Multi-Label Classification: Challenges, Pitfalls and Perspectives

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

MLD @ ICML 2010, Haifa, Israel, June 25, 2010

Joint work with
Krzysztof Dembczynski,
Willem Waegeman,
Weiwei Cheng.

Prelude

The distinguishing feature of multi-label classification (MLC): Instances may have **multiple** labels instead of only a **single** one!

X1	X2	X3	X4	Y1	Y2	Y3	Y4
0.34	0	10	174	0	0	**1**	0
1.45	0	32	277	0	**1**	0	0
1.22	1	46	421	0	0	0	**1**
0.74	1	25	165	0	**1**	0	0
0.95	1	72	273	**1**	0	0	0
1.04	0	33	158	0	0	**1**	0
0.92	1	81	382	?	?	?	?

In conventional classification, class labels are mutually exclusive.

Prelude

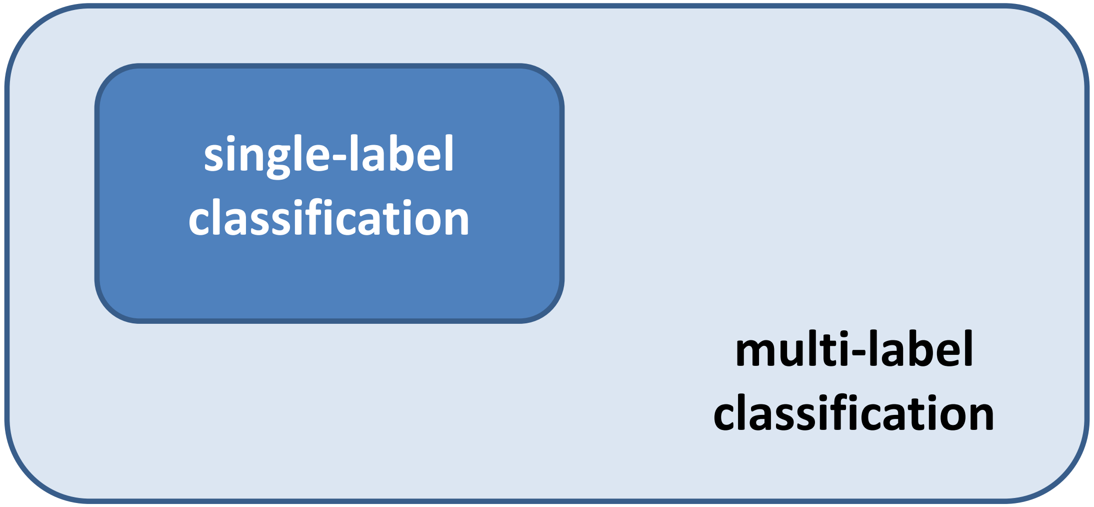
The distinguishing feature of multi-label classification (MLC): Instances may have **multiple** labels instead of only a **single** one!

X1	X2	X3	X4	Y1	Y2	Y3	Y4
0.34	0	10	174	0	**1**	**1**	0
1.45	0	32	277	0	**1**	0	**1**
1.22	1	46	421	0	0	0	1
0.74	1	25	165	0	**1**	**1**	**1**
0.95	1	72	273	**1**	0	**1**	0
1.04	0	33	158	**1**	**1**	**1**	0
0.92	1	81	382	?	?	?	?

Exploiting **label dependencies** (correlations) has become a major concern in MLC research.

Prelude

Note that conventional multi-class classification is a special case of MLC, not the other way around …



… and specific structure is normally exploited for restricted classes (e.g., linear programming).

Prelude

The idea of having „additional information“ comes from the „binary view“ (which is closely related to multi-task and transfer learning):
Multiple binary instead of a generalized multi-class problem.

X1	X2	X3	X4	English	French	Spanish	Dutch
0.34	0	10	174	0	**1**	**1**	0
1.45	0	32	277	0	**1**	0	**1**
1.22	1	46	421	0	0	0	1
0.74	1	25	165	0	**1**	**1**	**1**
0.95	1	72	273	**1**	0	**1**	0
1.04	0	33	158	**1**	**1**	**1**	0
0.92	1	81	382	?	?	?	?

Indeed, MLC problems can be „binarized“ in an obvious way, and it would be solved if **binary relevance** (BR) learning was enough …

Prelude

The typical MLC paper:

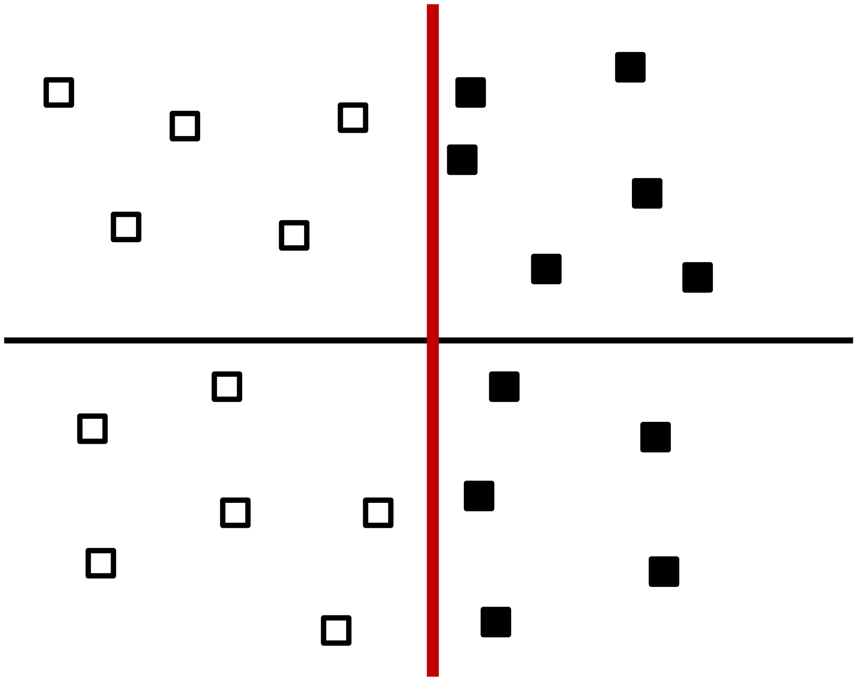
- A new method is proposed that **exploits label dependencies** in one way or the other.
- It is shown to have better **average accuracy** than existing approaches in terms of a bunch of **MLC loss functions**.

Criticisms:

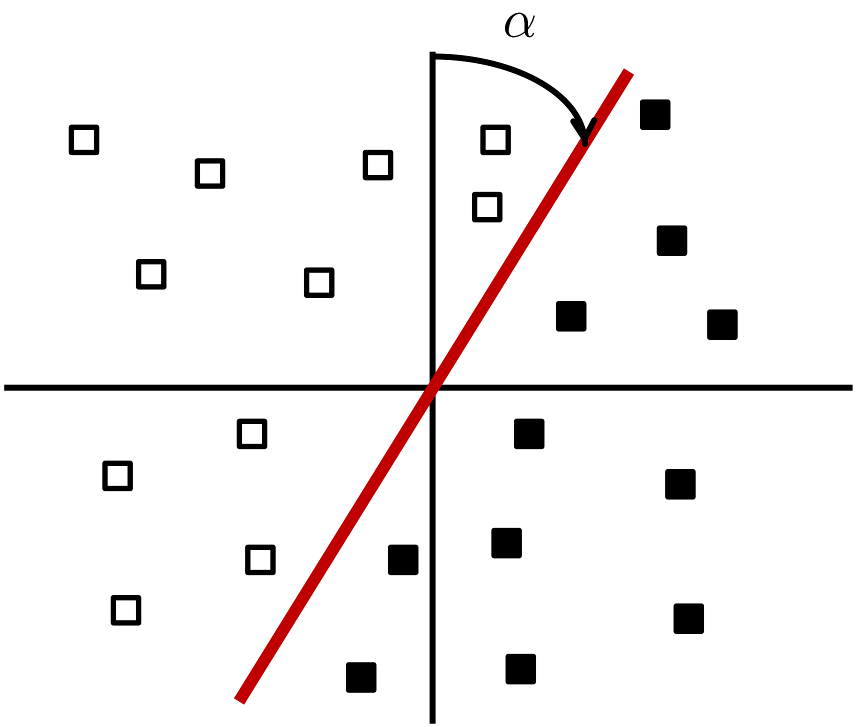
- Results are reported **on average** and not carefully analyzed: No investigation of the **reasons for improvement** and the conditions under which label dependencies are beneficial (some comments will follow).
- Implicitly, it seems to be assumed that one algorithm can be beneficial (if not optimal) **for multiple loss functions** (second part).
- The notion of “label correlation” is often used in a purely **intuitive** manner, without a precise **formal definition** (third part).

Running Example

first label (part. in ICML)

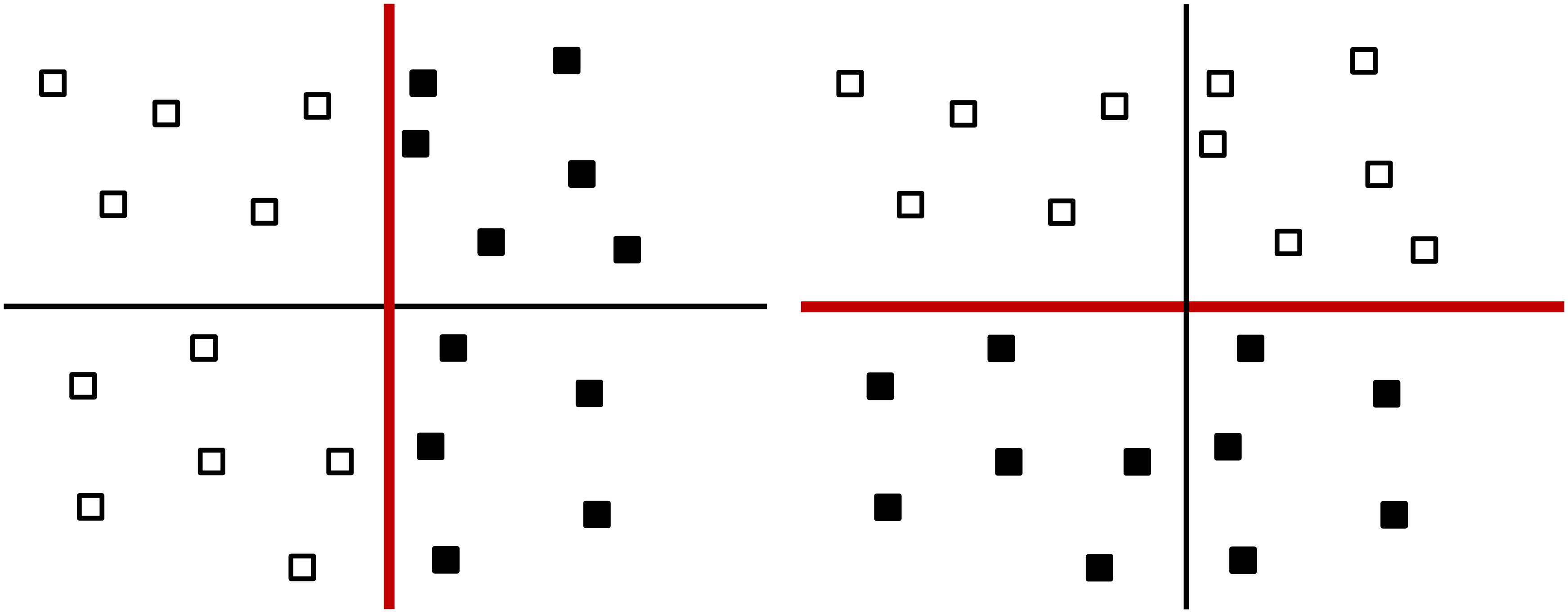


second label (part. in COLT)



Running Example

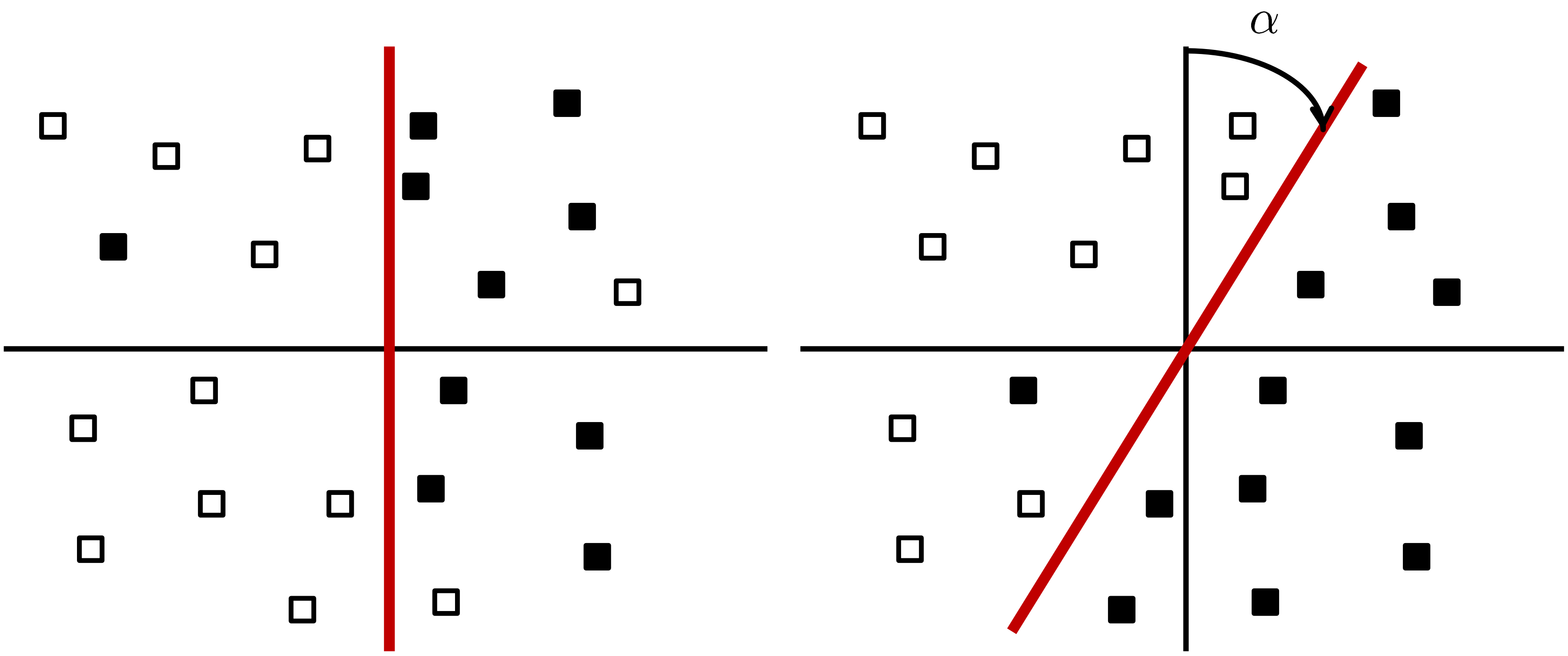
$\alpha = 0$ $\alpha = \pi$



Running Example

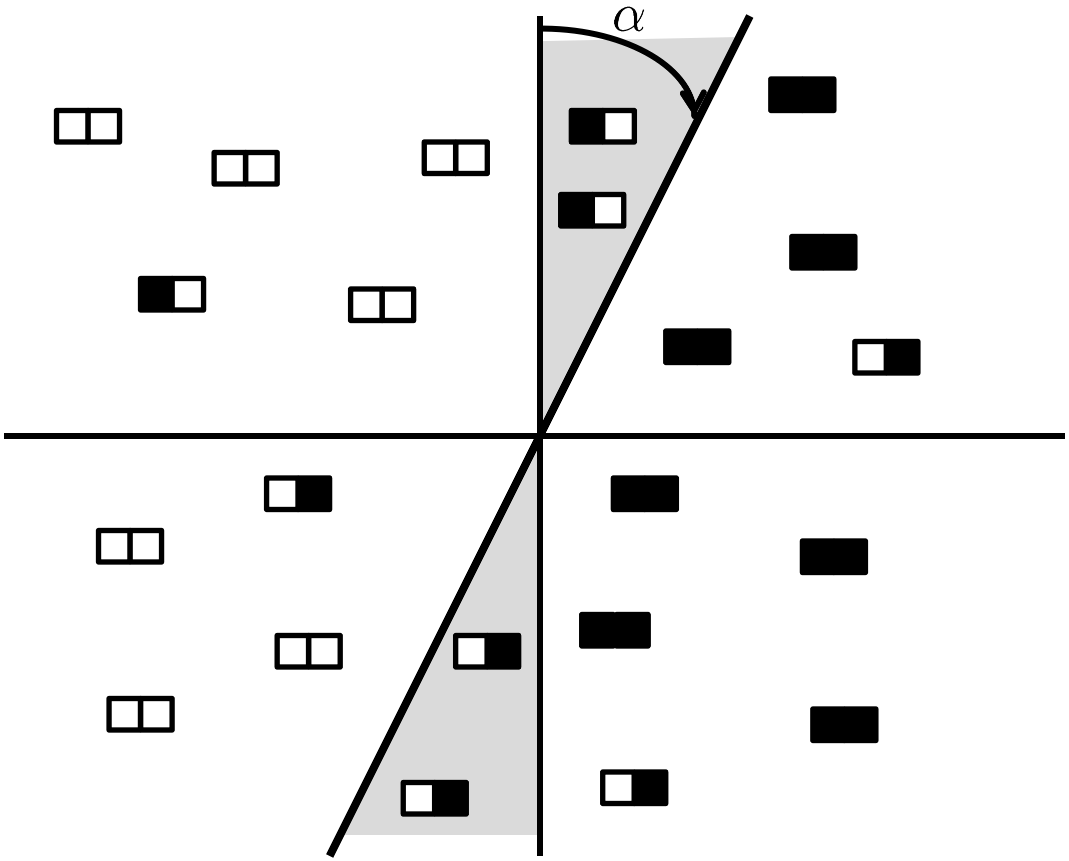
first label (part. in ICML)

second label (part. in COLT)



Adding noise by inverting each label with small probability (0.1).

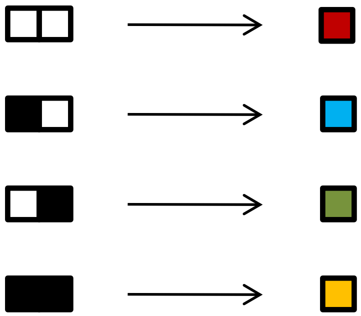
Running Example



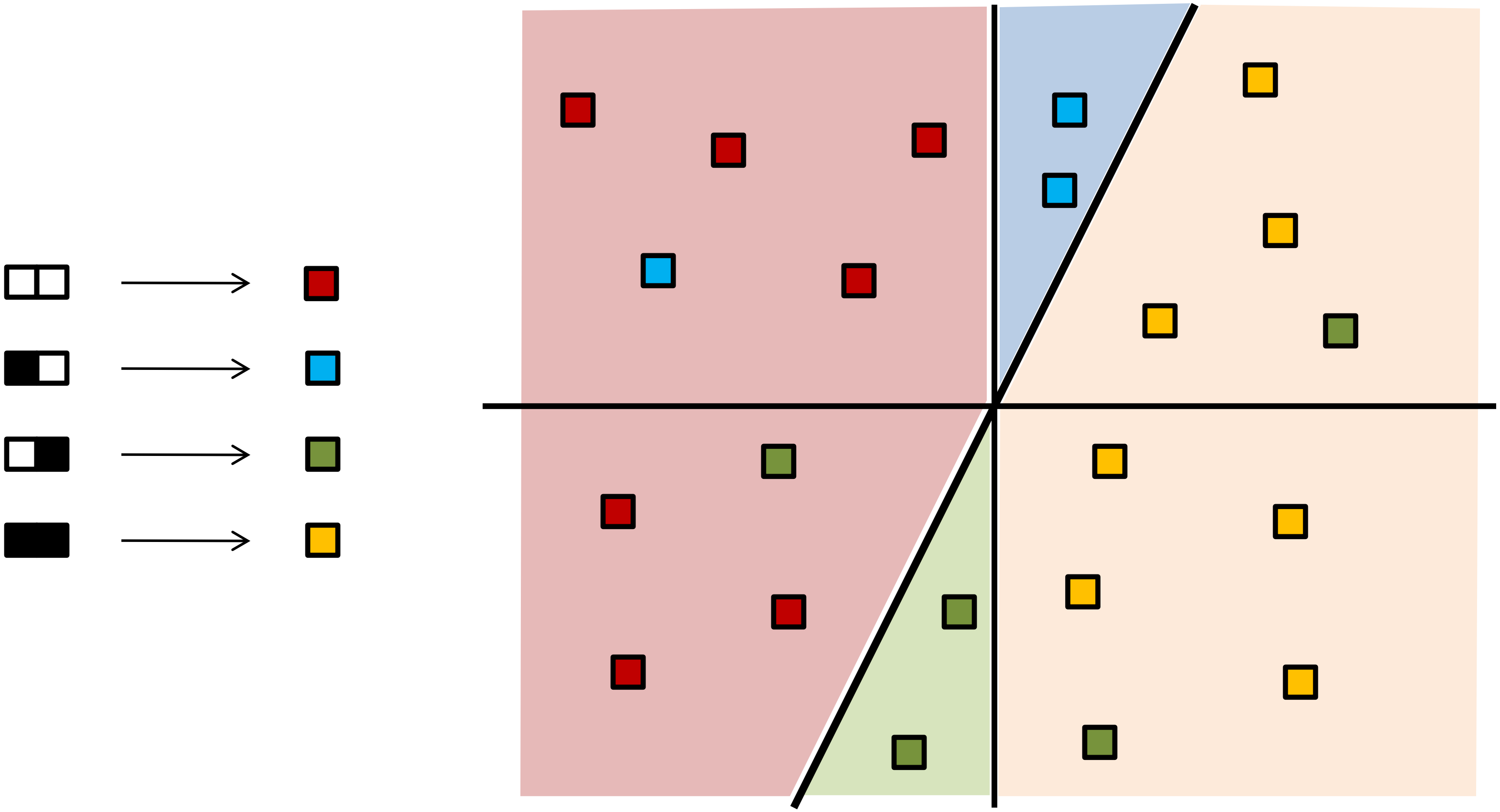
The Label Powerset (LP) Approach

Treat every potential labeling as a new (meta-)class

→ reduction to multi-class classification

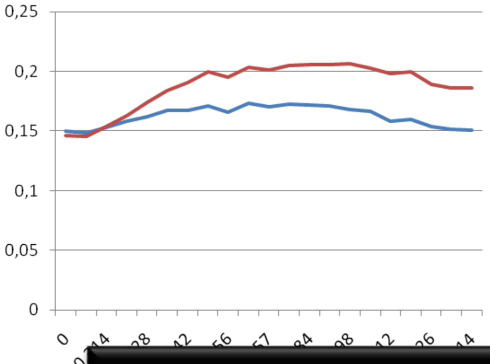


Running Example

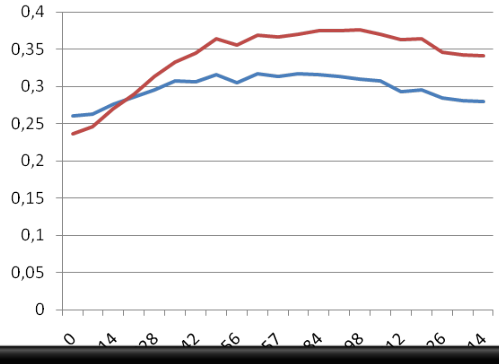


Potential advantage of taking label dependence into account comes at the cost of more classes, imbalanced distributions, more complex decision boundaries.

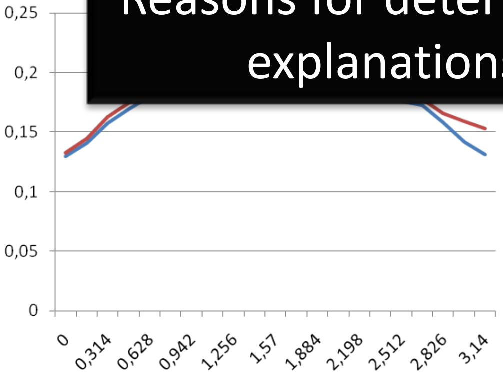
Hamming and Subset 0/1 Loss

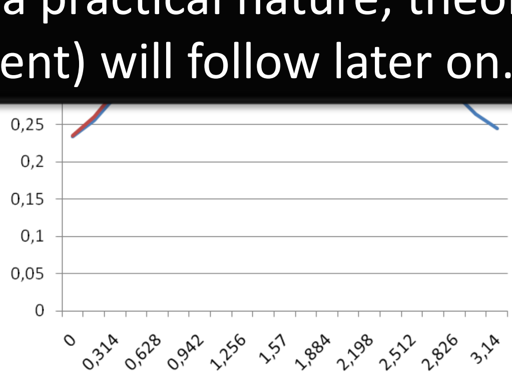


SVM (SMO)

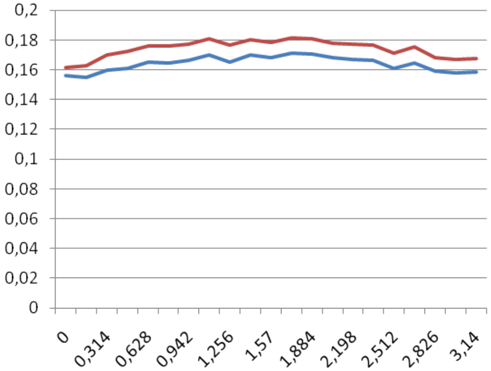


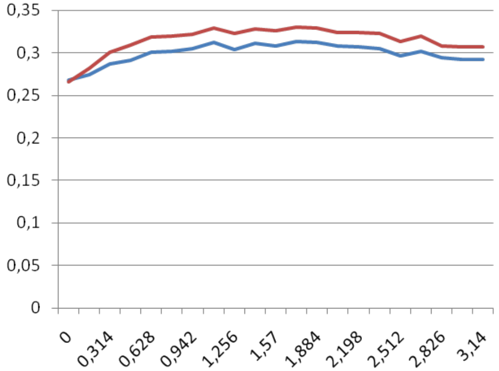
Reasons for deterioration are more of a practical nature; theoretical explanations (for non-improvement) will follow later on.





k-NN





AGENDA

1. MLC Loss Functions and Risk Minimization
2. Label Dependence
3. Connections to Related Fields (maybe)
4. Concluding Remarks

Some Notation

- Given an instance space $\mathcal{X}$ and label space $\mathcal{Y} = \{0,1\}^m$, an MLC examples $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$ is generated according to a probability distribution $\mathbf{P}(\mathbf{X}, \mathbf{Y})$.

- $\mathbf{P}_{\boldsymbol{x}}(\mathbf{Y}) = \mathbf{P}(\mathbf{Y} \,|\, \boldsymbol{x})$ is the conditional distribution of $\mathbf{Y}$ given $\mathbf{X} = \boldsymbol{x}$, and by $\mathbf{P}^{(i)}_{\boldsymbol{x}}(Y_i) = \mathbf{P}^{(i)}(Y_i \,|\, \boldsymbol{x})$ the corresponding marginal distribution of Y_i conditioned on $\boldsymbol{x}$:

$$\mathbf{P}^{(i)}_{\boldsymbol{x}}(1) = \mathbf{P}^{(i)}(1 \,|\, \boldsymbol{x}) = \sum_{\boldsymbol{y} \in \mathcal{Y} : y_i = 1} \mathbf{P}(\boldsymbol{y} \,|\, \boldsymbol{x})$$

- A multilabel classifier $\mathbf{h}$ is an $\mathcal{X} \to \mathcal{Y}$ mapping that assigns a (predicted) labeling to each instance $\boldsymbol{x} \in \mathcal{X}$. Thus, the output of a classifier $\mathbf{h}$ is a vector

$$\mathbf{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \ldots, h_m(\boldsymbol{x})).$$

Example of a Conditional Probability

$\mathbf{Y}_1$	$\mathbf{Y}_2$	$\mathbf{Y}_3$	$\mathbf{P}_{\boldsymbol{x}}(\mathbf{Y})$
0	0	0	0
0	0	1	0
0	1	0	0.4
1	0	0	0.3
0	1	1	0
1	0	1	0.3
1	1	0	0
1	1	1	0

Risk Minimization

The risk-minimizing model $\mathbf{h}^*$ is defined in a pointwise way by

$$\mathbf{h}^*(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}} \mathbb{E}_{\mathbf{Y}|\mathbf{X}} L(\mathbf{Y}, \boldsymbol{y}) \ ,$$

where $L(\cdot, \cdot)$ is a loss function defined on multi-label predictions.

MLC Loss Functions

The spectrum of losses in MLC is wider than in conventional classification!

- Subset 0/1 loss:

$$L_{0/1}(\boldsymbol{y}, \mathbf{h}(\boldsymbol{x})) = [\![\boldsymbol{y} \neq \mathbf{h}(\boldsymbol{x})]\!]$$

- Hamming loss:

$$L_H(\boldsymbol{y}, \mathbf{h}(\boldsymbol{x})) = \frac{1}{m} \sum_{i=1}^{m} [\![y_i \neq h_i(\boldsymbol{x})]\!]$$

- Rank loss:

$$L_R(\boldsymbol{y}, \mathbf{f}(\boldsymbol{x})) = \sum_{(i,j): y_i > y_j} \left([\![f_i < f_j]\!] + \frac{1}{2} [\![f_i = f_j]\!] \right)$$

MLC Loss Functions

- For the subset 0/1 loss, the risk-minimizing prediction is given by the mode of the distribution:

$$\mathbf{h}^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y}\in\mathcal{Y}} \mathbf{P}(\boldsymbol{y}\,|\,\boldsymbol{x})$$

- For the Hamming loss, the risk-minimizing prediction is given by

$$h_i^*(\boldsymbol{x}) = \arg\max_{b\in\{0,1\}} \mathbf{P}^{(i)}(b\,|\,\boldsymbol{x})$$

„joint mode“ versus **„combined mode“** (combination of the modes of the margins)

Hamming Loss vs. Subset 0/1 Loss

$\mathbf{Y}_1$	$\mathbf{Y}_2$	$\mathbf{Y}_3$	$\mathbf{P}_{\boldsymbol{x}}(\mathbf{Y})$
0	0	0	0
0	0	1	0
0	1	0	0.4
1	0	0	0.3
0	1	1	0
1	0	1	0.3
1	1	0	0
1	1	1	0

subset 0/1 risk minimizer:
(0, 1, 0)

would be the optimal choice of LP

$\mathbf{Y}_1$	$\mathbf{Y}_2$	$\mathbf{Y}_3$	$\mathbf{P}_{\boldsymbol{x}}(\mathbf{Y})$
0	0	0	0
0	0	1	0
0	1	**0**	0.4
1	**0**	**0**	0.3
0	1	1	0
1	**0**	1	0.3
1	1	0	0
1	1	1	0

Hamming risk minimizer:
(1, 0, 0)

would be the optimal choice of BR

Regret Analysis

Define the regret of a classifier $\boldsymbol{h}$ with respect to a target loss L as follows:

$$\begin{aligned} r_L(\boldsymbol{h}) &= R(\boldsymbol{h}) - R(\boldsymbol{h}^*) \\ &= \mathbb{E}_{\boldsymbol{X}\boldsymbol{Y}} L(\boldsymbol{Y}, \boldsymbol{h}(\boldsymbol{X})) - \mathbb{E}_{\boldsymbol{X}\boldsymbol{Y}} L(\boldsymbol{Y}, \boldsymbol{h}^*(\boldsymbol{X})), \end{aligned}$$

with R the risk and $\boldsymbol{h}_z^*$ the Bayes-optimal classifier with respect to L.

Specifically for Hamming and subset 0/1 loss:

$$\begin{aligned} r_H(\boldsymbol{h}) &= \mathbb{E}_{\boldsymbol{X}\boldsymbol{Y}} L_H(\boldsymbol{Y}, \boldsymbol{h}(\boldsymbol{X})) - \mathbb{E}_{\boldsymbol{X}\boldsymbol{Y}} L_H(\boldsymbol{Y}, \boldsymbol{h}_H^*(\boldsymbol{X})) \\ r_{0/1}(\boldsymbol{h}) &= \mathbb{E}_{\boldsymbol{X}\boldsymbol{Y}} L_{0/1}(\boldsymbol{Y}, \boldsymbol{h}(\boldsymbol{X})) - \mathbb{E}_{\boldsymbol{X}\boldsymbol{Y}} L_{0/1}(\boldsymbol{Y}, \boldsymbol{h}_{0/1}^*(\boldsymbol{X})) \end{aligned}$$

Regret Analysis

What is the „Hamming-regret“ of a subset 0/1 optimal classifier and vice versa?

$$r_H(\boldsymbol{h}^*_{0/1}) = \mathbb{E}_{\boldsymbol{X}\boldsymbol{Y}} L_H(\boldsymbol{Y}, \boldsymbol{h}^*_{0/1}(\boldsymbol{X})) - \mathbb{E}_{\boldsymbol{X}\boldsymbol{Y}} L_H(\boldsymbol{Y}, \boldsymbol{h}^*_H(\boldsymbol{X}))$$

use classifier tailored
for subset 0/1

seek to optimize
Hamming

$$r_{0/1}(\boldsymbol{h}^*_H) = \mathbb{E}_{\boldsymbol{X}\boldsymbol{Y}} L_{0/1}(\boldsymbol{Y}, \boldsymbol{h}^*_H(\boldsymbol{X})) - \mathbb{E}_{\boldsymbol{X}\boldsymbol{Y}} L_{0/1}(\boldsymbol{Y}, \boldsymbol{h}^*_{0/1}(\boldsymbol{X}))$$

use classifier tailored
for Hamming

seek to optimize
subset 0/1

Regret Analysis

Proposition: The following upper bound holds:

$$\mathbb{E}_{\boldsymbol{Y}} L_{0/1}(\boldsymbol{Y}, \boldsymbol{h}^*_H(\boldsymbol{x})) - \mathbb{E}_{\boldsymbol{Y}} L_{0/1}(\boldsymbol{Y}, \boldsymbol{h}^*_{0/1}(\boldsymbol{x})) < 0.5.$$

Moreover, this bound is tight, i.e.,

$$\sup_{\mathbf{P}} \left(\mathbb{E}_{\boldsymbol{Y}} L_{0/1}(\boldsymbol{Y}, \boldsymbol{h}^*_H(\boldsymbol{x})) - \mathbb{E}_{\boldsymbol{Y}} L_s(\boldsymbol{Y}, \boldsymbol{h}^*_{0/1}(\boldsymbol{x}))\right) = 0.5,$$

where the supremum is taken over all probability distributions on $\mathcal{Y}$.

K. Dembczynski et al. *Regret Analysis for Performance Metrics in Multi-Label Classication: The Case of Hamming and Subset Zero-One Loss*. ECML-2010.

Regret Analysis

Proposition: The following upper bound holds for $m > 3$:

$$\mathbb{E}_{\mathbf{Y}} L_H(\boldsymbol{Y}, \boldsymbol{h}^*_{0/1}(\boldsymbol{x})) - \mathbb{E}_{\mathbf{Y}} L_H(\boldsymbol{Y}, \boldsymbol{h}^*_H(\boldsymbol{x})) < \frac{m-2}{m+2}.$$

Moreover, this bound is tight, i.e.

$$\sup_{\mathbf{P}}(\mathbb{E}_{\mathbf{Y}} L_H(\boldsymbol{Y}, \boldsymbol{h}^*_{0/1}(\boldsymbol{x})) - \mathbb{E}_{\mathbf{Y}} L_H(\boldsymbol{Y}, \boldsymbol{h}^*_H(\boldsymbol{x}))) = \frac{m-2}{m+2},$$

where the supremum is taken over all probability distributions on $\mathcal{Y}$.

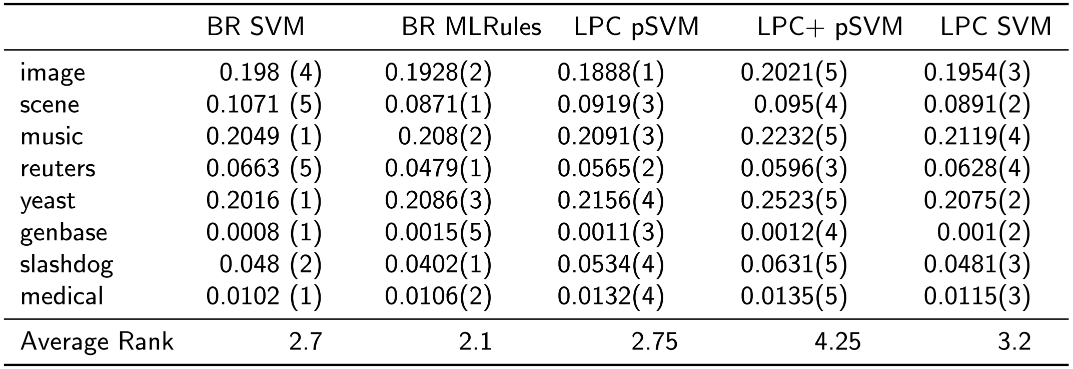
A classifier tailored for subset 0/1 loss may perform extremely poor in terms of Hamming loss (at least theoretically)!

K. Dembczynski et al. *Regret Analysis for Performance Metrics in Multi-Label Classication: The Case of Hamming and Subset Zero-One Loss*. ECML-2010.

Regret Analysis: Empirical Evidence

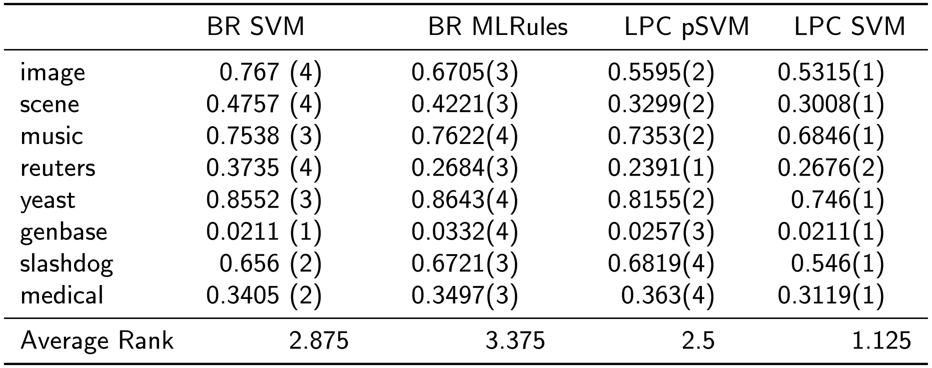
Empirically, we could confirm that BR is often better (!) than LP in terms of Hamming loss but worse in terms of subset 0/1 loss.

Hamming



	BR SVM	BR MLRules	LPC pSVM	LPC+ pSVM	LPC SVM
image	0.198 (4)	0.1928(2)	0.1888(1)	0.2021(5)	0.1954(3)
scene	0.1071 (5)	0.0871(1)	0.0919(3)	0.095(4)	0.0891(2)
music	0.2049 (1)	0.208(2)	0.2091(3)	0.2232(5)	0.2119(4)
reuters	0.0663 (5)	0.0479(1)	0.0565(2)	0.0596(3)	0.0628(4)
yeast	0.2016 (1)	0.2086(3)	0.2156(4)	0.2523(5)	0.2075(2)
genbase	0.0008 (1)	0.0015(5)	0.0011(3)	0.0012(4)	0.001(2)
slashdog	0.048 (2)	0.0402(1)	0.0534(4)	0.0631(5)	0.0481(3)
medical	0.0102 (1)	0.0106(2)	0.0132(4)	0.0135(5)	0.0115(3)
Average Rank	2.7	2.1	2.75	4.25	3.2

subset 0/1

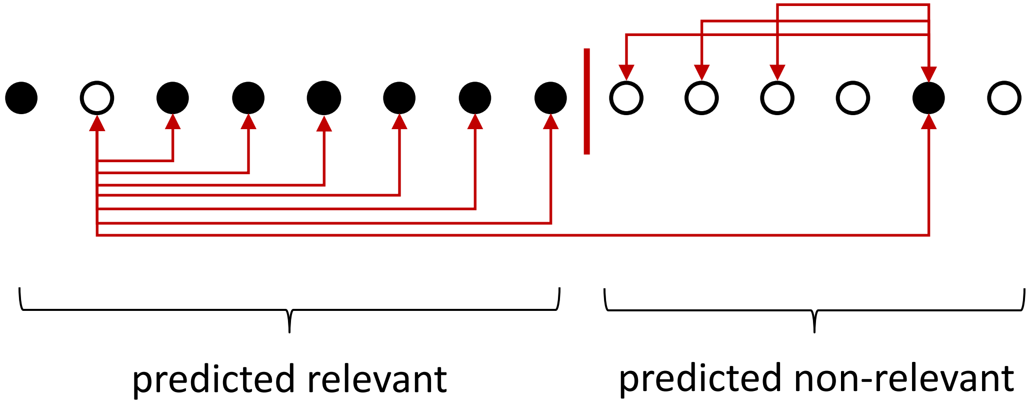


	BR SVM	BR MLRules	LPC pSVM	LPC SVM
image	0.767 (4)	0.6705(3)	0.5595(2)	0.5315(1)
scene	0.4757 (4)	0.4221(3)	0.3299(2)	0.3008(1)
music	0.7538 (3)	0.7622(4)	0.7353(2)	0.6846(1)
reuters	0.3735 (4)	0.2684(3)	0.2391(1)	0.2676(2)
yeast	0.8552 (3)	0.8643(4)	0.8155(2)	0.746(1)
genbase	0.0211 (1)	0.0332(4)	0.0257(3)	0.0211(1)
slashdog	0.656 (2)	0.6721(3)	0.6819(4)	0.546(1)
medical	0.3405 (2)	0.3497(3)	0.363(4)	0.3119(1)
Average Rank	2.875	3.375	2.5	1.125

Regret Analysis

Similar observations have been made for classification vs. ranking (e.g., learning AUC-optimizing classifiers):

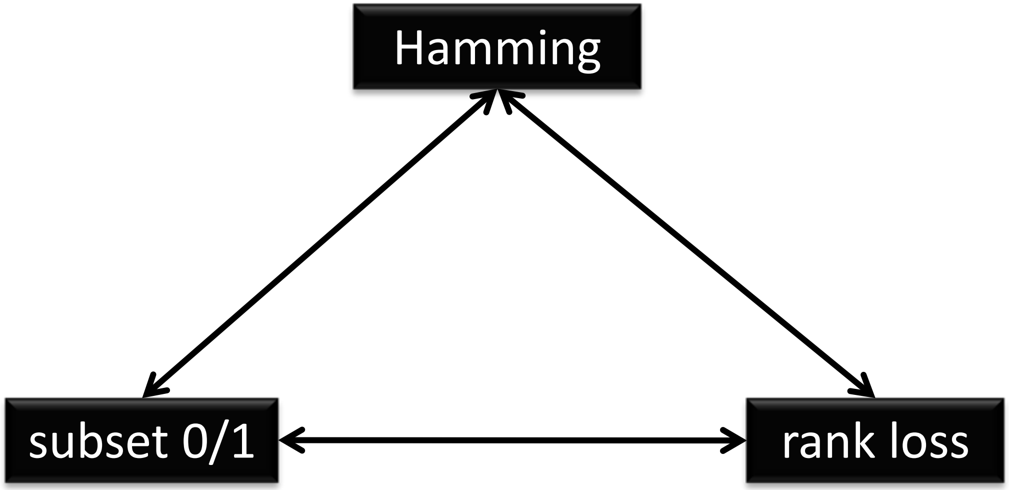
A good classifier is not necessarily a good ranker!



2 classification but
10 ranking errors

Regret Analysis

Single labels: getting individual labels right, decomposable loss



All labels: global loss, evaluating a prediction as a whole

Pairs of labels: correct sorting and comparison of labels

What Do We Learn From These Observation?

- Even though the true differences will normally be smaller than suggested by these worst case estimates, or may even shrink to zero under specific assumptions (e.g., probability of mode > ½, conditional independence), **different MLC loss functions will generally call for different classifiers**.

- Stated differently, **a single classifier is unlikely to be optimal for various loss functions** at the same time.

- Thus, empirical studies suggesting the opposite should arouse suspicion …

RAKEL: Problem Reduction

Train a label powerset classifier for each k-subset of labels, or a random subset thereof.

X1	X2	X3	X1	Y1	Y2	Y3	Y4
0.34	0	10	174	0	1	**1**	0
1.45	0	32	277	0	1	0	**1**
1.22	1	46	421	0	0	0	1
0.74	1	25	165	0	1	**1**	**1**
0.95	1	72	273	1	0	**1**	0
1.04	0	33	158	1	1	**1**	0
0.92	1	81	382	0	1	**1**	**1**

G. Tsoumakas and I. Vlahavas. *Random k-labelsets: An ensemble method for multilabel classication*. ECML 2007.

RAKEL: Prediction

	Y1	Y2	Y3	Y4
(Y1, Y2) classifier →	0	1		
(Y1, Y3) classifier →	1		1	
(Y1, Y4) classifier →	1			0
(Y2, Y3) classifier →		0	1	
(Y2, Y4) classifier →		0		1
(Y3, Y4) classifier →			0	1
	1	0	1	1

majority vote
(or thresholding)

What is RAKEL Estimating?

Y_1	Y_2	Y_3	Y_4	$P_{\boldsymbol{x}}(Y)$
0	0	0	0	3/64
0	0	0	1	6/64
0	0	1	0	2/64
0	1	0	0	3/64
1	0	0	0	6/64
0	0	1	1	8/64
0	1	0	1	8/64
1	0	0	1	0
0	1	1	0	9/64
1	0	1	0	2/64
1	1	0	0	5/64
0	1	1	1	1/64
1	0	1	1	8/64
1	1	0	1	3/64
1	1	1	0	0
1	1	1	1	0

Y_1	Y_2	$P_{\boldsymbol{x}}(Y)$
0	0	19/64
0	1	21/64
1	0	16/64
1	1	8/64

← Bayes prediction for Rakel

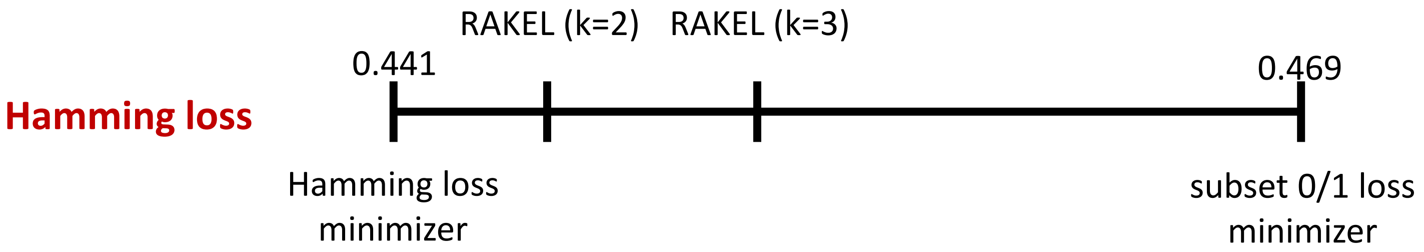
Rakel prediction: (0, 0, 1, 1)

Hamming loss minimizer: (0, 0, 0, 1)
subset 0/1 loss minimizer: (0, 1, 1, 0)

What is RAKEL Estimating?

- For k=1, RAKEL equals **Binary Relevance** (hence tailored for Hamming).
- For k=m, RAKEL equals **Label Powerset** (hence tailored for subset 0/1).
- One may conjecture that, by varying k, RAKEL smoothly „interpolates“ between Hamming and subset 0/1.

- Experiment for m=4, with probability distributions are selected uniformly at random.



Hamming loss

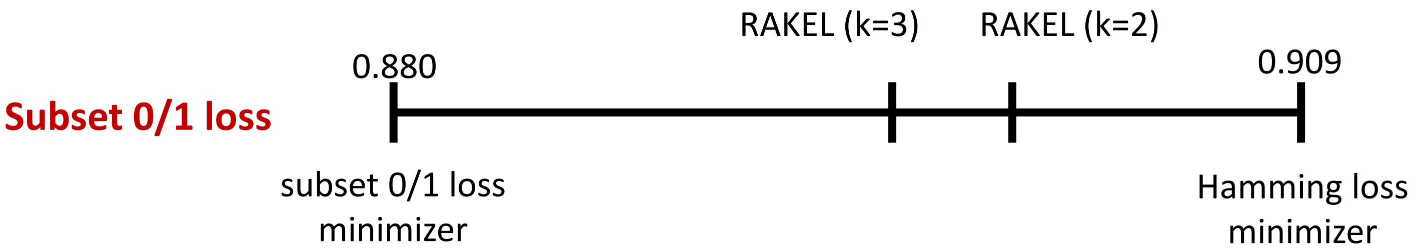
RAKEL (k=2) RAKEL (k=3)

0.441

0.469

Hamming loss minimizer

subset 0/1 loss minimizer



Subset 0/1 loss

RAKEL (k=3) RAKEL (k=2)

0.880

0.909

subset 0/1 loss minimizer

Hamming loss minimizer

AGENDA

1. MLC Loss Functions and Risk Minimization
2. **Label Dependence**
3. Connections to Related Fields
4. Concluding Remarks

Risk Minimization

The risk-minimizing model $\mathbf{h}^*$ is defined in a pointwise way by

$$\mathbf{h}^*(\boldsymbol{x}) = \arg\min_{\boldsymbol{y}} \mathbb{E}_{\mathbf{Y}|\mathbf{X}} L(\mathbf{Y}, \boldsymbol{y}) \ ,$$

where $L(\cdot, \cdot)$ is a loss function defined on multi-label predictions.

- Subset $0/1$ loss:

$$\mathbf{h}^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} \mathbf{P}_{\boldsymbol{x}}(\boldsymbol{y})$$

- Hamming loss:

$$h_i^*(\boldsymbol{x}) = \arg\max_{b \in \{0,1\}} \mathbf{P}_{\boldsymbol{x}}^{(i)}(b)$$

Exploiting label dependence is probably more important for subset 0/1 than for Hamming!

What Do We Mean by Label Dependence?

Standard stochastic independence: The joint distribution is the product of the marginals.

Conditional distribution for an instance $\boldsymbol{x}$ form the $(1,1)$ region, assuming independent error terms ϵ_1 and ϵ_2:

$\mathbf{P}_{\boldsymbol{x}}(\mathbf{Y})$	0	1	$\mathbf{P}_{\boldsymbol{x}}^{(1)}(Y_1)$
0	0.01	0.09	0.10
1	0.09	0.81	0.90
$\mathbf{P}_{\boldsymbol{x}}^{(1)}(Y_2)$	0.10	0.90	1

→ conditional label independence

What Do We Mean by Label Dependence?

Standard stochastic independence: The joint distribution is the product of the marginals.

Assuming a close dependency, namely $\epsilon_1 = \epsilon_2$:

$\mathbf{P}_{\boldsymbol{x}}(\mathbf{Y})$	0	1	$\mathbf{P}^{(1)}_{\boldsymbol{x}}(Y_1)$
0	0.1	0	0.1
1	0	0.9	0.9
$\mathbf{P}^{(1)}_{\boldsymbol{x}}(Y_2)$	0.1	0.9	1

→ conditional label dependence

How to Exploit Label Dependence?

- The **conditional distribution** is our target, as it allows for (Bayes) optimal prediction (regardless of the loss function).
- In other words, we can „exploit“ this distribution, respectively the label dependencies it implies, for **optimal decision making**.
- However, by „exploiting label correlation“ one normally means using them for **learning**, not for decision making.
- But how to exploit conditional label dependence from this point of view, given that it refers to a **single instance**?

Unconditional Label Dependence

Taking a global view on label correlation:

X1	X2	X3	X1	Y1	Y2	Y3	Y4
0.34	0	10	174	0	**1**	**1**	0
1.45	0	32	277	0	**1**	0	**1**
1.22	1	46	421	0	0	0	1
0.74	1	25	165	0	**1**	**1**	**1**
0.95	1	72	273	**1**	0	**1**	0
1.04	0	33	158	**1**	**1**	**1**	0
0.92	1	81	382	0	**1**	**1**	**1**

Unconditional label dependence refers to the joint distribution $\mathbf{P}(\mathbf{Y})$ instead of $\mathbf{P}_{\boldsymbol{x}}(\mathbf{Y}) = \mathbf{P}(\mathbf{Y} \,|\, \boldsymbol{x})$.

Connections between the Dependencies

$$\mathbf{P}(\mathbf{Y}) \;=\; \int_{\mathcal{X}} \mathbf{P}(\mathbf{Y} \,|\, \boldsymbol{x})\, d\mu(\boldsymbol{x})$$

unconditional dependence = „average“ conditional dependence

$$\underset{\substack{\uparrow \\ \text{conditional}}}{\mathbf{P}(\mathbf{Y} \,|\, \mathbf{X})} \;\propto\; \mathbf{P}(\mathbf{X} \,|\, \mathbf{Y}) \times \underset{\substack{\uparrow \\ \text{unconditional}}}{\mathbf{P}(\mathbf{Y})}$$

→ exploiting unconditional dependence, typically via regularization, for predicting conditional distributions (or functions thereof)

Stacking

stacking classifier

$$(y_1, y_2) \mapsto (y_1^*, y_2^*)$$

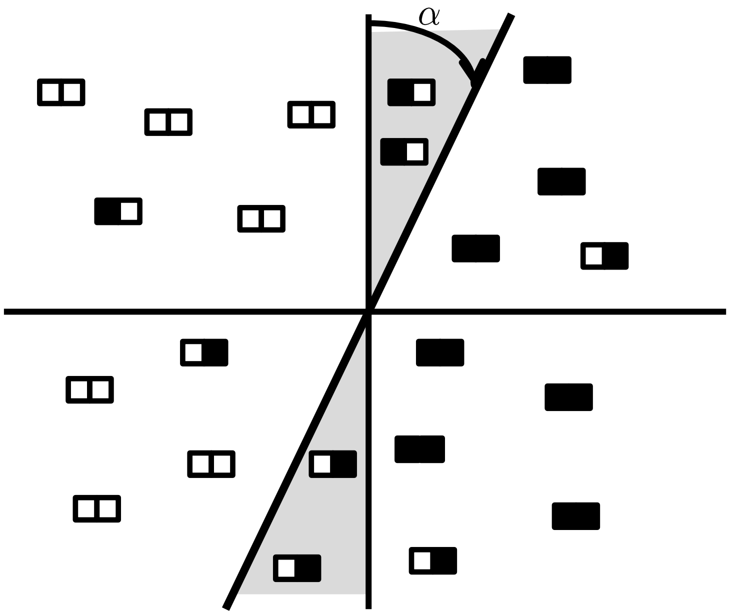
$$\uparrow$$

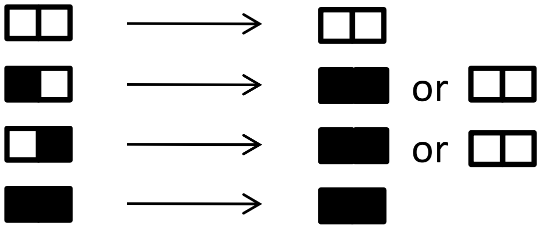
$$(x_1, x_2) \mapsto (h_1(x_1, x_2), h_2(x_1, x_2)) = (y_1, y_2)$$

base classifier (BR)

		predicted		true	
x_1	x_2	y_1	y_2	y_1	y_2
0.34	-0.45	0	0	1	0
0.45	0.56	1	1	1	1
-0.22	0.82	0	0	0	1
0.74	-0.12	0	1	0	0

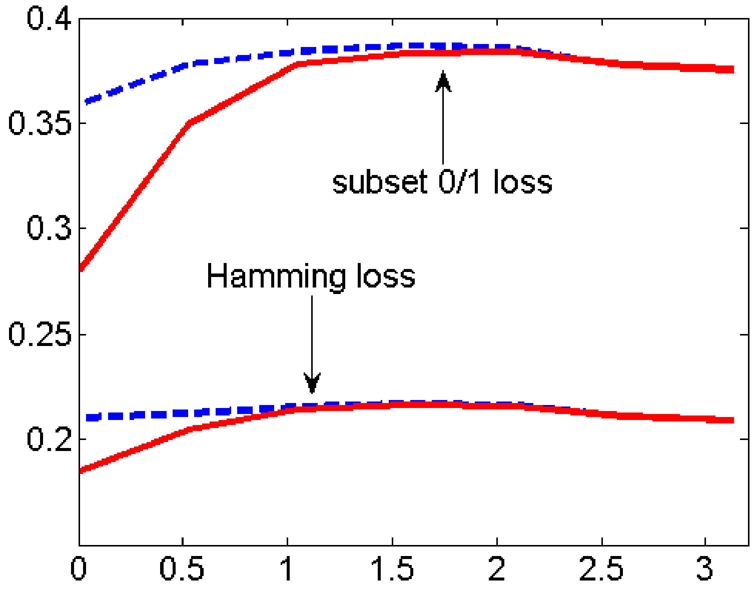
Stacking in the Running Example





Improvement **on average!**

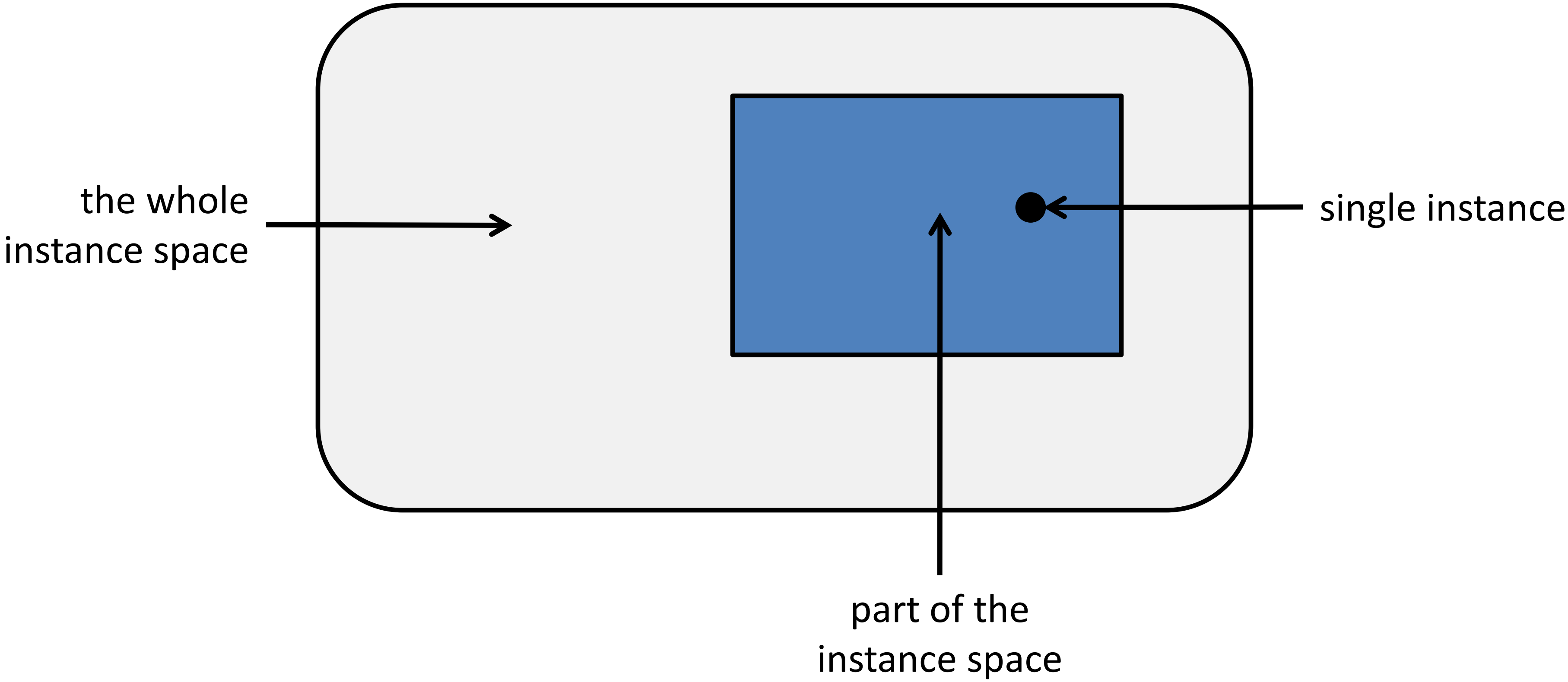
Average loss as a function of the angle, with and without stacking



- base classifier = 1-NN
- stacking by table classifier
- number of training examples = 50

In-Between Conditional and Unconditional Dependence

Extended stacking classifier: $(x_1, x_2 \,|\, y_1, y_2) \mapsto (y_1^*, y_2^*)$



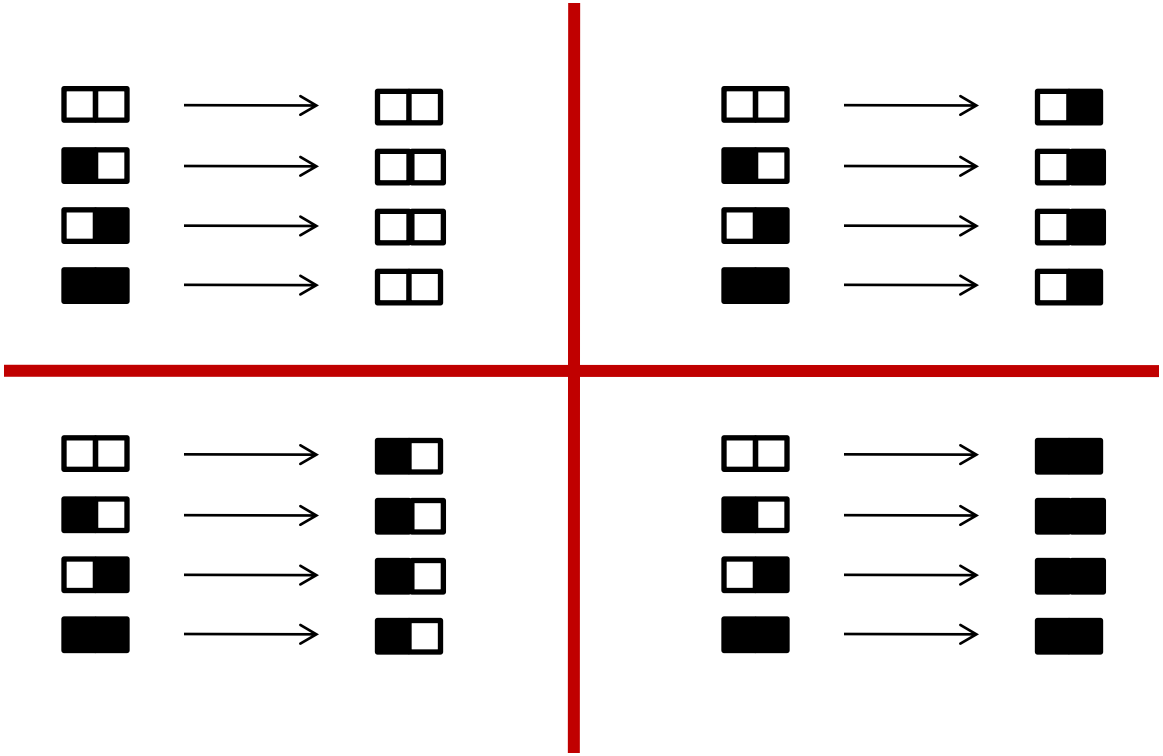
the whole instance space

single instance

part of the instance space

In-Between Conditional and Unconditional Dependence

Extended stacking classifier: $(x_1, x_2 \,|\, y_1, y_2) \mapsto (y_1^*, y_2^*)$



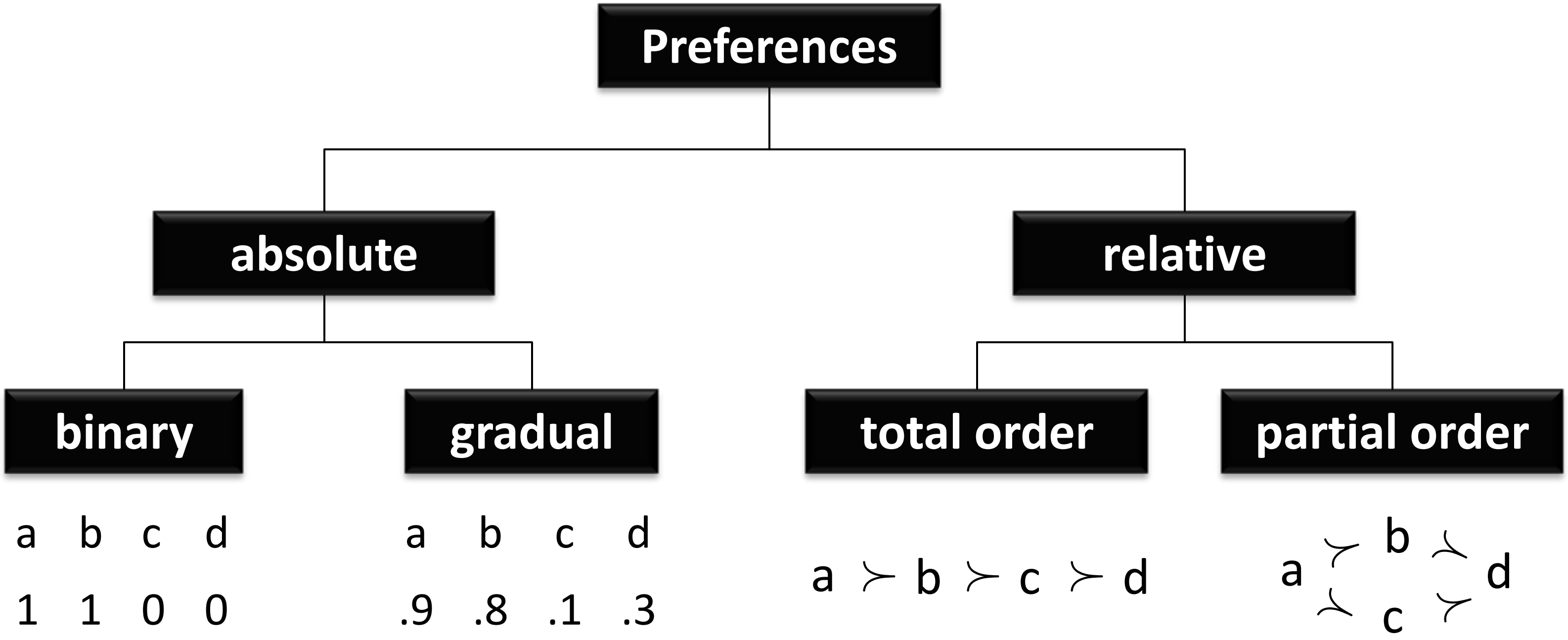
AGENDA

1. MLC Loss Functions and Risk Minimization
2. Label Dependence
3. **Connections to Related Fields**
4. Concluding Remarks

Related Methods and Research Fields

- **Stacking, Bayesian inference** and related regularization methods.
- **Multi-task learning** and **transfer learning** (also based on regularization)
- **Reduced rank regression** for **multivariate regression** with unconditionally dependent outputs.
- **Label dimensionality reduction**, e.g., kernel dependency estimation.
- **Structured output prediction**
- **Conditional random fields** and **graphical models**
- **Preference Learning**

Preference Learning



Preferences

absolute

relative

binary

gradual

total order

partial order

a	b	c	d
1	1	0	0

a	b	c	d
.9	.8	.1	.3

$a \succ b \succ c \succ d$

Preference Learning

Training	Prediction	
binary	binary	multi-label classification
binary	total order	multi-label classification (ranking)
graded	graded	graded muti-label classification (ICML 2010)
partial order	total order	label ranking
partial order	partial order	ranking with abstention (ECML 2010)

Calibrated Label Ranking

Combining absolute and relative evaluation:

$$\underbrace{a \succ b \succ c}_{\substack{\text{relevant}\\ \text{positive}\\ \text{liked}}} \;|\; \underbrace{d \succ e \succ f \succ g}_{\substack{\text{irrelevant}\\ \text{negative}\\ \text{disliked}}}$$

F. Fürnkranz et al. Multilabel classification via calibrated label ranking. Machine Learning 73(2), 2008.

Summary and Conclusions

- **Comparing MLC methods** is difficult, since the effect of „taking label correlations into account“ is hard to isolate from other changes that become necessary.

- **MLC loss functions** are of a quite different nature, and it‘s arguably impossible to minimize all of them by the same algorithm …

- **Label dependence** can be considered at different levels (conditional, unconditional, in-between), and the concrete mechanisms of „taking label correlations into account“ call for an explanation.

- MLC is related to many **other subfields of ML** (and statistics), and existing work in these fields should not be ignored.