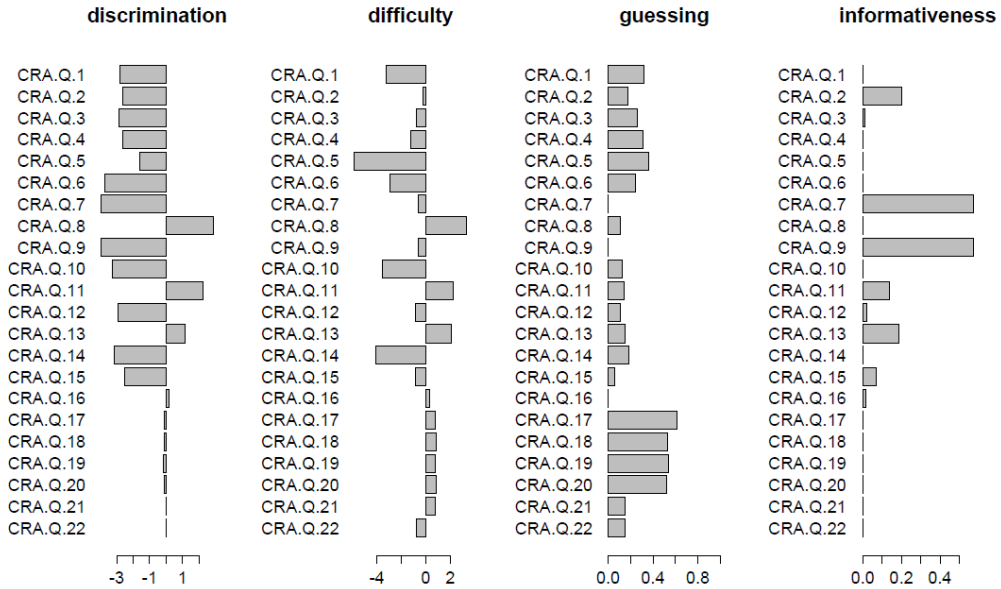
Experiment: Using service provider's real project assessment data

- Two types of project assessment data
  - CRA (contract risk assessment)
  - PBA (project baseline assessment)

- Data size
  - CRA: *M* = 22, *N*=262
  - PBA: *M* = 56, *N*=1056



Result (1): Estimated IRT parameters providing practical information on the usefulness of each question



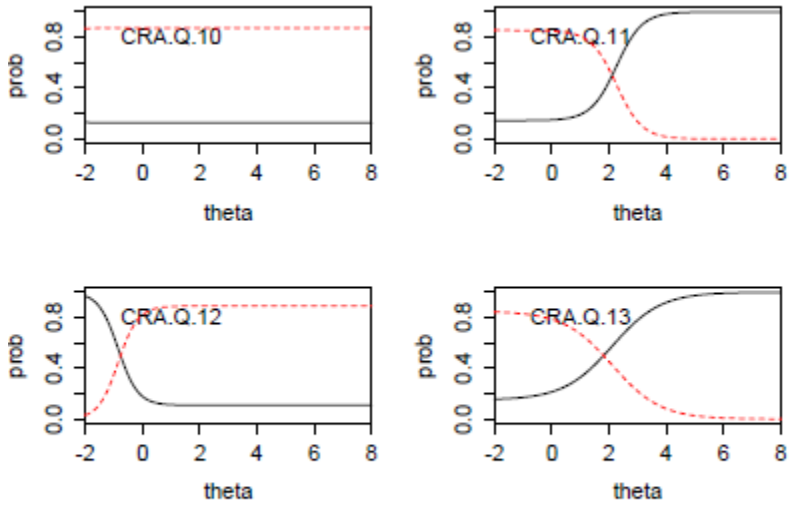


Fig. 8. Examples of ICCs for the CRA data.

Result (2): Achieved comparable or even better accuracy

- Performance metric: F-value
  - harmonic mean between troubled project accuracy and non-troubled project accuracy
- Outperformed baseline
  - Max margin nearest neighbors
  - Logistic regression
  - Simple k-NN
  - SVM
  - Decision tree (C5.0)
  - Neural network

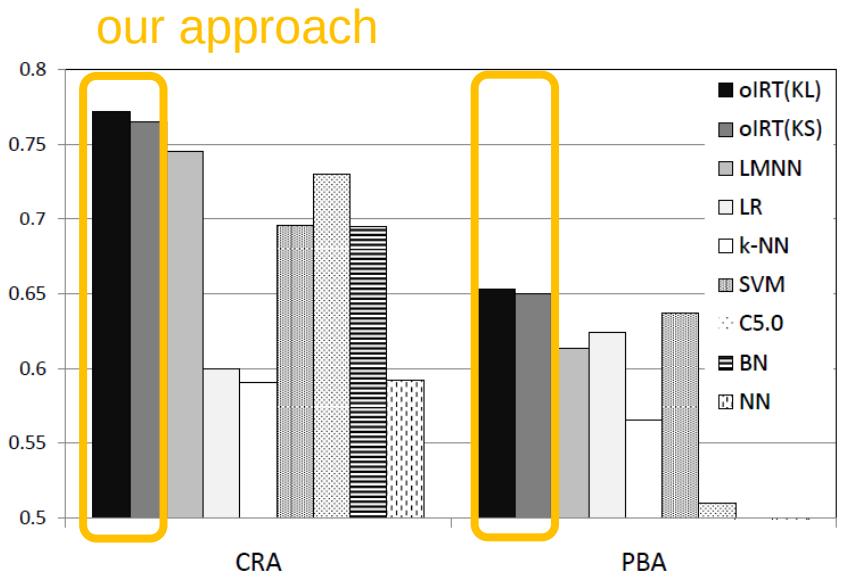


Fig. 9. Comparison of F-values (BN and NN are not visible for PBA).

Conclusion

- Proposed a shallow but fully-interpretable prediction method for questionnaire data
- Extended IRT to the supervised setting
- Proposed a new metric learning criteria of entropy equation to define a proper metric space
- Applied the method to real project review data to show practical utility

This deck of slides will be uploaded on my Web site (ide-research.net)

Thank you!

Another application: employee evaluation

- Input data, x : questionnaire answers on employee's performance
  - Questions are like "Has he/she made good enough contributions to teamwork?"
  - Managers put evaluation on individual questions
- Outcome value, y : termination or not

	Good?
Q1	**yes**
Q2	no
Q3	yes
…	

y **:**
stay or leave?

For natural interpretability, we employ the item response theory (IRT) in psychometrics