Graded Multilabel Classification: The Ordinal Case

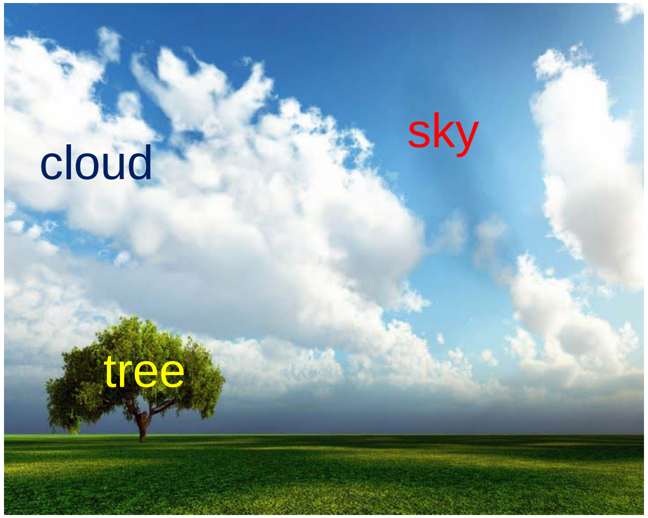
Weiwei Cheng, Krzysztof Dembczynski, Eyke Hüllermeier

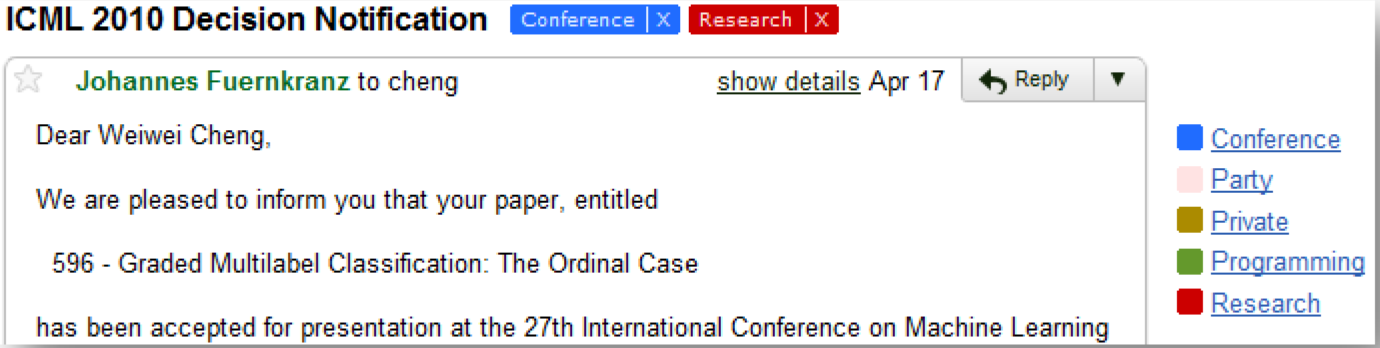
Knowledge Engineering & Bioinformatics Lab
Department of Mathematics and Computer Science
University of Marburg, Germany

Philipps Universität Marburg

ICML 2010, Haifa, Jun.24.2010

Multilabel Classification




ICML 2010 Decision Notification Conference X Research X

Johannes Fuernkranz to cheng show details Apr 17 Reply

Dear Weiwei Cheng,

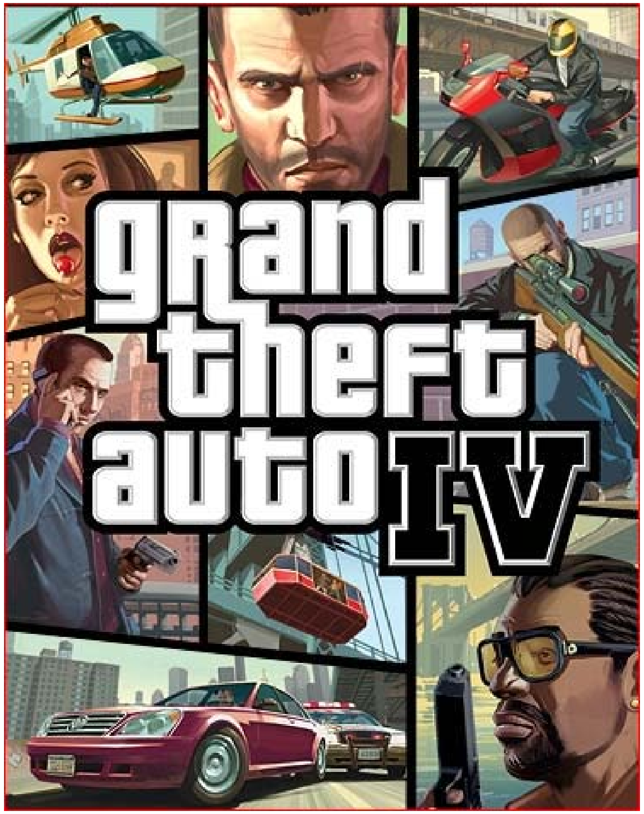
We are pleased to inform you that your paper, entitled

596 - Graded Multilabel Classification: The Ordinal Case

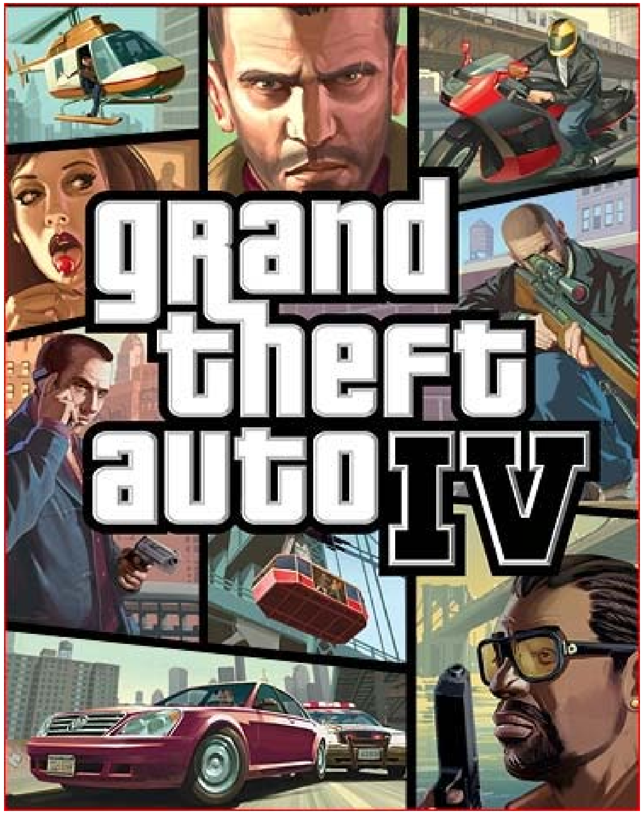
has been accepted for presentation at the 27th International Conference on Machine Learning

Conference
Party
Private
Programming
Research

Grand Theft Auto



Grand Theft Auto



Shooting

★★★ *completely*

Racing

★★☆ *almost*

Fighting

★☆☆ *somewhat*

Role-playing

☆☆☆ *not at all*

Another Example

VIDEOLECTURES.NET

EXCHANGE IDEAS / SHARE KNOWLEDGE

Combining Instance-Based Learning and Logistic Regression for Multi-Label Classification

as author at Sessions,
74 views

Lecture rating

People found this lecture:

Worth seeing

because it is:

Valuable and informative

Well presented

Easily understandable

Acceptably recorded

Graded Multilabel Classification

- Instance $\mathbf{x} \in \mathcal{X}$ can belong to each class $\lambda \in \mathcal{L}$ to a certain degree ($\longrightarrow$ idea of graded class membership in the spirit of fuzzy set theory)

- A graded multilabel classifier is a mapping $\mathcal{X} \longrightarrow M$, where M is a set of graded membership degrees, belonging to [0,1] (instead of {0,1}).

- Often, an ordinal scale of membership degrees is convenient, i.e. $M = \{m_0, m_1, \ldots, m_k\}$ with

$$0 = m_0 < m_1 < \ldots < m_k = 1.$$

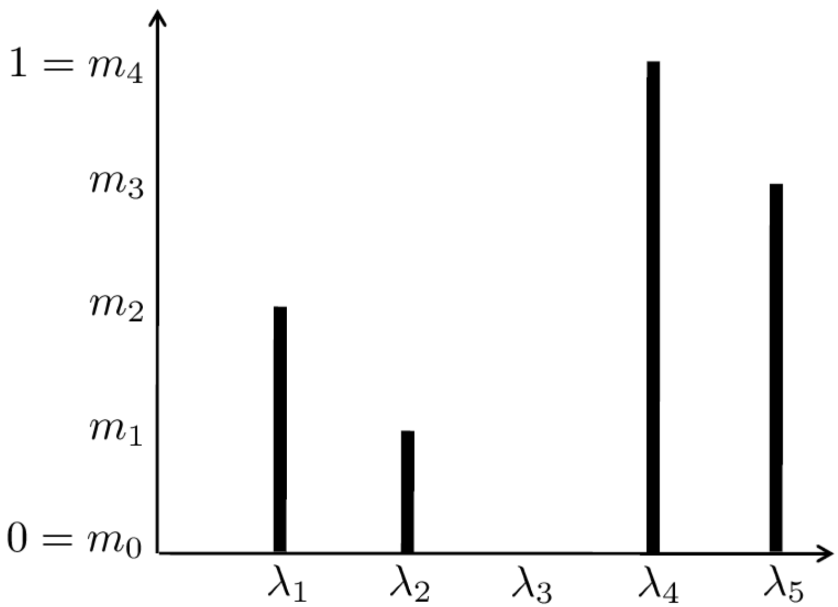
Graded Multilabel Classification

The general idea of reduction in machine learning:
Reduce a complex problem to one or several simpler problems, preferably those for which good algorithms already exist.

We propose two reduction schemes for graded multilabel classification:

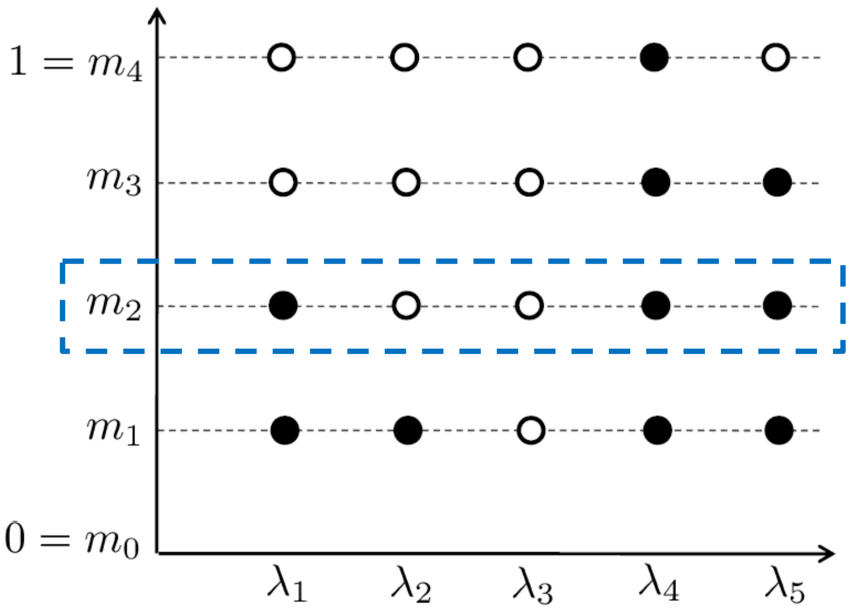
- vertical reduction leads to solving $|L|$ ordinal classification problems
- horizontal reduction leads to solving $|M|$ standard multilabel classification problems

Vertical Reduction



- Induce one classifier $h_i : \mathcal{X} \longrightarrow M$ for each label λ_i.
- h_i is solving an ordinal classification problem.
- Overall, we are solving |L| such problems.
- The simplest approach is “graded relevance”, however, to take dependencies between labels into account, these problems should not be solved independently of each other.

Horizontal Reduction

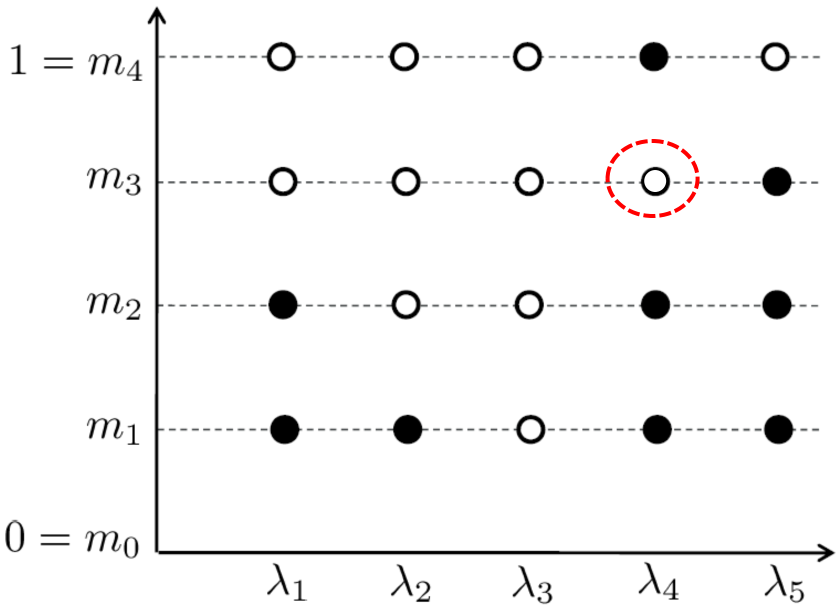


- M can be represented "horizontally" in terms of its level-cuts, e.g., $\boxed{[L_{\mathbf{x}}]_{m_2} = \{\lambda_1, \lambda_4, \lambda_5\}.}$
  $\longrightarrow$ *problems obtained by thresholding on the membership scale*

- For each level $\alpha \in \{m_1, m_2, \ldots, m_k\}$, learn the mapping
  $$h^{(\alpha)} : \mathcal{X} \longrightarrow 2^M, \mathbf{x} \mapsto [\mathcal{L}]_\alpha .$$

- Overall, we are solving k standard multilabel classification problems.

Horizontal Reduction

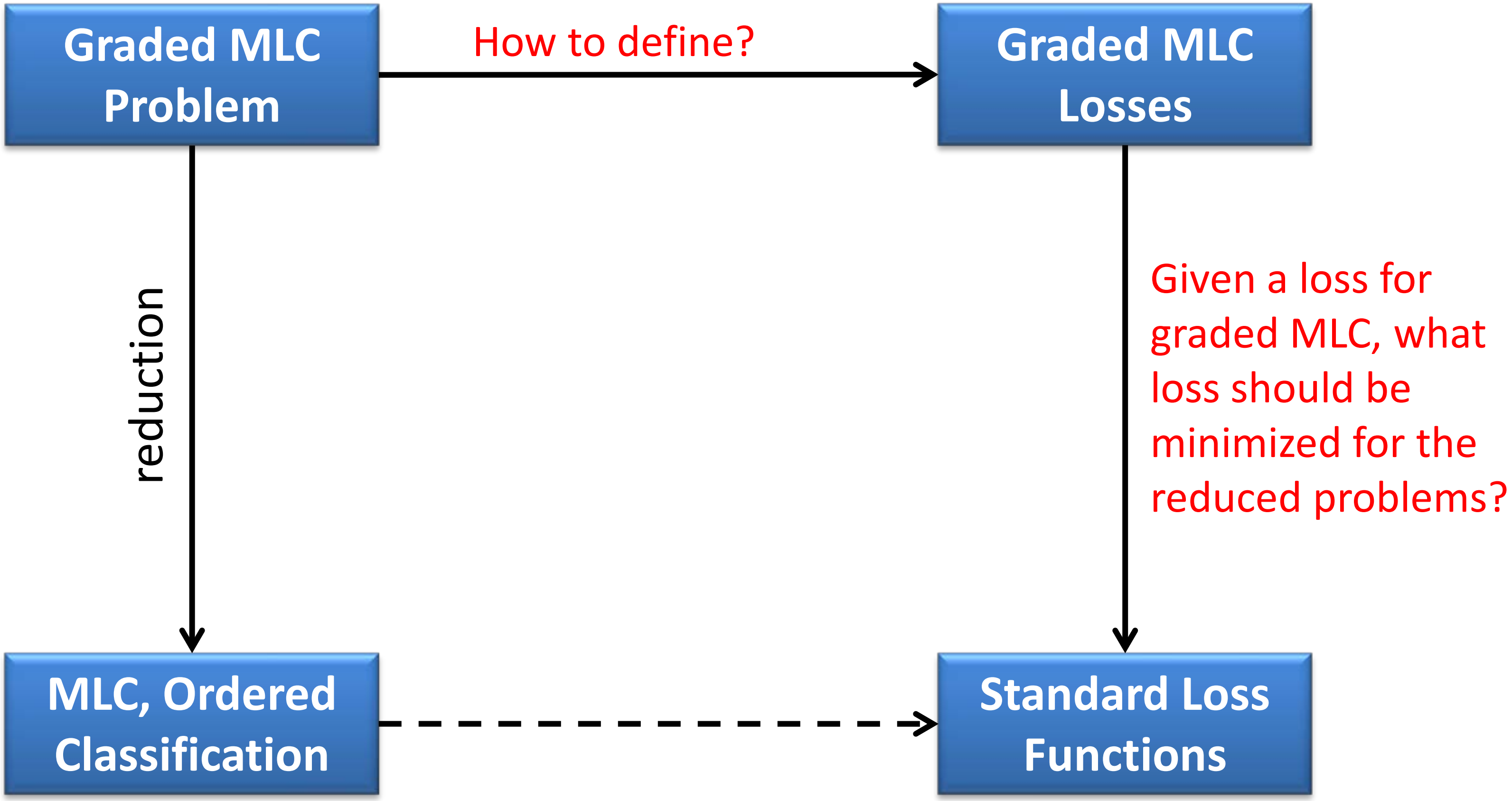


- Predictions should be consistent in the sense that $\left(h^{(m_i)}(\mathbf{x}) = 1\right) \Rightarrow \left(h^{(m_i - 1)}(\mathbf{x}) = 1\right)$.

Non-trivial!

Once $h^{(m_1)}, \ldots, h^{(m_k)}$ are trained consistently, predictions are recovered by $h(\mathbf{x})(\lambda) = \max\left\{m_i \in M \mid \lambda \in h^{(m_i)}(\mathbf{x})\right\}$.

Loss functions



Example: Hamming Loss

$$E_H(h(\mathbf{x}), L_{\mathbf{x}}) = \frac{1}{|\mathcal{L}|} |h(\mathbf{x}) \Delta L_{\mathbf{x}}| = \frac{1}{|\mathcal{L}|} \sum_{i=1}^{|\mathcal{L}|} \begin{cases} 0 & h(\mathbf{x})(\lambda_i) = L_{\mathbf{x}}(\lambda_i) \\ 1 & h(\mathbf{x})(\lambda_i) \neq L_{\mathbf{x}}(\lambda_i) \end{cases}$$

Hamming loss = average over *label-wise* losses

Label-wise loss in the graded (ordinal) case?

- Standard 0/1 loss:

$$E_{0/1}(m_i, m_j) = \begin{cases} 0 & m_i = m_j \\ 1 & m_i \neq m_j \end{cases}$$

- Absolute error:

$$AE(m_i, m_j) = |i - j|$$

Example: Hamming Loss

This leads to two variants:

$$E_H(h(\mathbf{x}), L_\mathbf{x}) = \frac{1}{|\mathcal{L}|} \sum_{i=1}^{|\mathcal{L}|} E_{0/1}(h(\mathbf{x})(\lambda_i), L_\mathbf{x}(\lambda_i))$$

$$E_H(h(\mathbf{x}), L_\mathbf{x}) = \frac{1}{|\mathcal{L}|} \sum_{i=1}^{|\mathcal{L}|} AE(h(\mathbf{x})(\lambda_i), L_\mathbf{x}(\lambda_i))$$

This is already a “vertical” expression of the GMLC loss, i.e., an expression of the form

$$\mathrm{A}\left(\{\, l\,(H(\mathbf{x})(\lambda_i), L_\mathbf{x}(\lambda_i))\}_{i=1}^{|\mathcal{L}|}\right)$$

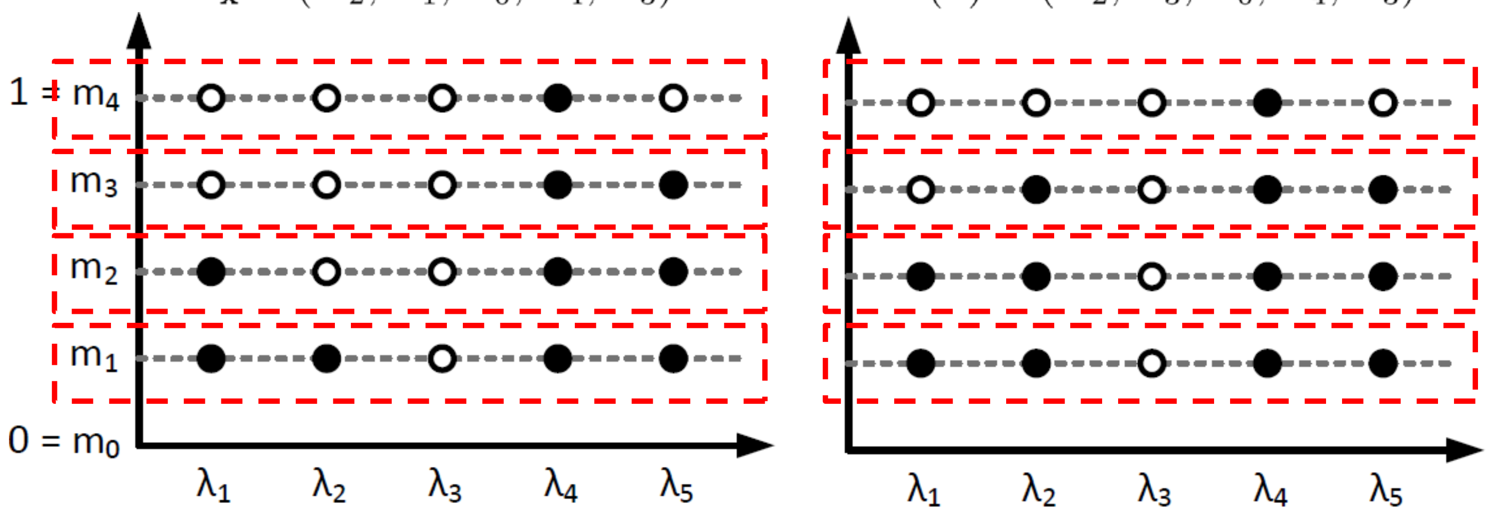
A: aggregation operator
$l(\cdot)$: loss defined on $\mathcal{L}$

Example: Hamming Loss

Is there also a “horizontal” expression of the following form?

$$\mathrm{A}\left(\{\, L\left([H(\mathbf{x})]_{m_i}, [L_{\mathbf{x}}]_{m_i}\right)\}_{i=1}^{k}\right)$$

$L_{\mathbf{x}} = (m_2, m_1, m_0, m_4, m_3)$ $\qquad$ $H(\mathbf{x}) = (m_2, m_3, m_0, m_4, m_3)$



Loss functions

This is indeed the case for *absolute error*, since

$$\sum_{i=1}^{|\mathcal{L}|} \mathrm{AE}(h(\mathbf{x})(\lambda_i), L_{\mathbf{x}}(\lambda_i)) = \sum_{i=1}^{k} \underbrace{|[h(\mathbf{x})]_{m_i} \Delta [L_{\mathbf{x}}]_{m_i}|}_{\text{standard Hamming loss for the i-th MLC reduction}}$$

$$E_H(h(\mathbf{x}), L_{\mathbf{x}}) = \frac{1}{|\mathcal{L}|} \sum_{i=1}^{|\mathcal{L}|} E_{0/1}(h(\mathbf{x})(\lambda_i), L_{\mathbf{x}}(\lambda_i))$$

For *0-1 loss*, there is **no** representation of that kind (there does not exist an aggregation/loss pair (A,L) with the desired properties).

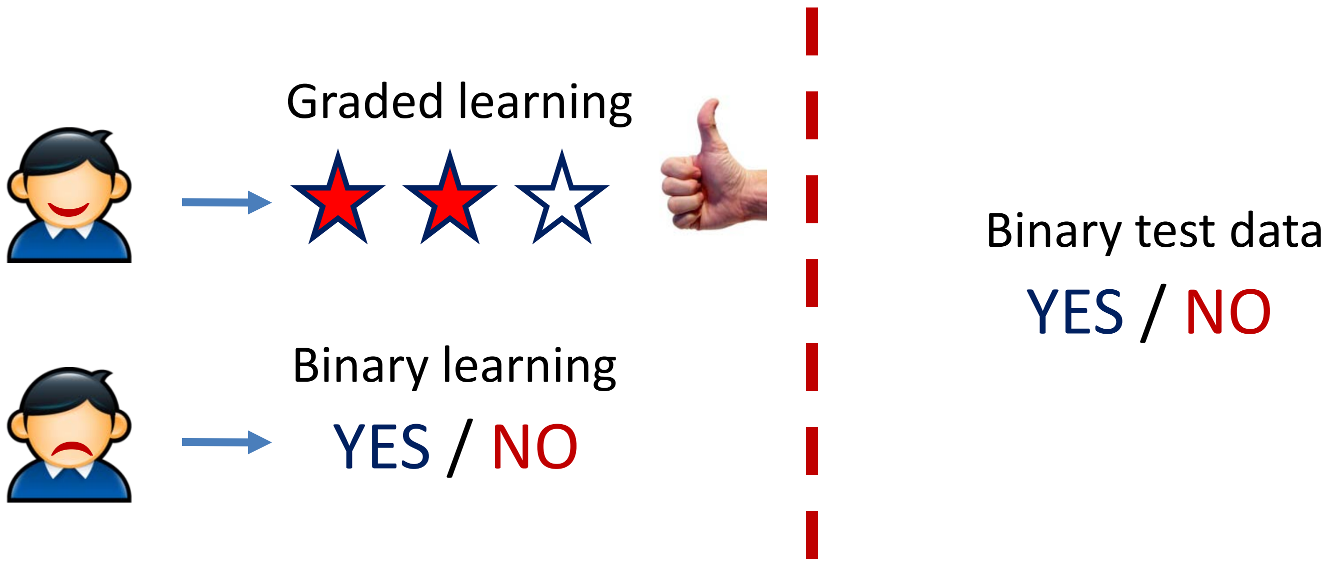
$\longrightarrow$ **It is not amenable to the horizontal reduction scheme.**

Likewise, there are generalized losses with a horizontal but no vertical representation.

Experiment – Goal

Showing the **usefulness** of the graded setting.

- We provide empirical evidence showing that labeling on graded scales offers useful extra information (binary learning *VS*. graded learning)
- We claim that training a learner on graded data can be useful even if only a binary prediction is actually requested.

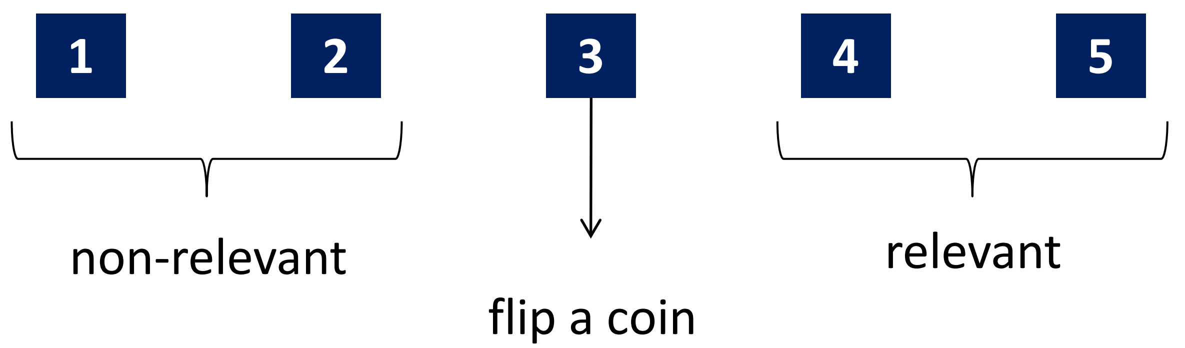


Experiment – Data

BeLa-E data set (Abele & Stief, 2004)

- Degrees of importance of the future job's different properties provided by grad students, e.g., *reputation*, *job security*, *income*, etc..
- Degrees are given in an ordinal scale from 5 to 1.
- 1930 instances, 50 attributes (48 job properties, 2 for sex and age).

Binarization (mimicking a person forced to decide):



Experiment – Setting & Results

- A subset of features is randomly chosen as labels.
- Binary learning: the whole data is binarized
- Graded learning: only predictions and test data are binarized

- 10-fold cross validation with 50 randomly generated problems.
- Paired *t*-test shows significance at level of 5%.

- *Both, vertical and horizontal, decompositions work well.*
- *Graded training shows significant advantage over binary training.*

Summary

- We proposed graded multilabel classification (GMLC) as an extension of conventional multilabel classification, since label relevance is often a matter of degree.
- We proposed two meta-techniques for GMLC, *vertical* and *horizontal* decomposition (as well as a combination).
- We proposed extensions of MLC loss functions and studied their usability with the two reduction schemes.
- We provided empirical evidence for the usefulness of learning from graded multilabel data.

Thanks!

Knowledge Engineering & Bioinformatics (KEBI)
Mathematics and Computer Science
University of Marburg

http://www.uni-marburg.de/fb12/kebi