

Learning from Multi-Label Data

(多标记学习)

Min-Ling Zhang

School of Computer Science & Engineering,
Southeast University, China

URL: <http://cse.seu.edu.cn/people/zhangml/>

Email: zhangml@seu.edu.cn



MLA'10, Nanjing

Outline

- Multi-Label Learning (MLL)
- Learning Algorithms
- Advanced Topics
- Resources

ECML/PKDD 2009 Tutorial on

“Learning from Multi-Label Data”

by Grigorios Tsoumakas, Min-Ling Zhang, Zhi-Hua Zhou

<http://www.ecmlpkdd2009.net/program/tutorials/learning-from-multi-label-data/>

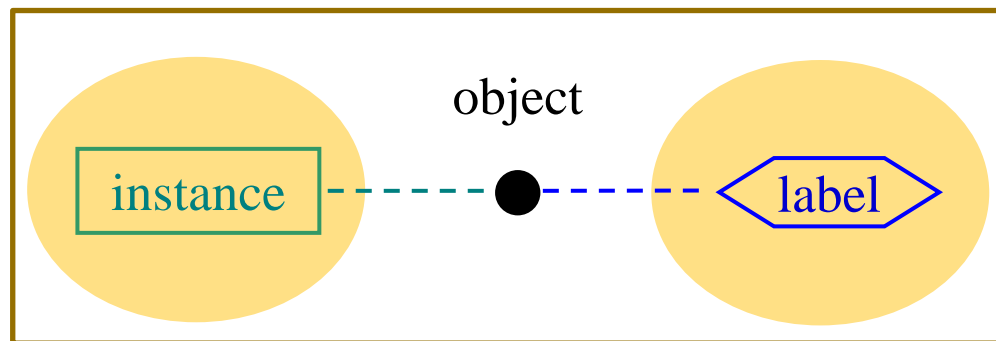


Outline

- Multi-Label Learning (MLL)
 - What is MLL?
 - Challenge & Philosophy
 - Evaluation Metrics
- Learning Algorithms
- Advanced Topics
- Resources



Traditional Supervised Learning



- **Input space**: represented by **a single instance** (feature vector) characterizing its properties
- **Output space**: associated with **a single label** characterizing its semantics

Basic assumption

real-world objects are unambiguous

Multi-Label Objects

South Africa's FIFA World Cup – a success at home and abroad



Success, pride and unity – three words which are being used by the people of South Africa to describe the effects that staging the 2010 FIFA World Cup™ has had on their country. These feelings are supported by the positive experiences of international fans who visited South Africa during the event, as highlighted in post-event research commissioned by FIFA.

Back in December 2008, FIFA commissioned a six-wave study of South African residents with the aim of tracking public opinion towards the tournament from the initial build-up through to the final whistle and then beyond. The picture that emerged following the final wave of the survey is of a country that took increasing pride in a tournament which was considered not only a huge success in its own right, but also an important event in terms of promoting national unity.

When asked in 2008 whether they thought the FIFA World Cup would be able to bring the South African people even closer together, 75 per cent of those asked said they believed this was a possibility. The post-event findings suggest that the event strengthened this sentiment, with 91 per cent of South Africans claiming their country is now more unified. The post-tournament results also showed an upswing in national confidence, with nine out of ten feeling that their country had a stronger sense of self-belief post-tournament and 87 per cent feeling more confident than ever before in their nation's capabilities.

Sports

Africa

Economics

Diego Forlán

Octopus Paulo

.....

**Multiple
labels**

Multi-Label Objects - More



Sunset

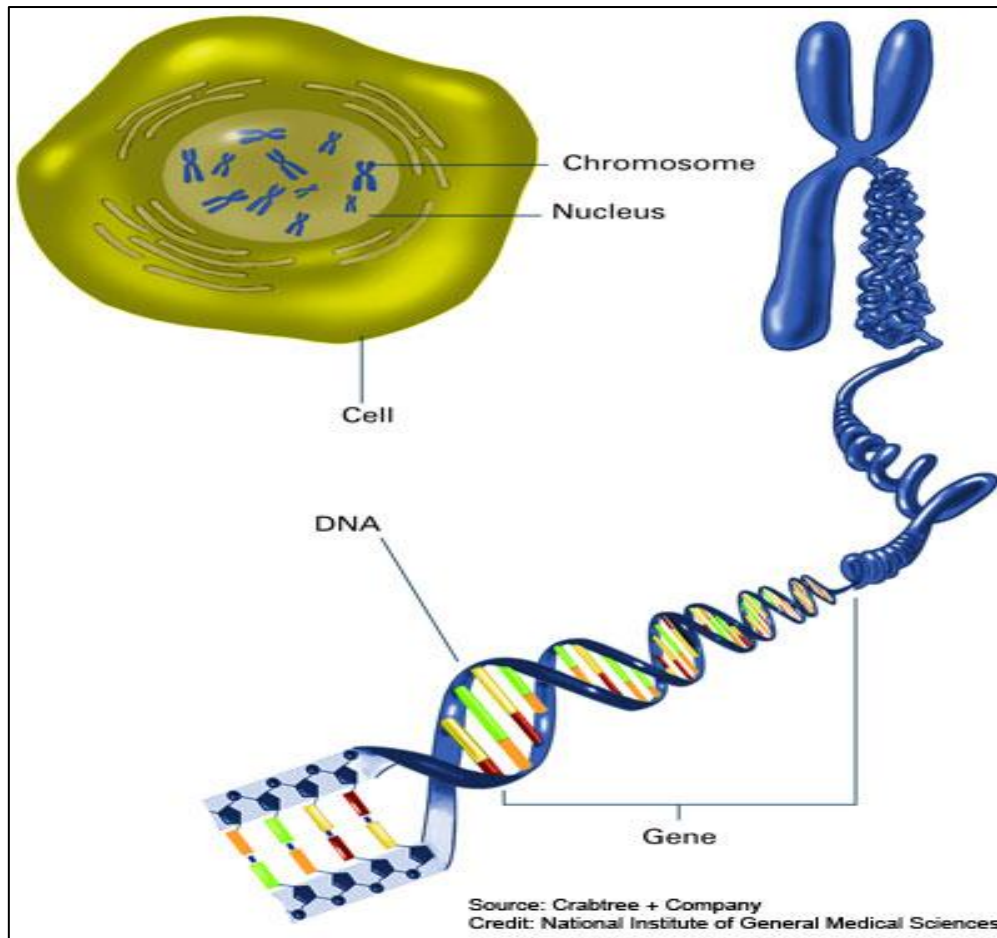
Clouds

Trees

Countryside

.....

Multi-Label Objects - More



Metabolism

Transcription

Protein
synthesis

.....

Multi-Label Objects - More



Piano

Classical music

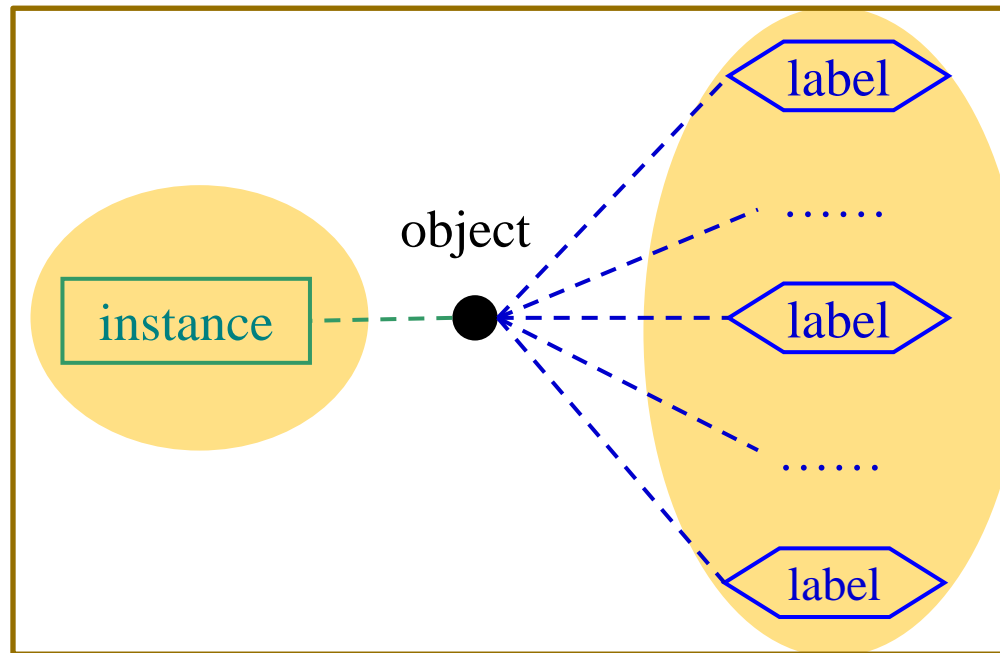
Mozart

Austria

.....

Multi-label objects are ubiquitous !

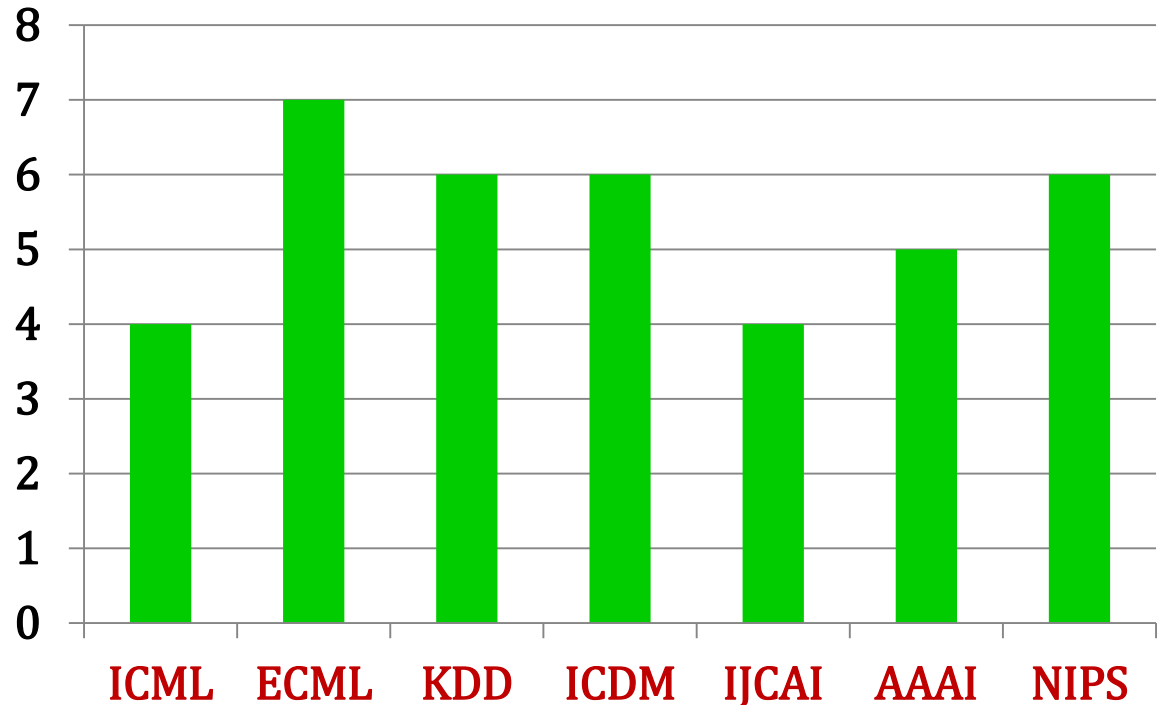
Multi-Label Learning (MLL)



Multi-Label Learning (MLL)

Snapshots on MLL - Papers

Distributions of
papers in major ML-
related conferences
with keyword
"multi-label" in title
(2007-2010)



30+ papers +

1 ECML'09 Best Student Paper

1 NIPS'09 Best Student Paper Honorable Mention



Snapshots on MLL - Events

Tutorial

“Learning from Multi-Label Data”

in conjunction with *ECML/PKDD 2009*

Workshops

MLD'09 - <http://lpis.csd.auth.gr/workshops/mld09/>

in conjunction with *ECML/PKDD 2009*

MLD'10 - <http://cse.seu.edu.cn/conf/MLD10/>

in conjunction with *ICML/COLT 2010*

Machine Learning Journal Special Issue

<http://mlkd.csd.auth.gr/events/ml2010si.html>

expected to appear next year



Snapshots on MLL - Applications

- Text Categorization
- Automatic annotation for multimedia contents
 - Image, Audio, Video
- Bioinformatics
- World Wide Web
- Information Retrieval
- Directed marketing
-



So, MLL is an
important topic
&
fast growing !

Formal Definition of MLL

Settings

\mathcal{X} : d -dimensional feature space \mathbb{R}^d

\mathcal{Y} : label space with q labels $\{1, 2, \dots, q\}$

Inputs

\mathcal{D} : training set with m examples $\{(\mathbf{x}_i, Y_i) \mid 1 \leq i \leq m\}$

$\mathbf{x}_i \in \mathcal{X}$ is a d -dimensional feature vector $(\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{id})^T$

$Y_i \subseteq \mathcal{Y}$ is the label set associated with \mathbf{x}_i

Outputs

h : multi-label predictor $\mathcal{X} \rightarrow 2^{\mathcal{Y}}$



Formal Definition of MLL - Cont.

Alternative Outputs

f : a ranking function $\mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$

Here, $f(\mathbf{x}, y)$ returns the “**confidence**” of labeling \mathbf{x} with y

Given a **threshold function** $t : \mathcal{X} \rightarrow \mathbb{R}$, we have

$$h(\mathbf{x}) = \{y \mid f(\mathbf{x}, y) > t(\mathbf{x})\}$$

Here, $t(\mathbf{x})$ produces a bipartition of label space \mathcal{Y} into **relevant** label set and **irrelevant** label set

Caveat here:

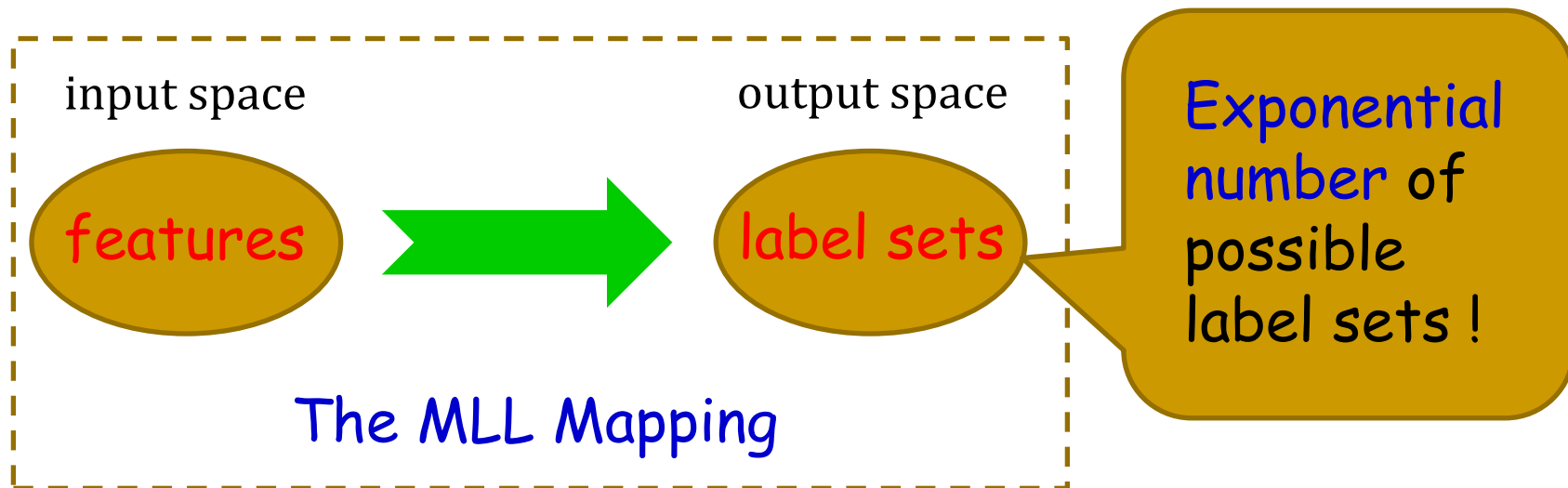
MLL \nRightarrow Label Ranking

Outline

- Multi-Label Learning (MLL)
 - What is MLL?
 - Challenge and Philosophy
 - Evaluation Metrics
- Learning Algorithms
- Advanced Topics
- Resources



The Major Challenge



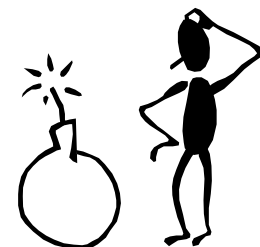
$q=5 \rightarrow 32$ label sets

$q=10 \rightarrow \sim 1\text{k}$ label sets

$q=20 \rightarrow \sim 1\text{M}$ label sets

.....

How can we
take on this
challenge?



The Basic Philosophy

Exploiting Label Correlations



For instance:

An image labeled as **lions** and **grassland** would be **likely** annotated with label **Africa**

A document labeled as **politics** would be **unlikely** labeled as **entertainment**

Order of Correlations

First-Order Strategy

Tackle MLL problem in a label-by-label style,
ignore the co-existence of other labels

e.g.: decomposing MLL into q number of independent
binary classification problems

Pros:

conceptually simple, efficient and easy to implement

Cons:

label correlations totally ignored, less effective



Order of Correlations - Cont.

Second-Order Strategy

Tackle MLL problem by considering pairwise relations between labels

e.g.: ranking between relevant and irrelevant labels,
interaction between a pair of labels, etc.

Pros:

correlations exploited, relatively effective

Cons:

correlations may go beyond second-order



Order of Correlations - Cont.

High-Order Strategy

Tackle MLL problem by considering high-order relations between labels

e.g.: among all the possible labels, among a subset of labels, etc.

Pros:

more appropriate for realistic correlations

Cons:

high model complexity, less scalable

Outline

- Multi-Label Learning (MLL)
 - What is MLL?
 - Challenge and Philosophy
 - Evaluation Metrics
- Learning Algorithms
- Advanced Topics
- Resources



Categories of Multi-Label Metrics

Traditional single-label metrics not directly applicable for MLL evaluation

Example-based Metrics [Shapire & Singer, MLJ00]

1. Calculate separately for each test example
2. Return the **mean value** across the test set

Label-based Metrics [Tsoumakas & Vlahavas, ECML'07]

1. Calculate separately for each possible label
2. Return the **macro-/micro- averaged value** across all labels



Example-based Metrics

Y_i : the actual label set associated with \mathbf{x}_i

P_i : the predicted label set for \mathbf{x}_i , i.e. $h(\mathbf{x}_i)$

$\text{rank}_f(\mathbf{x}_i, y)$: return the rank of y according to $f(\mathbf{x}_i, \cdot)$

Subset Accuracy: $\frac{1}{m} \sum_{i=1}^m \mathbb{I}[P_i = Y_i]$ \uparrow ($\mathbb{I}[\text{true}] = 1, \mathbb{I}[\text{false}] = 0$)

fraction of examples with consistent predicted label set (usually rather low when q is large)

Hamming Loss: $\frac{1}{m} \sum_{i=1}^m \frac{|P_i \Delta Y_i|}{q}$ \downarrow (Δ : symmetric difference)

fraction of misclassified example-label pairs

Example-based Metrics - Cont.

One-error: $\frac{1}{m} \sum_{i=1}^m \mathbb{I}[\arg \max_{y \in \mathcal{Y}} f(\mathbf{x}_i, y)] \notin Y_i$ ↓

fraction of examples whose top-ranked label not being relevant

Coverage: $\frac{1}{m} \sum_{i=1}^m \max_{y \in Y_i} \text{rank}_f(\mathbf{x}_i, y) - 1$ ↓

“average depth” of the lowest-ranked relevant label

Ranking Loss: $\frac{1}{m} \sum_{i=1}^m \frac{|\{(y_1, y_2) | f(\mathbf{x}_i, y_1) < f(\mathbf{x}_i, y_2) \mid (y_1, y_2) \in Y_i \times \bar{Y}_i\}|}{|Y_i| |\bar{Y}_i|}$ ↓

fraction of mis-ordered label pairs

Average Precision: $\frac{1}{m} \sum_{i=1}^m \frac{1}{|Y_i|} \sum_{y \in Y_i} \frac{|\{y' \mid f(\mathbf{x}_i, y') \geq f(\mathbf{x}_i, y), y' \in Y_i\}|}{\text{rank}_f(\mathbf{x}_i, y)}$ ↑

macro-averaged precision for relevant labels

Label-based Metrics

The single-label contingency table

the j -th label		actual output	
		YES	NO
classifier output	YES	TP_j	FP_j
	NO	FN_j	TN_j

TP_j : True Positive

FP_j : False Positive

TN_j : True Negative

FN_j : False Negative

$B(TP_j, FP_j, FN_j, TN_j)$: binary metrics derived from the table

e.g.: $\text{Accuracy} = B(TP_j, FP_j, FN_j, TN_j) = \frac{TP_j + TN_j}{TP_j + FP_j + FN_j + TN_j}$

$$\text{Precision} = B(TP_j, FP_j, FN_j, TN_j) = \frac{TP_j}{TP_j + FP_j}$$

Label-based Metrics - Cont.

q contingency tables (one per label)

the 1st label		actual output	
		YES	NO
classifier output	YES	TP_1	FP_1
	NO	FN_1	TN_1

.....

the q -th label		actual output	
		YES	NO
classifier output	YES	TP_q	FP_q
	NO	FN_q	TN_q

Macro-averaging: ["equal weight" for labels]

$$B_{\text{macro}} = \frac{1}{q} \sum_{j=1}^q B(TP_j, FP_j, FN_j, TN_j)$$

Micro-averaging: ["equal weight" for examples]

$$B_{\text{micro}} = B\left(\sum_{j=1}^q TP_j, \sum_{j=1}^q FP_j, \sum_{j=1}^q FN_j, \sum_{j=1}^q TN_j\right)$$

Notes on Multi-Label Metrics

Which Metric(s) Should One Use?

- ❑ Existing MLL metrics cover a broad spectrum
- ❑ No such “general-purpose” MLL metrics
- ❑ Depend on concrete application domains
 - ✓ **Retrieval**: Label-based metrics may be preferred, e.g. micro-avg precision
 - ✓ **Classification**: Example-based metrics may be preferred, e.g. hamming loss

What are their relationships? [Dembczyński et al., ECML'10]

Is the optimization of some metrics related to the optimization of other ones?



Outline

- Multi-Label Learning (MLL)
- Learning Algorithms
 - A Brief Taxonomy
 - Problem Transformation Methods
 - Algorithm Adaptation Methods
- Advanced Topics
- Resources



A Category of MLL Algorithms

Problem Transformation Methods

- ❑ Transform the MLL task into other well-established learning tasks
- ❑ MLL \rightarrow Binary classification; MLL \rightarrow Label ranking; MLL \rightarrow Multi-class classification;

Fit Data
to Algorithm

Algorithm Adaptation Methods

- ❑ Extend popular learning techniques to deal with multi-label data directly
- ❑ Multi-label lazy learner; Multi-label kernel learner; Multi-label Bayesian learner;

Fit Algorithm
to Data

Upcoming Algorithms

Problem Transformation Methods

- ❑ **First-order:** Binary Relevance [Boutell et al., PRJ04]
- ❑ **Second-order:** Calibrated Label Ranking [Fürnkranz et al. MLJ08]
- ❑ **High-order:** Random k -labelsets [Tsoumakas & Vlahavas, ECML'07]

Algorithm Adaptation Methods

- ❑ **First-order:** ML- k NN [Zhang & Zhou, PRJ07]
- ❑ **Second-order:** Rank-SVM [Elisseeff & Weston, NIPS'02]
- ❑ **High-order:** LEAD [Zhang & Zhang, KDD'10]



Outline

- Multi-Label Learning (MLL)
- Learning Algorithms
 - A Brief Taxonomy
 - Problem Transformation Methods
 - Algorithm Adaptation Methods
- Advanced Topics
- Resources

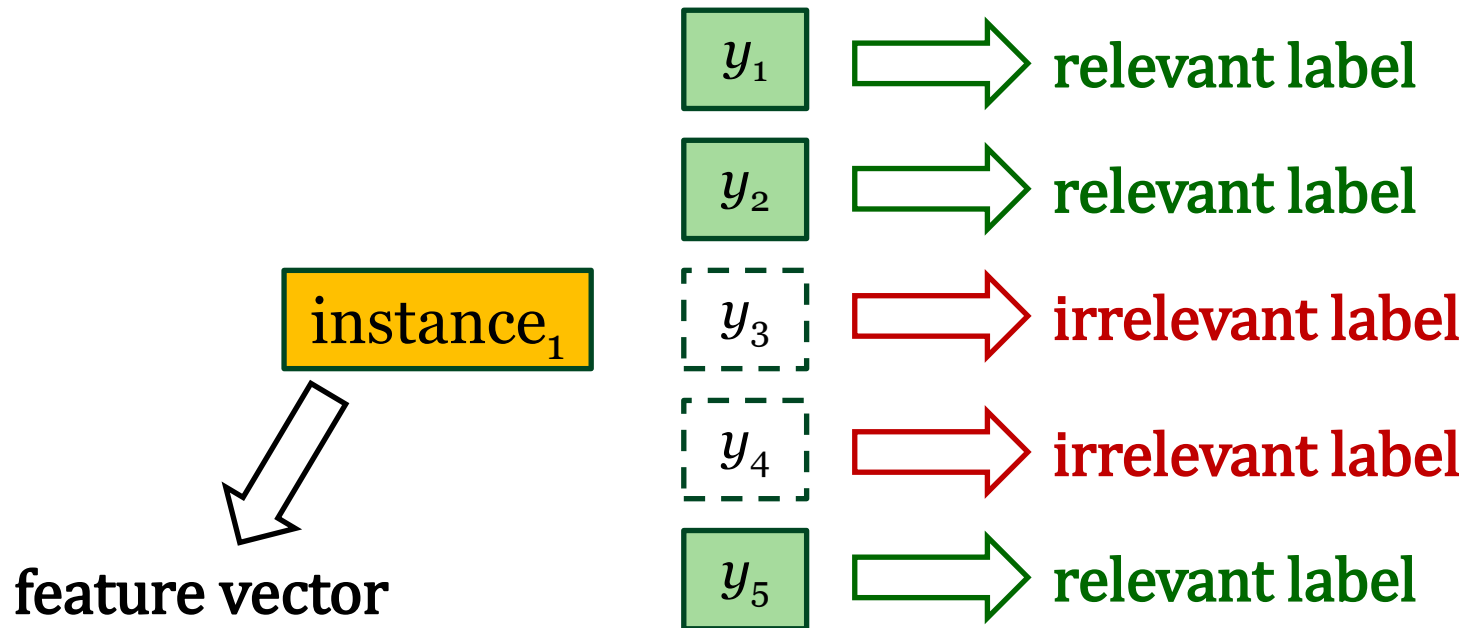


Binary Relevance [Boutell et al., PRJ04]

Basic Idea

decompose MLL into q independent binary problems

Illustrative Data

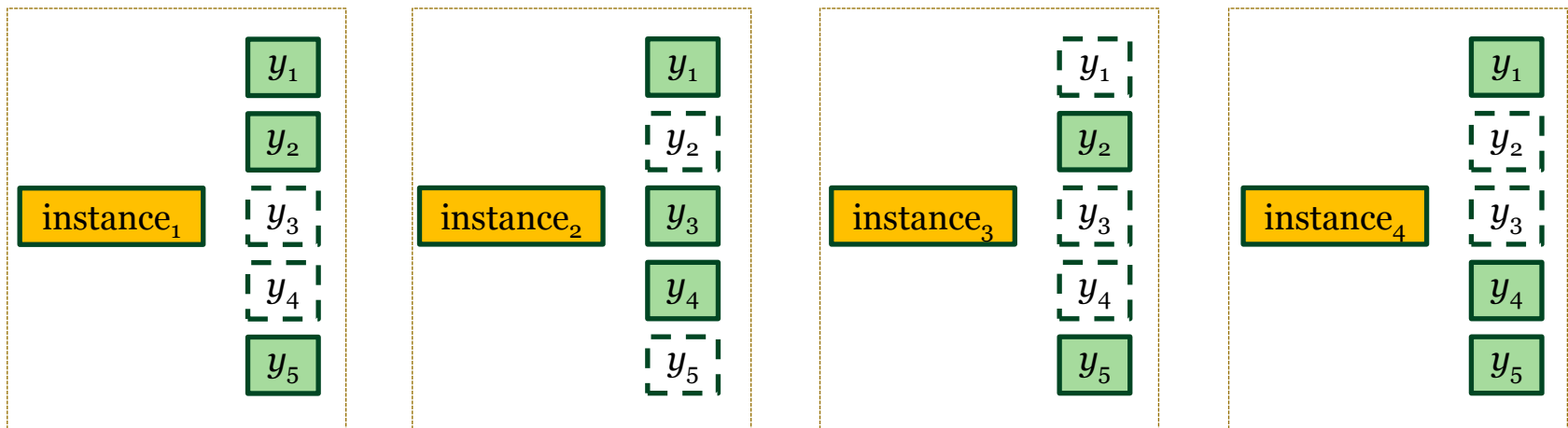


Binary Relevance [Boutell et al., PRJ04]

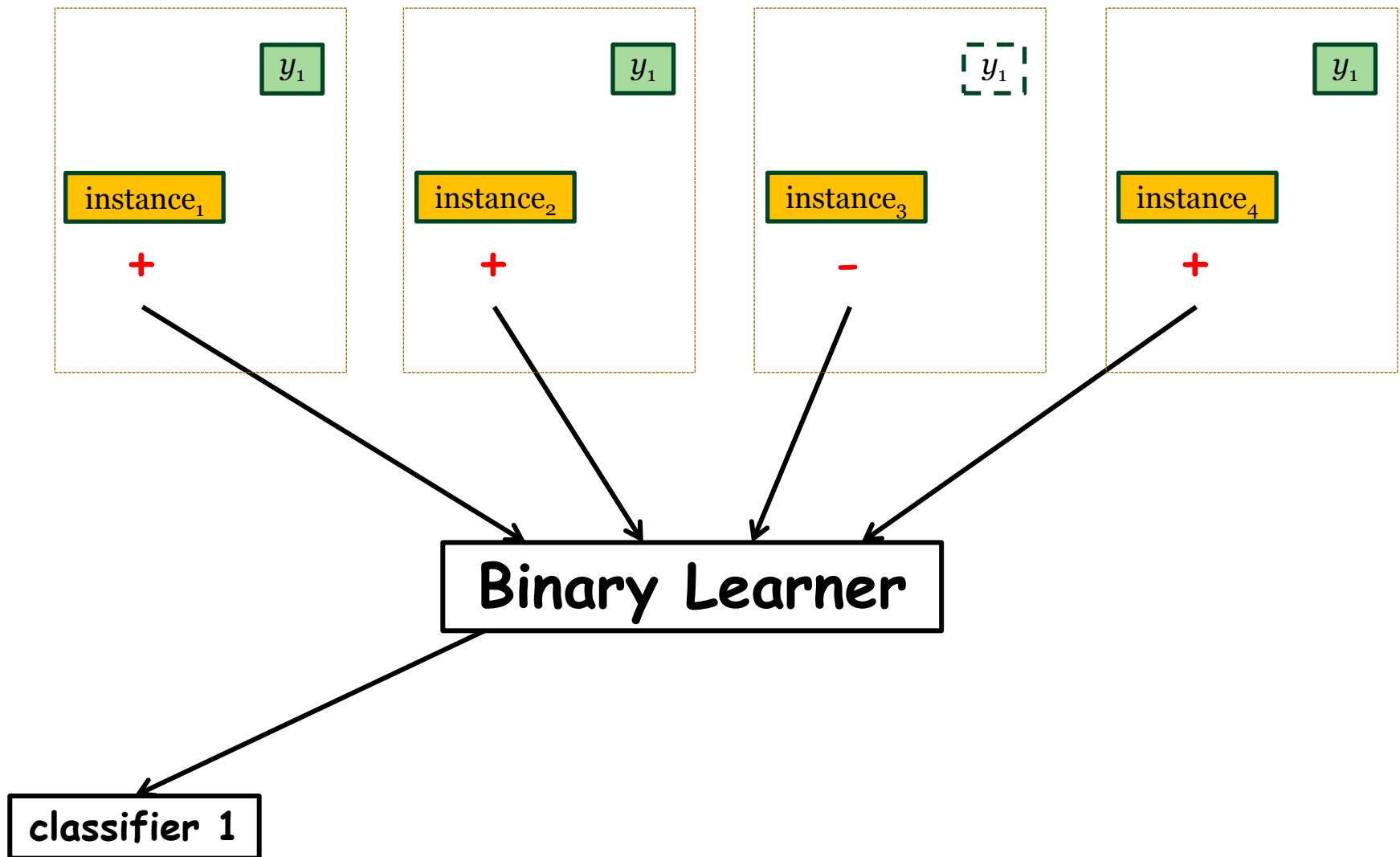
Basic Idea

decompose MLL into q independent binary problems

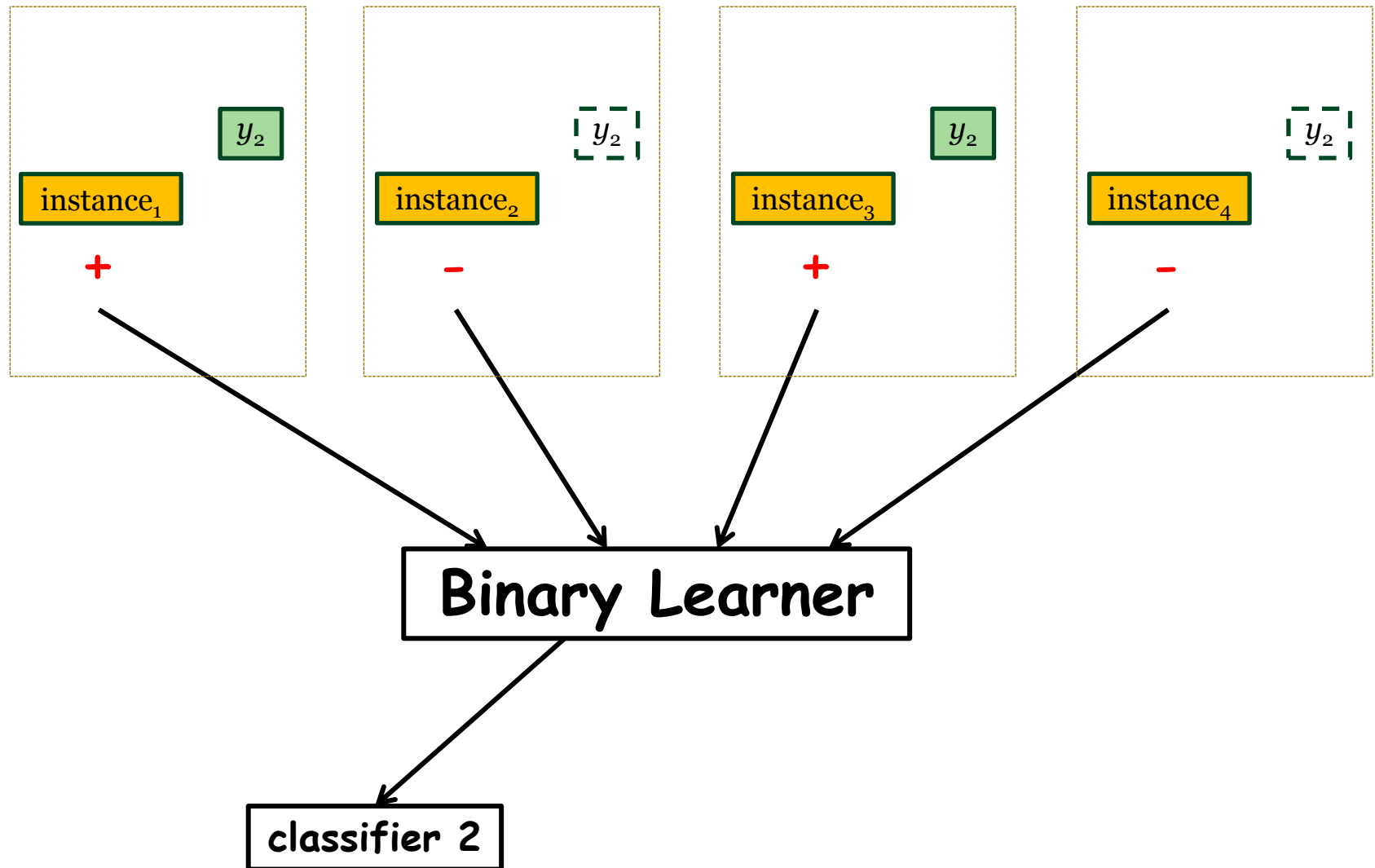
Illustrative Data



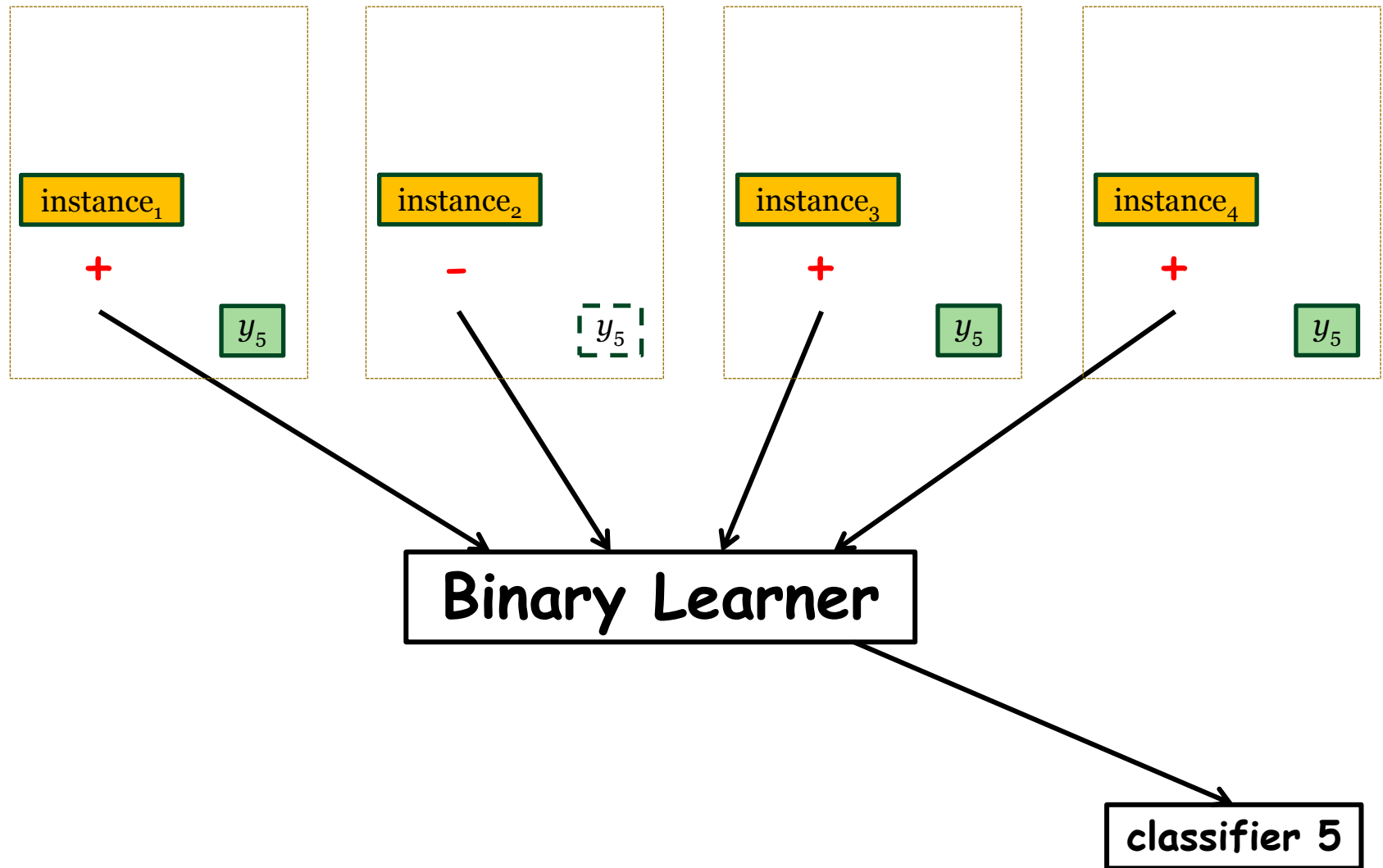
Binary Relevance - Cont.



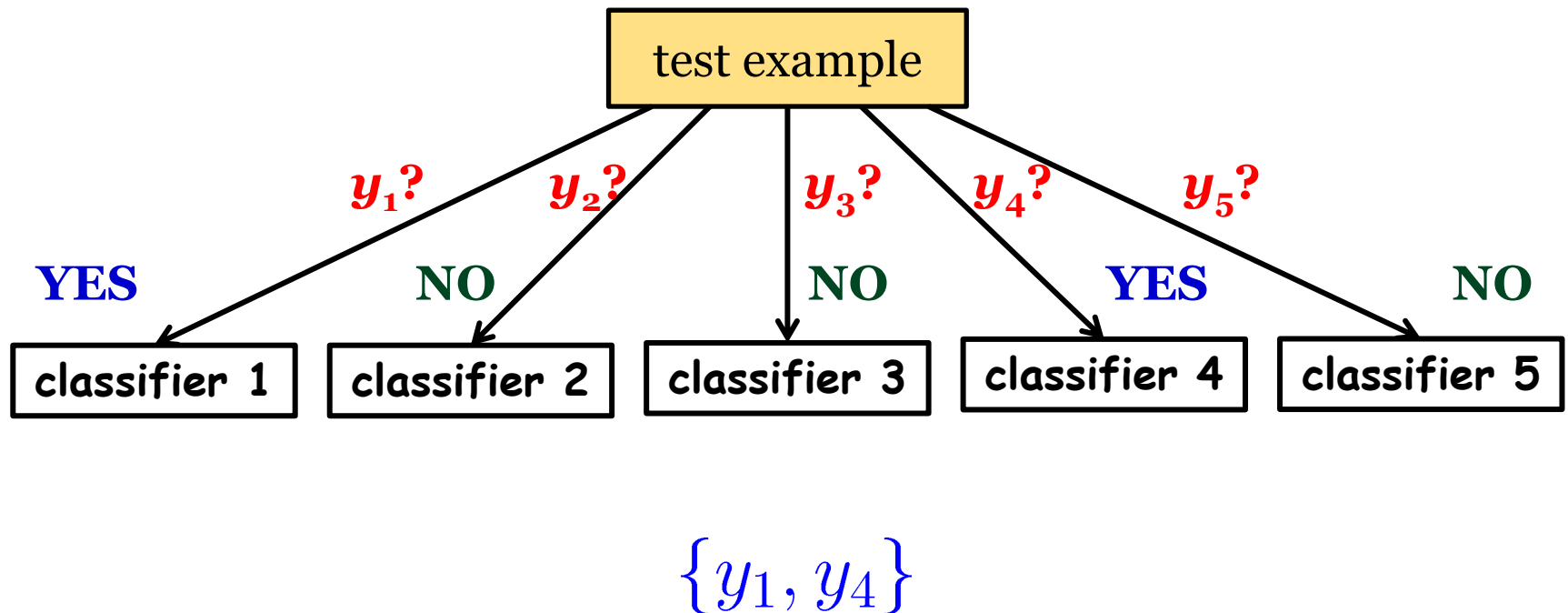
Binary Relevance - Cont.



Binary Relevance - Cont.



Binary Relevance - Cont.



Calibrated Label Ranking [Fürnkranz et al. MLJ08]

Basic Idea

transform MLL into a label ranking problem by pairwise comparison

Ranking by Pairwise Comparison

Learn $q(q-1)/2$ binary models, one for each label pair (y_j, y_k) , $1 \leq j < k \leq q$

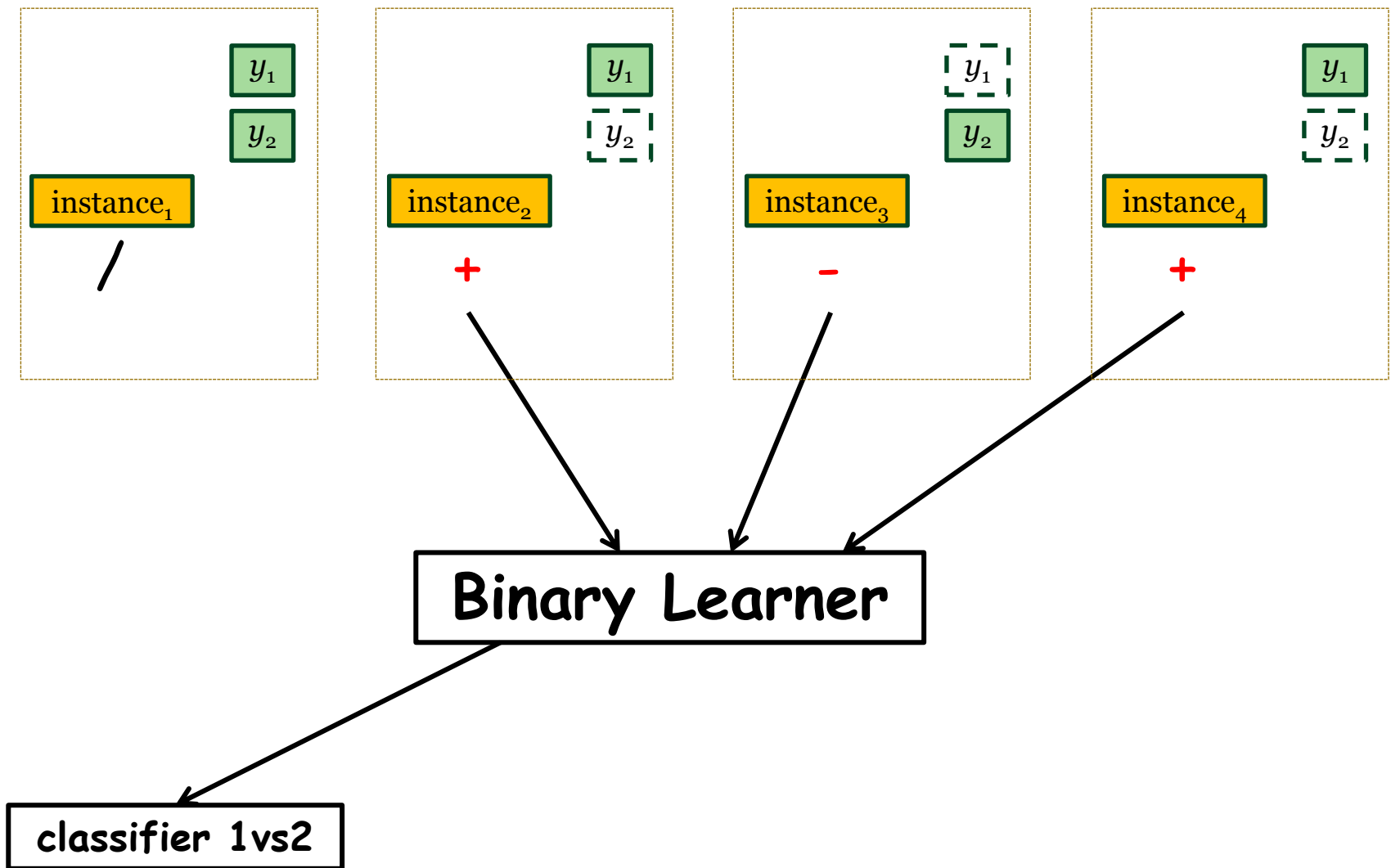
Training set for binary model (y_j, y_k) :

- ❑ x_i used as positive example if $y_j \in Y_i$ and $y_k \notin Y_i$
- ❑ x_i used as negative example if $y_j \notin Y_i$ and $y_k \in Y_i$
- ❑ Otherwise, x_i is ignored

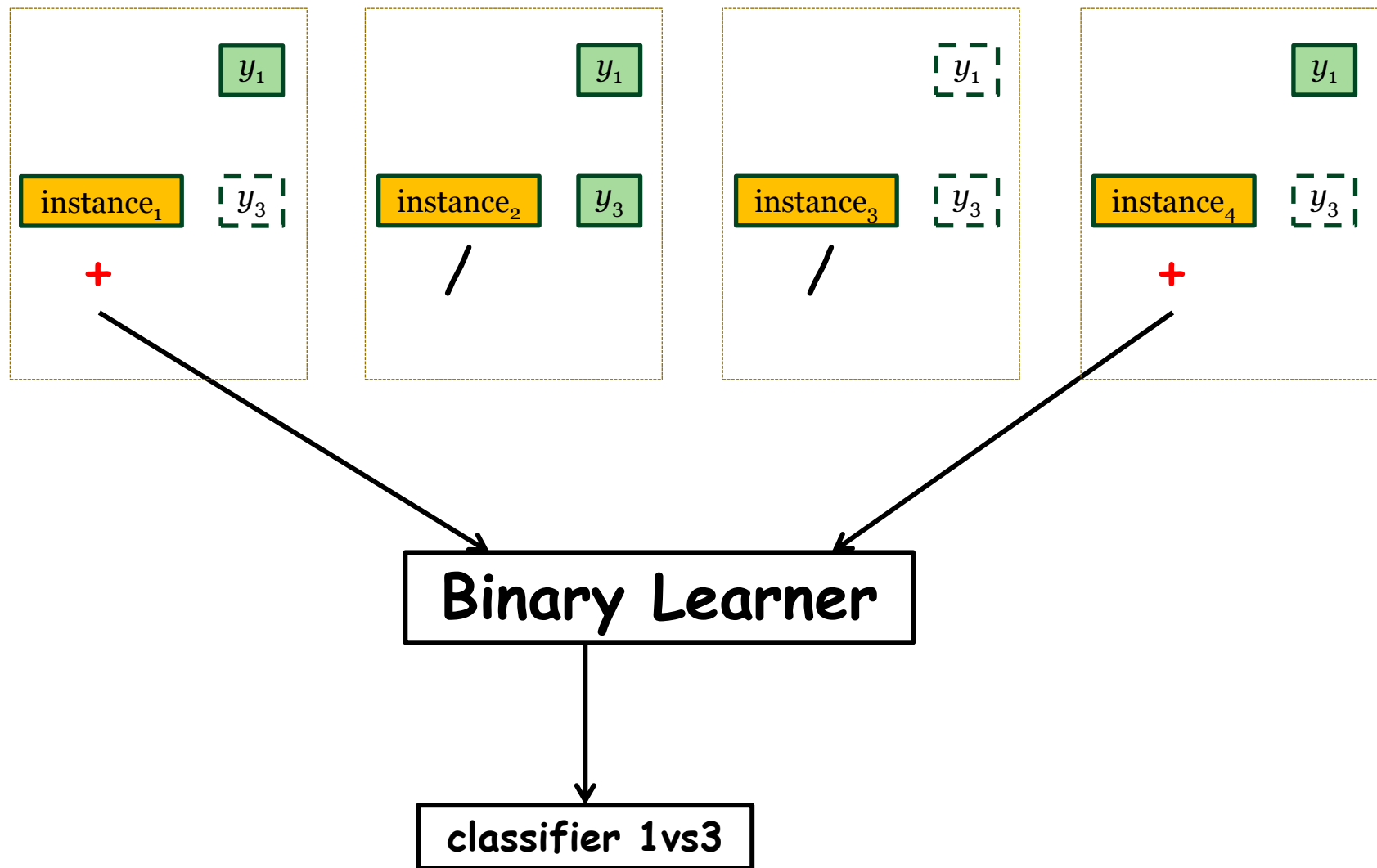
Invoke each binary model during testing, then rank all the labels based on their received votes



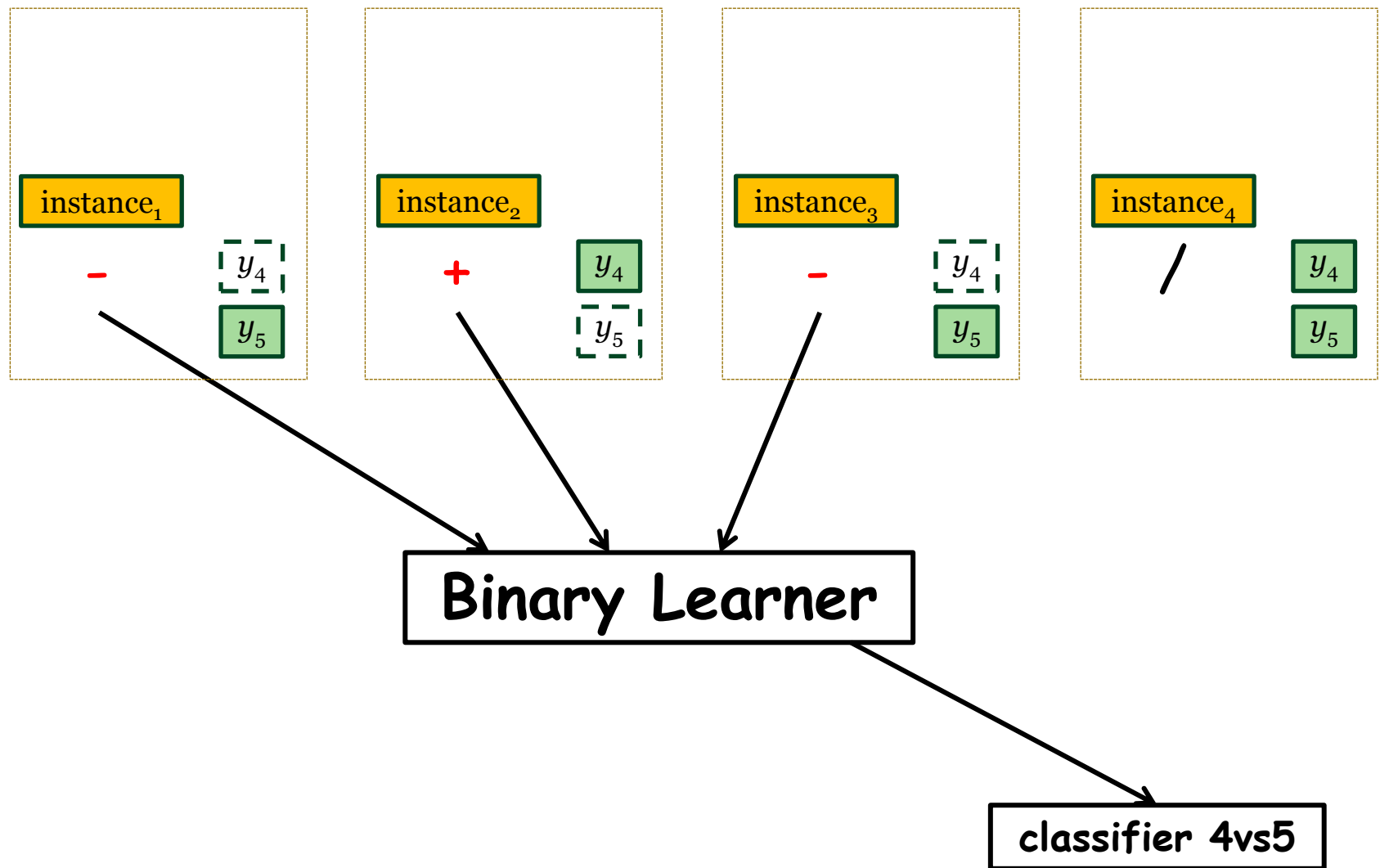
Calibrated Label Ranking - Cont.



Calibrated Label Ranking - Cont.



Calibrated Label Ranking - Cont.



Calibrated Label Ranking - Cont.

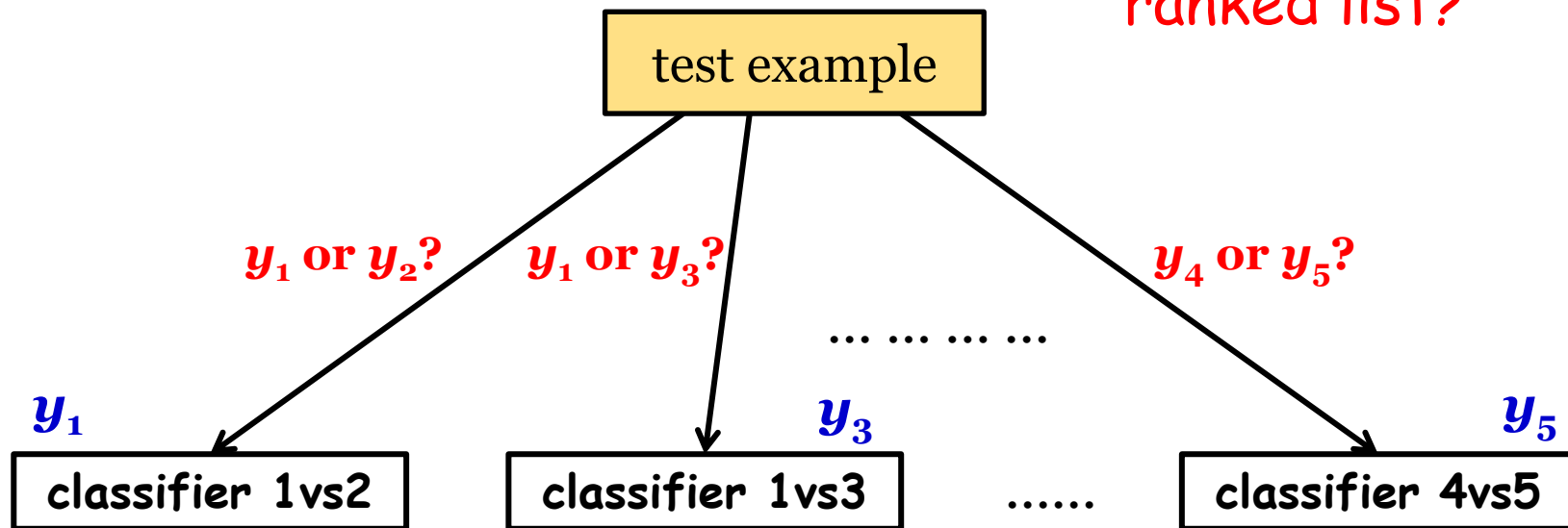
labels	votes
y_1	2
y_2	4
y_3	1
y_4	0
y_5	3

Ranking

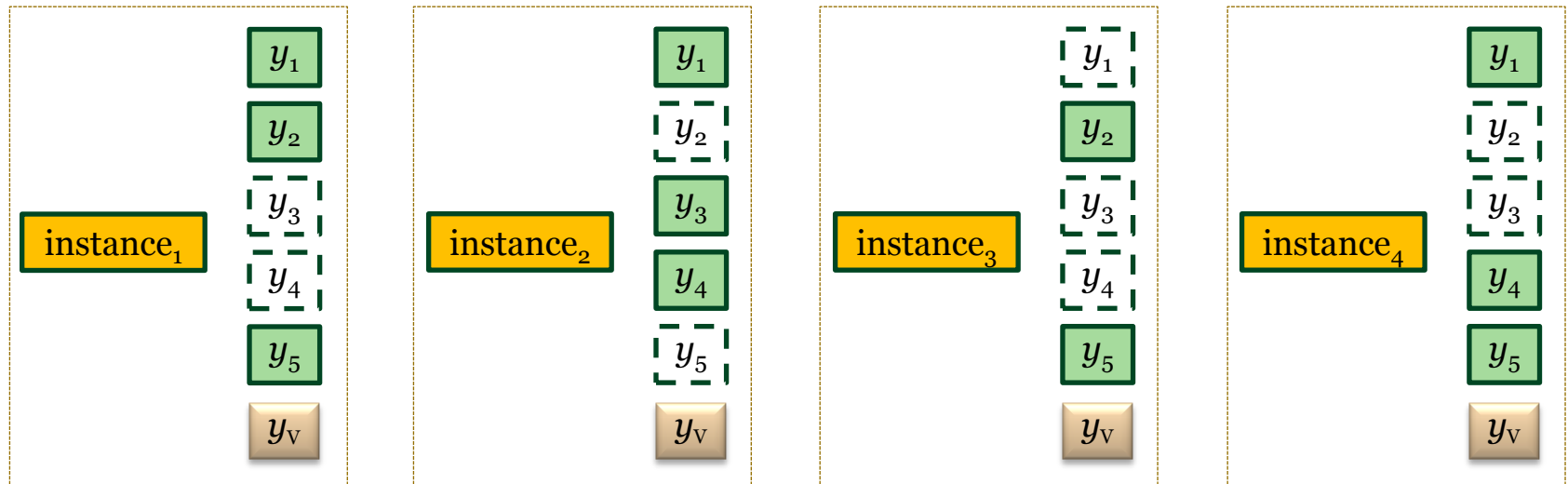
$y_2 \succ y_5 \succ y_1 \succ y_3 \succ y_4$



But, where should we bi-partition the ranked list?

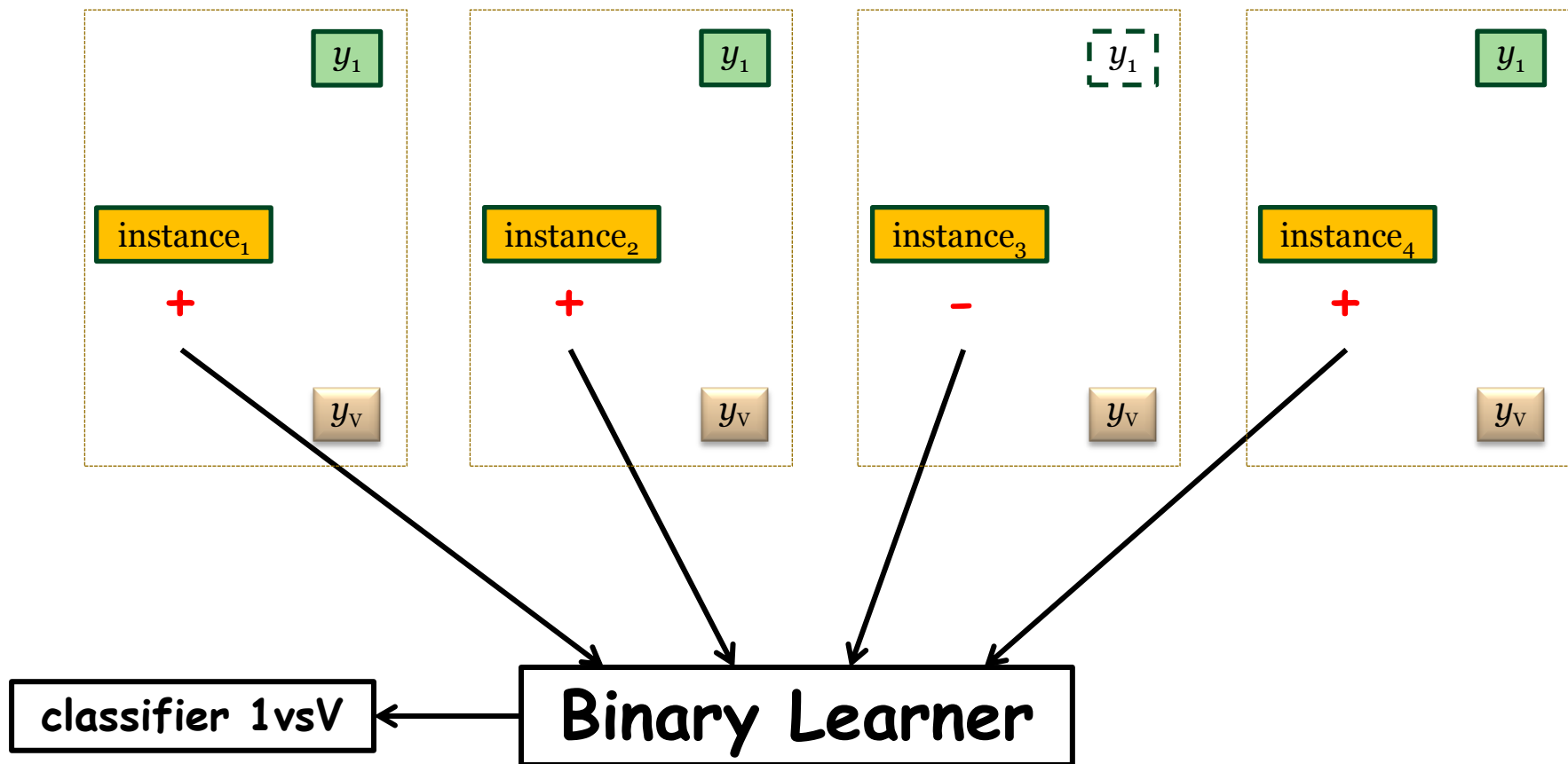


Calibrated Label Ranking - Cont.

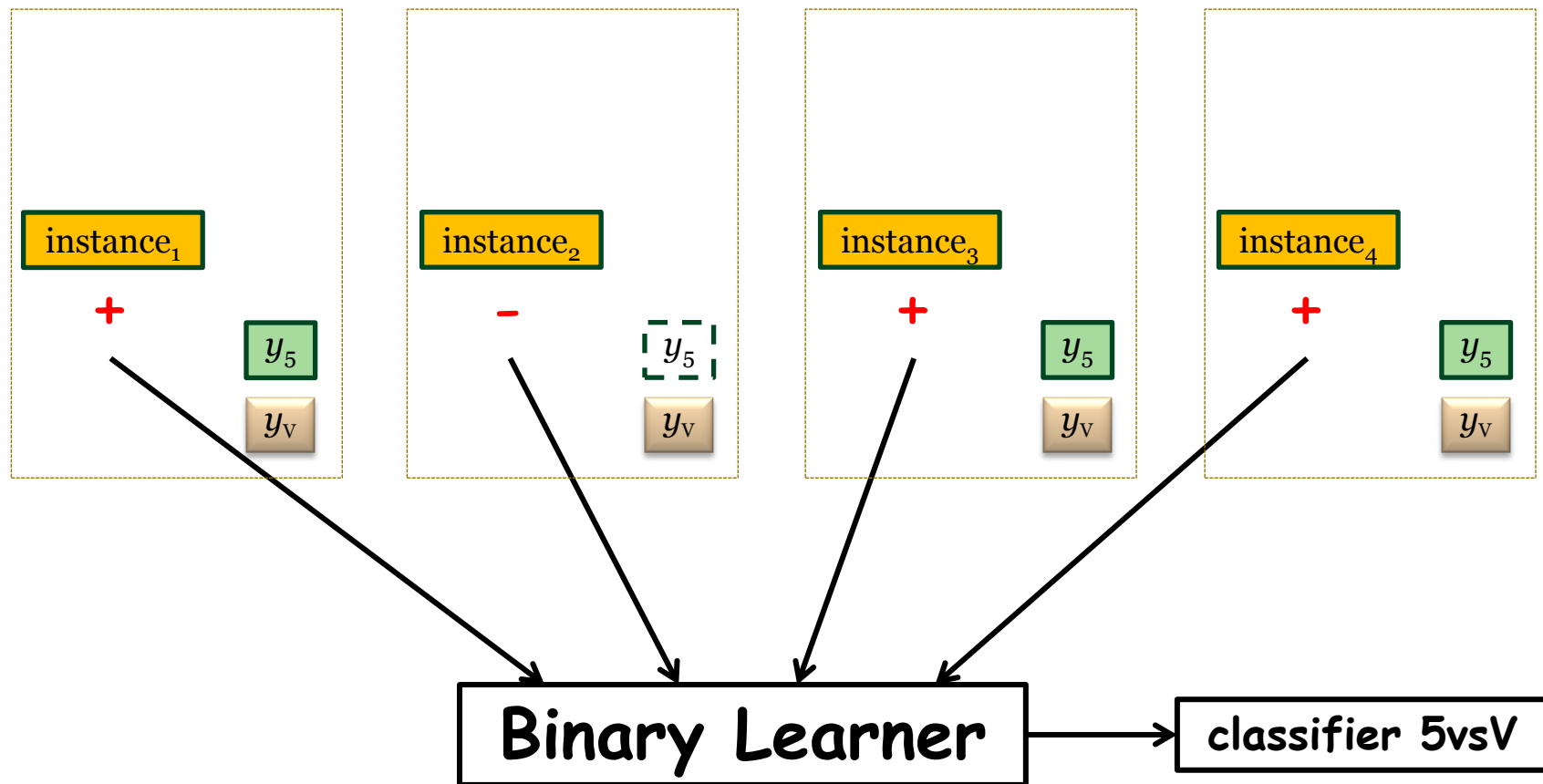


Add a virtual label y_v to each of the training examples, which serves as an artificial splitting point between relevant and irrelevant labels

Calibrated Label Ranking - Cont.



Calibrated Label Ranking - Cont.



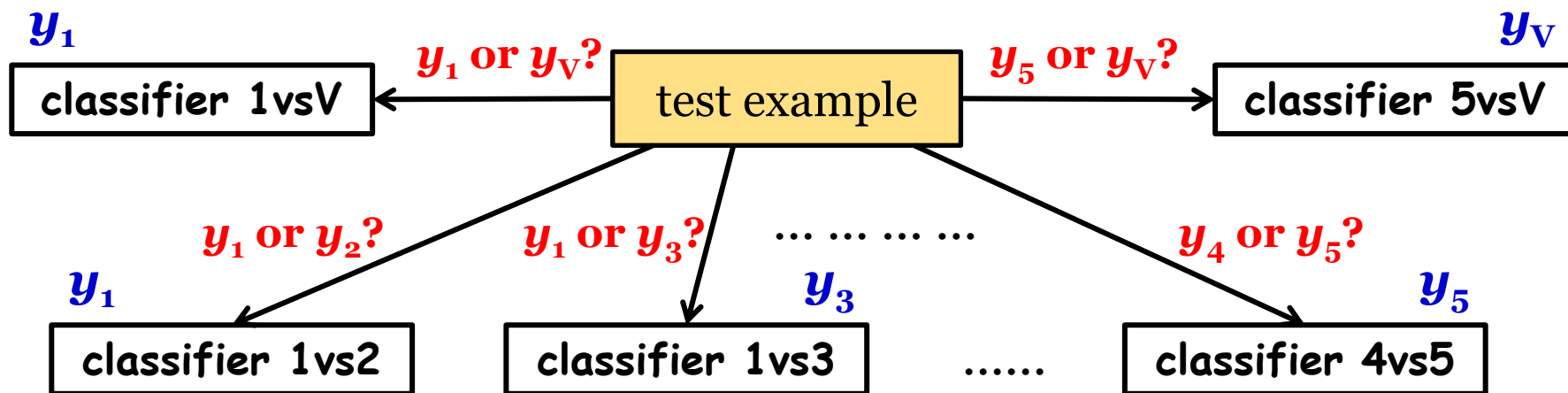
Calibrated Label Ranking - Cont.

labels	votes
y_1	2
y_2	5
y_3	1
y_4	0
y_5	4
y_V	3

Calibrated
Ranking

$y_2 \succ y_5 \succ y_V \succ y_1 \succ y_3 \succ y_4$

bi-partition
point



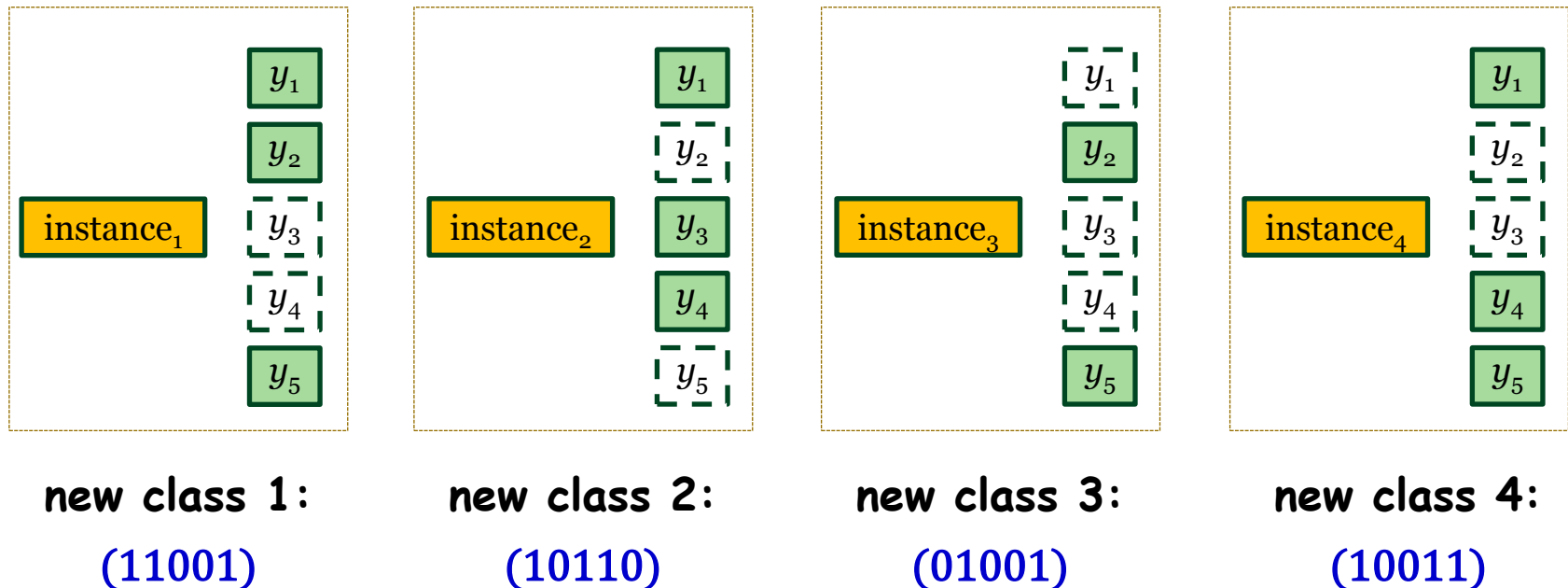
Random k -Labelsets [Tsoumakas & Vlahavas, ECML'07]

Basic Idea

transform MLL into an ensemble of single-label multi-class problems

Label Powerset (LP)

Treat each label set appearing in training set as a new class



Random k -Labelsets - Cont.

Basic Idea

transform MLL into an ensemble of single-label multi-class problems

Label Powerset (LP)

Upper bound of # new classes: $\min(m, 2^q)$

When q is large:

- Limited (even single) number of training examples per new class

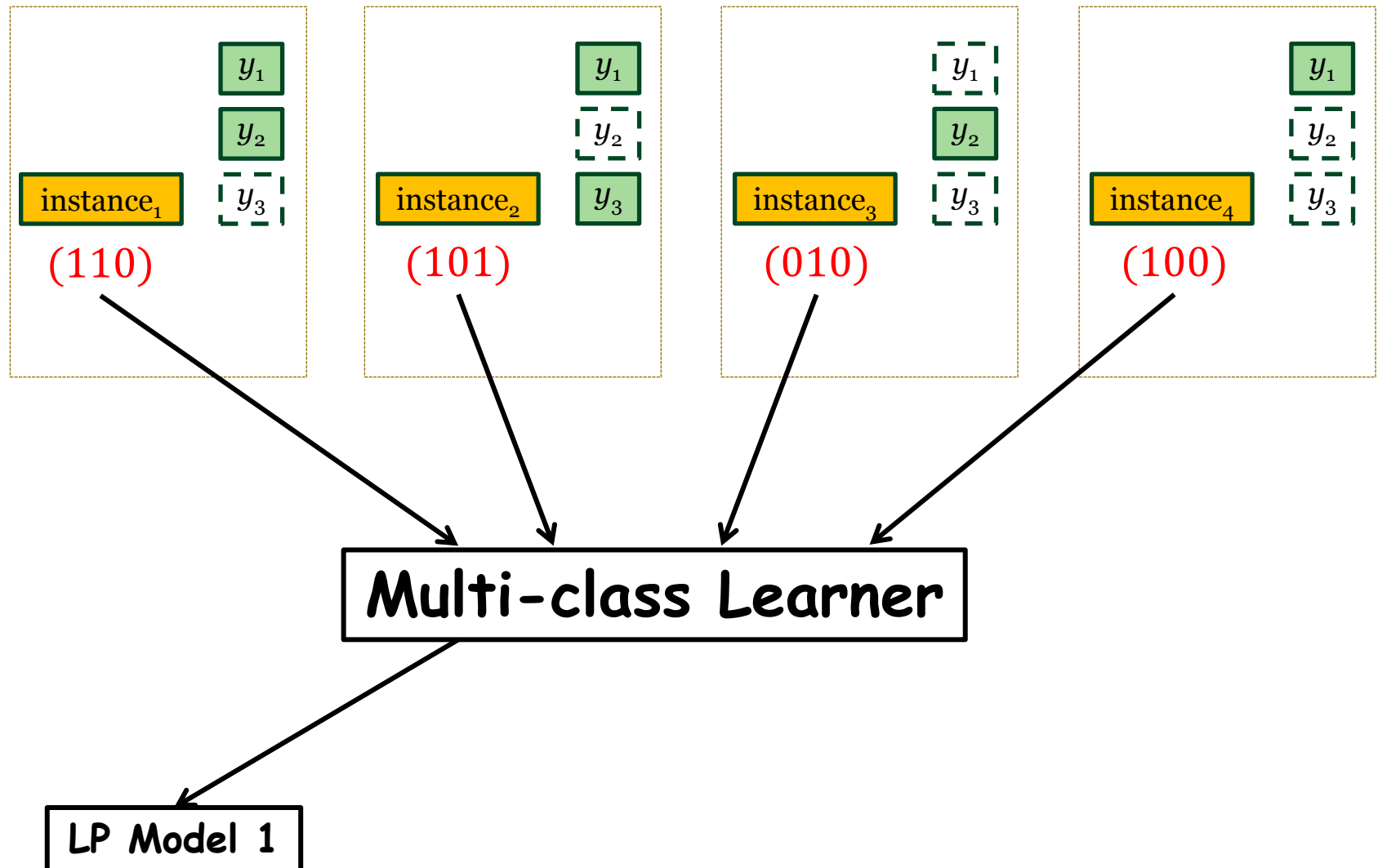
- High complexity, incapable of predicting unseen label sets

k -Labelsets

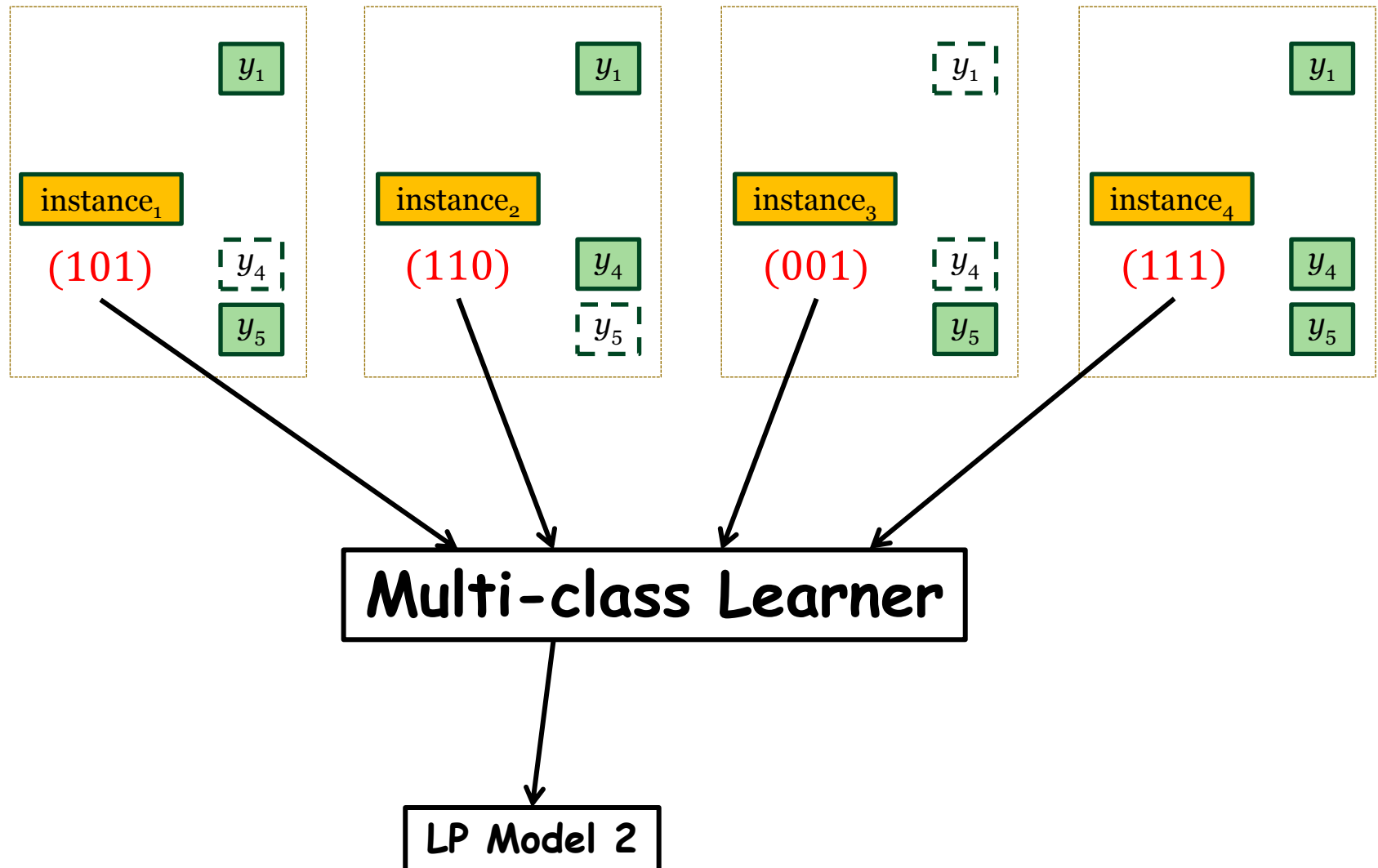
Randomly pick a subset of k labels (e.g. $k=3$), and invoke the LP method

Build an ensemble of LP models, and predict by voting and thresholding

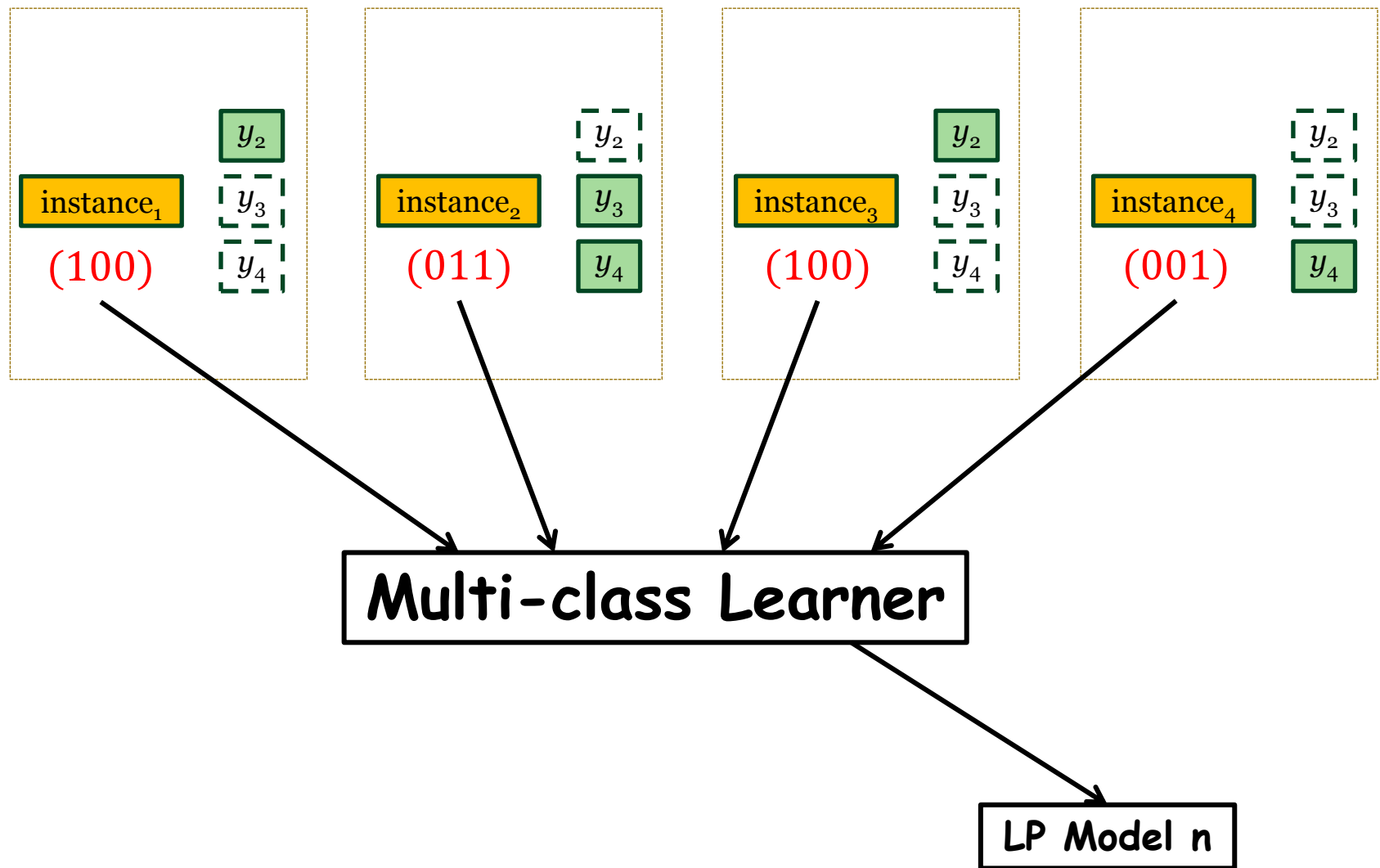
Random k -Labelsets - Cont.



Random k -Labelsets - Cont.

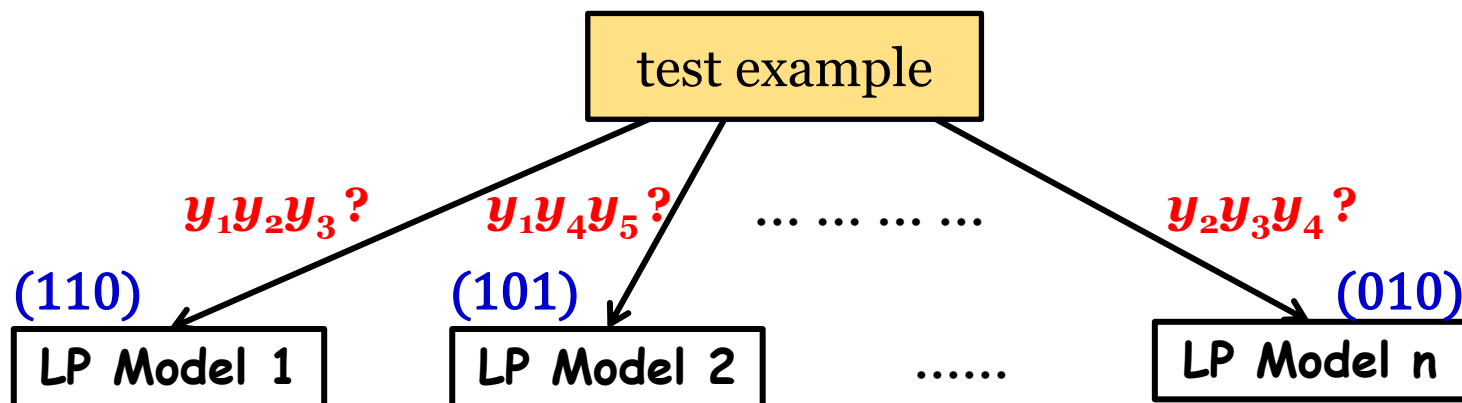
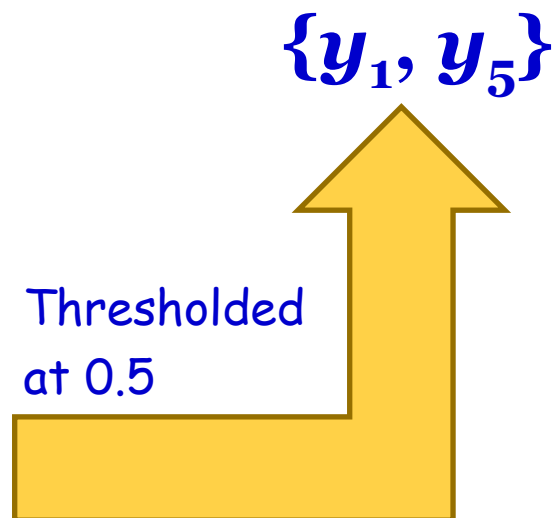


Random k -Labelsets - Cont.



Random k -Labelsets - Cont.

LP Model	k -labelsets	Prediction				
		y_1	y_2	y_3	y_4	y_5
h_1	$\{y_1, y_2, y_3\}$	1	1	0	-	-
h_2	$\{y_1, y_4, y_5\}$	1	-	-	0	1
h_3	$\{y_2, y_4, y_5\}$	-	0	-	1	1
h_4	$\{y_2, y_3, y_4\}$	-	0	0	0	-
averaged voting		2/2	1/3	0/2	1/3	2/2



Other Problem Transformation Style Methods

■ First-order

- [Schapire & Singer, MLJ00; Comité et al., MLDM'03;]

■ Second-order

- [Schapire & Singer, MLJ00; Brinker & Hüllermeier, NIPS'05w; Brinker et al., ECAI'06;]

■ High-order

- [Godbole & Sarawagi, PAKDD'04; Ji et al., KDD'08; Read et al., ICDM'08; Read et al., ECML'09; Dembczyński et al., ICML'10;]



Outline

- Multi-Label Learning (MLL)
- Learning Algorithms
 - A Brief Taxonomy
 - Problem Transformation Methods
 - Algorithm Adaptation Methods
- Advanced Topics
- Resources



ML- k NN [Zhang & Zhou, PRJ07]

Basic Idea

k NN rule + MAP reasoning with neighbors' labeling information

Algorithmic Setting

\mathcal{N} : the k nearest neighbors identified

C_j : # examples in \mathcal{N} having the j -th label

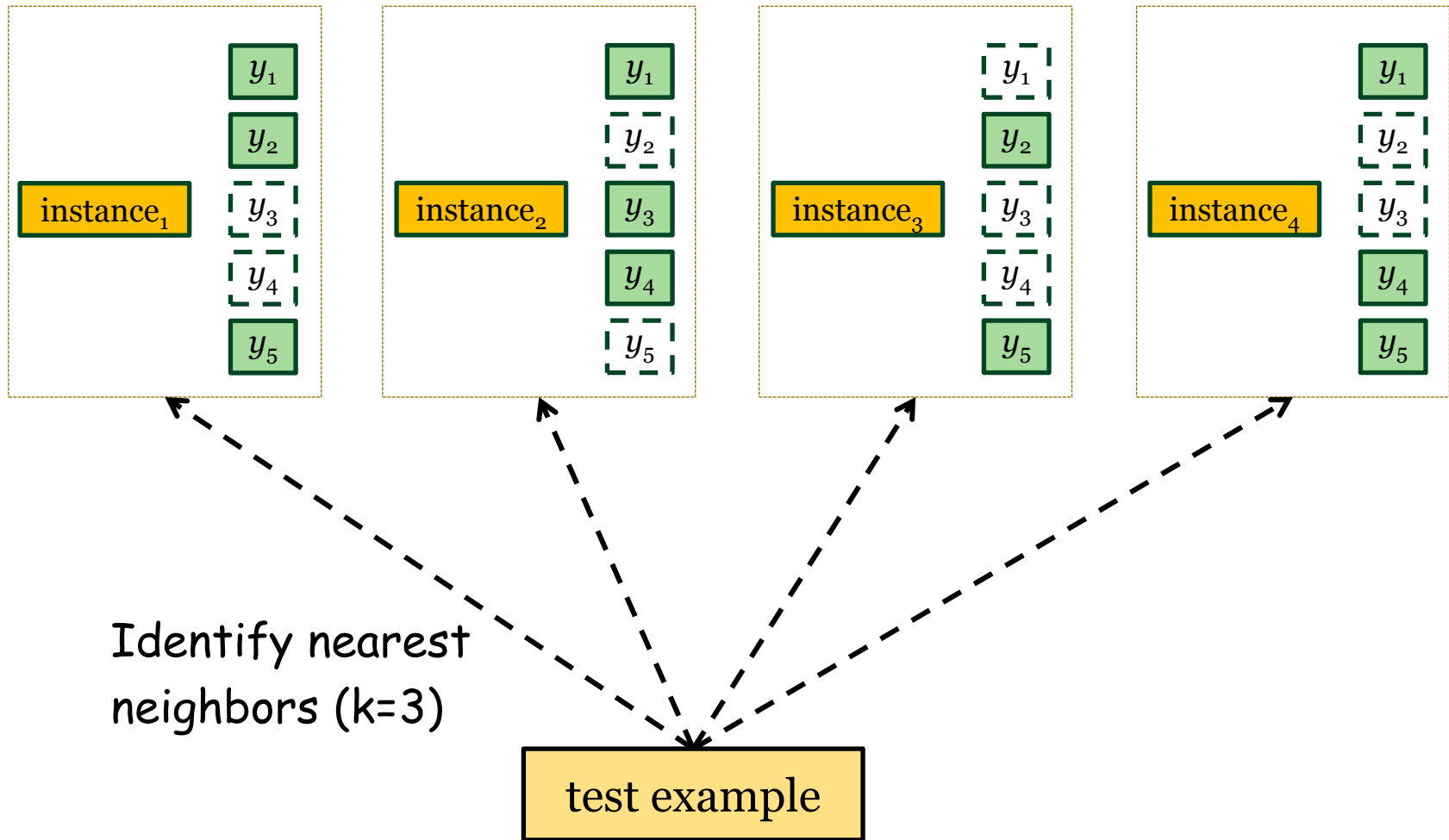
H_j : the event that test example having the j -th label

$$\mathbb{P}(H_j \mid C_j) > \mathbb{P}(\neg H_j \mid C_j) \quad \longrightarrow \quad H_j \text{ holds}$$

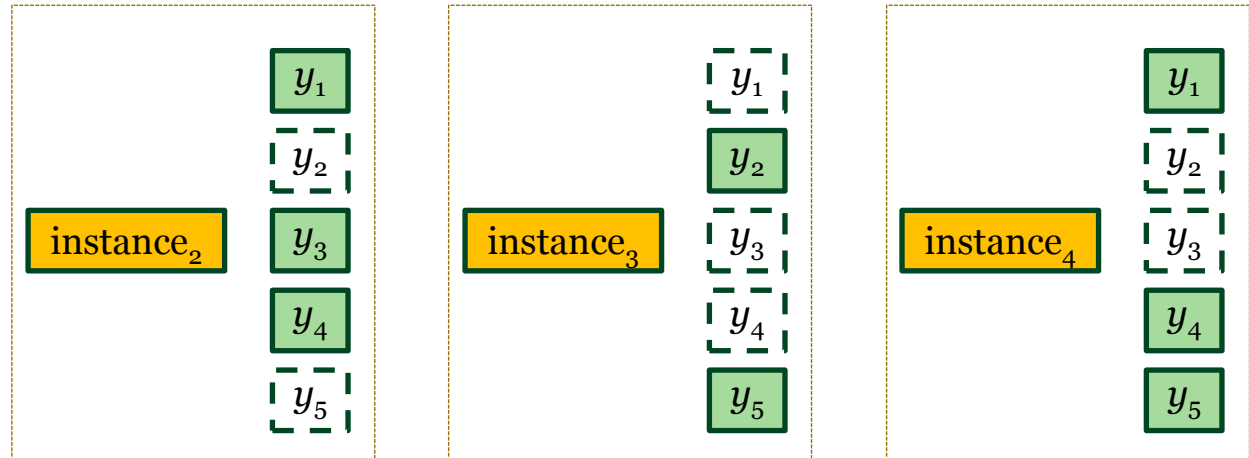
$$j = 1, 2, \dots, q$$



ML- k NN - Cont.



ML- k NN - Cont.



ML- k NN - Cont.

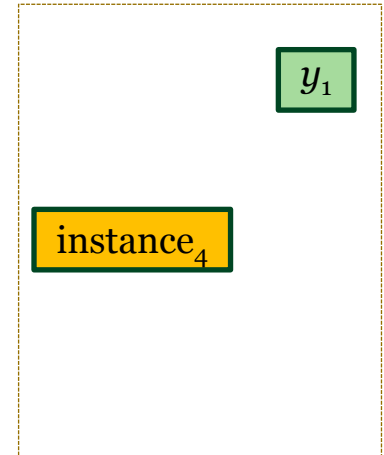
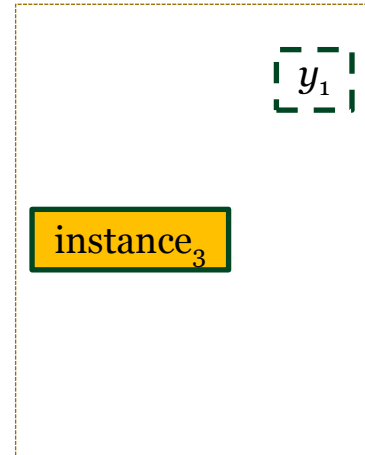
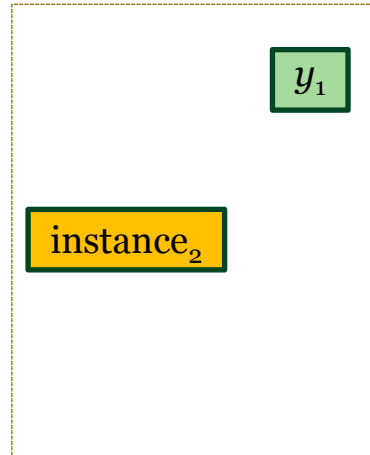
To Compute:

$$\mathbb{P}(H_1)$$

$$\mathbb{P}(\neg H_1)$$

$$\mathbb{P}(C_1 = 2 \mid H_1)$$

$$\mathbb{P}(C_1 = 2 \mid \neg H_1)$$



Does H_1 hold?

$$\frac{\mathbb{P}(H_1 \mid C_1)}{\mathbb{P}(\neg H_1 \mid C_1)} = \frac{\mathbb{P}(H_1) \cdot \mathbb{P}(C_1 \mid H_1)}{\mathbb{P}(\neg H_1) \cdot \mathbb{P}(C_1 \mid \neg H_1)}$$

estimate from
training set by
"frequency counting"

ML- k NN - Cont.

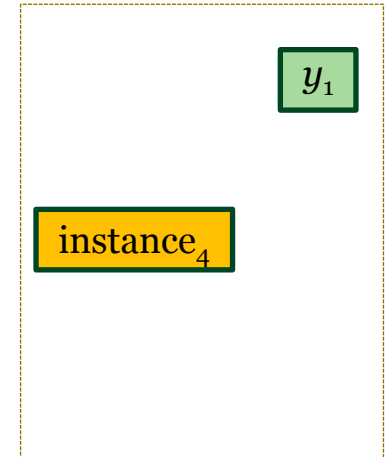
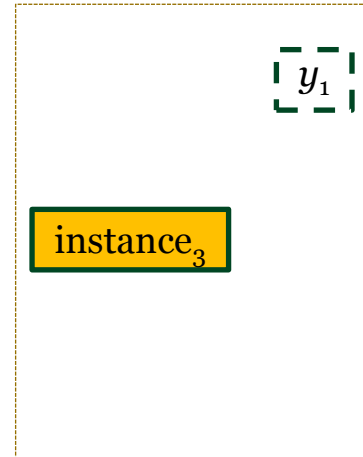
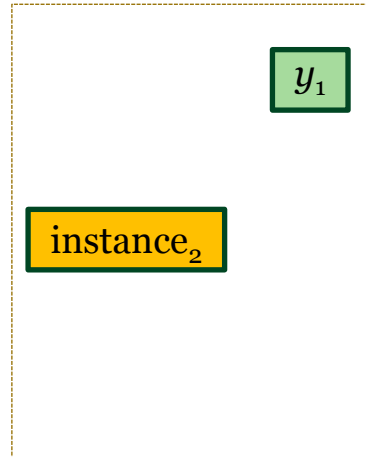
To Compute:

$$\mathbb{P}(H_1)$$

$$\mathbb{P}(\neg H_1)$$

$$\mathbb{P}(C_1 = 2 \mid H_1)$$

$$\mathbb{P}(C_1 = 2 \mid \neg H_1)$$



Does H_1 hold? ✓ true

$$\frac{\mathbb{P}(H_1 \mid C_1)}{\mathbb{P}(\neg H_1 \mid C_1)} = \frac{\mathbb{P}(H_1) \cdot \mathbb{P}(C_1 \mid H_1)}{\mathbb{P}(\neg H_1) \cdot \mathbb{P}(C_1 \mid \neg H_1)} > 1$$

estimate from
training set by
"frequency counting"

Extension to high-order [Cheng & Hüllermeier, ECML'09]

$$\frac{\mathbb{P}(H_1 \mid C_1)}{\mathbb{P}(\neg H_1 \mid C_1)} \longrightarrow \frac{\mathbb{P}(H_1 \mid C_1, C_2, \dots, C_q)}{\mathbb{P}(\neg H_1 \mid C_1, C_2, \dots, C_q)}$$

Rank-SVM [Elisseeff & Weston, NIPS'02]

Basic Idea

Train a collection of SVMs (one per label) by minimizing ranking loss

Algorithmic Setting

(\mathbf{w}_j, b_j) : the linear classifier for the j -th label

$\langle \mathbf{w}_j, \mathbf{x} \rangle + b_j > 0 \iff j$ -th label being relevant

Object function to minimize:

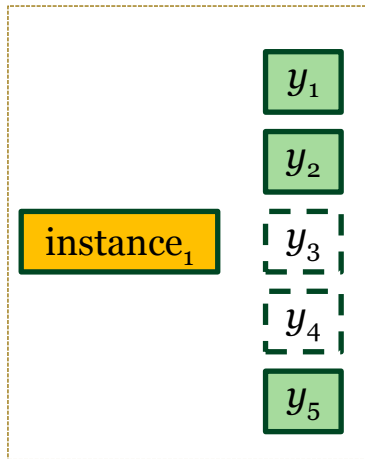
$$\sum_{j=1}^q \|\mathbf{w}_j\|^2 + C \cdot \sum_{i=1}^m \frac{1}{|Y_i| |\bar{Y}_i|} \sum_{(j,k) \in Y_i \times \bar{Y}_i} \text{hinge}(\langle \mathbf{w}_j - \mathbf{w}_k, \mathbf{x}_i \rangle + b_j - b_k)$$

model
complexity

empirical
ranking loss



Rank-SVM - Cont.



$$(w_1, b_1)$$

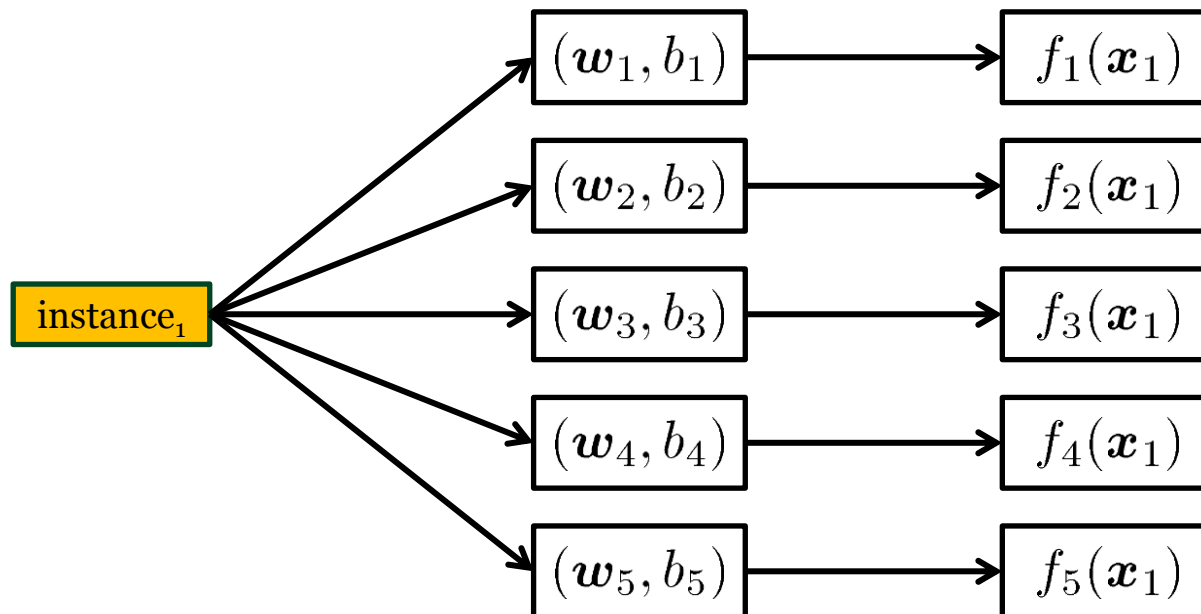
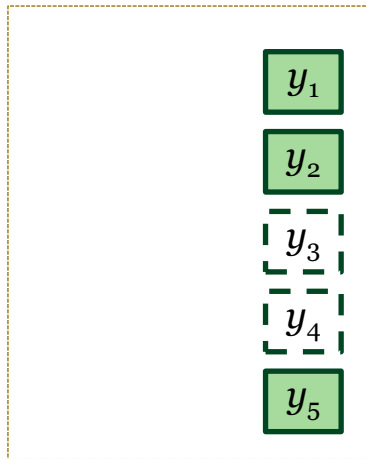
$$(w_2, b_2)$$

$$(w_3, b_3)$$

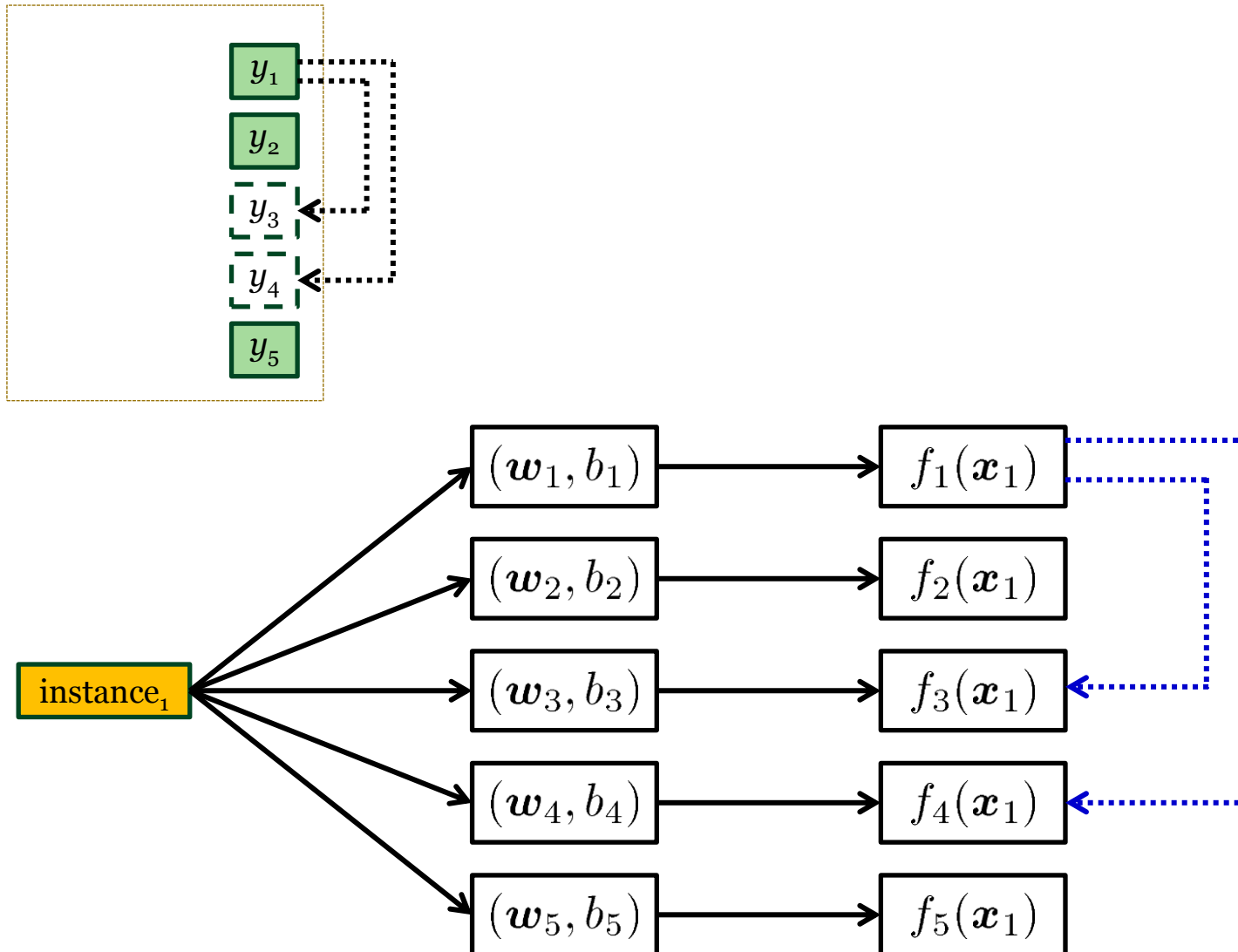
$$(w_4, b_4)$$

$$(w_5, b_5)$$

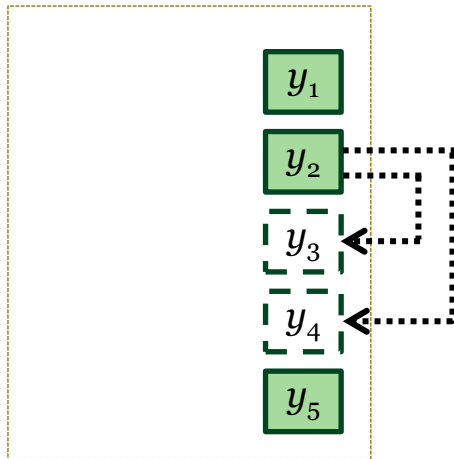
Rank-SVM - Cont.



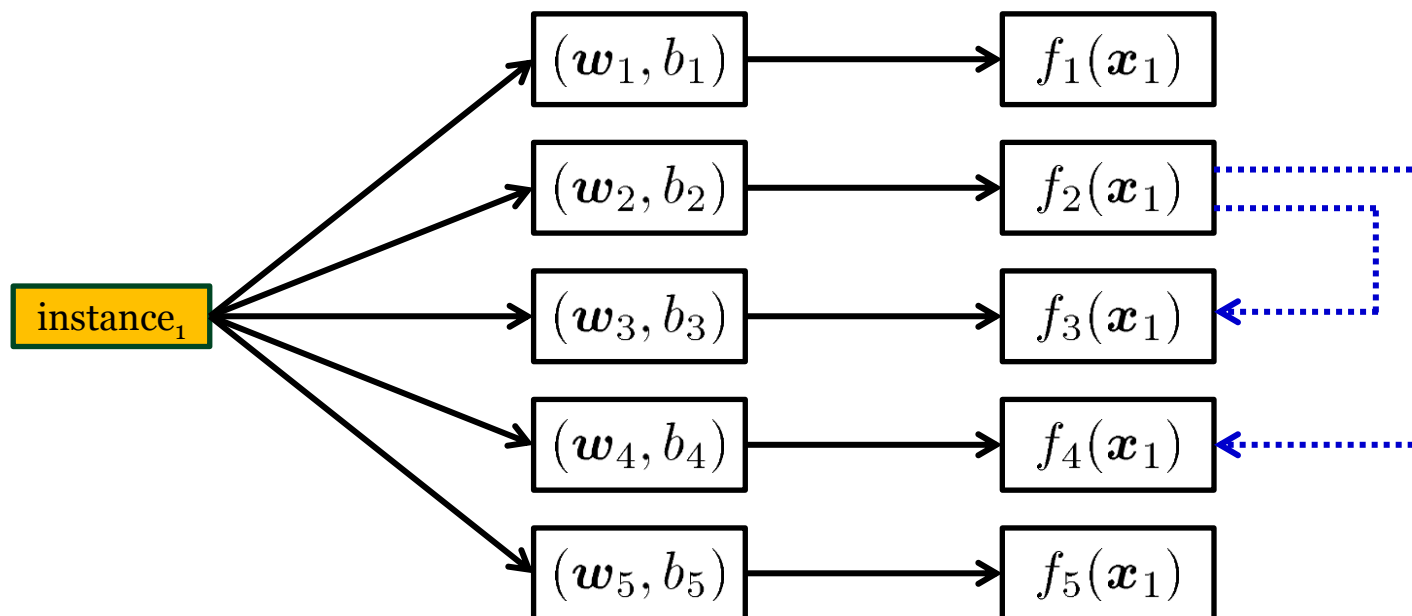
Rank-SVM - Cont.



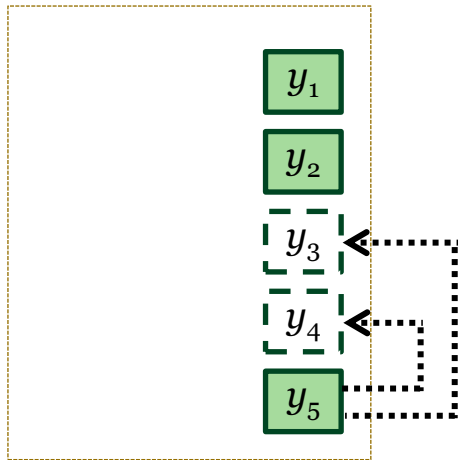
Rank-SVM - Cont.



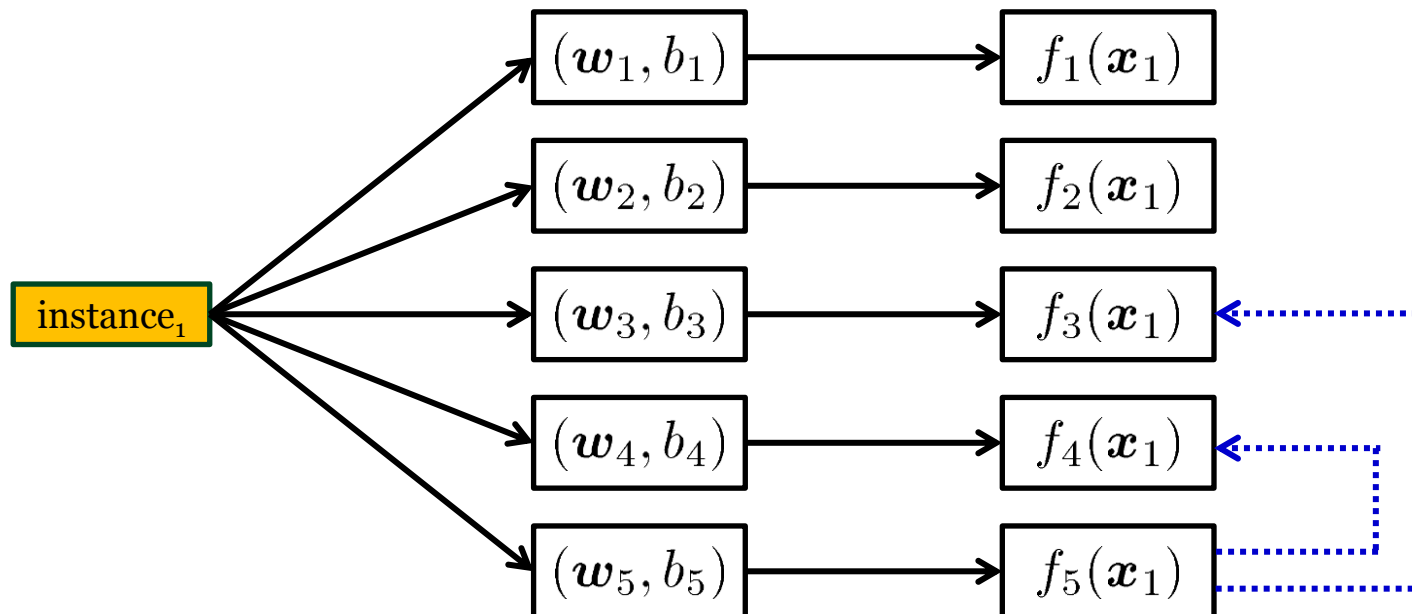
Rank-SVM seeks to minimize the ranking loss metric, which is equivalent to impose the constraint that outputs on relevant labels should be larger than those on irrelevant labels



Rank-SVM - Cont.



Objective function optimized within QP formulation by introducing slack variables, and then solved in its dual form by incorporating kernel trick



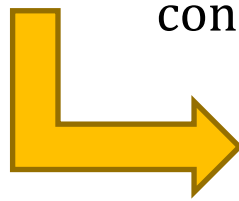
LEAD [Zhang & Zhang, KDD'10]

Basic Idea

Model dependencies among labels via Bayesian network structure

Algorithmic Setting

$p(\mathbf{y} \mid \mathbf{x})$: the joint probabilities of all labels $\mathbf{y} = (y_1, y_2, \dots, y_q) \in \{0, 1\}^q$
conditioned on the feature vector \mathbf{x}

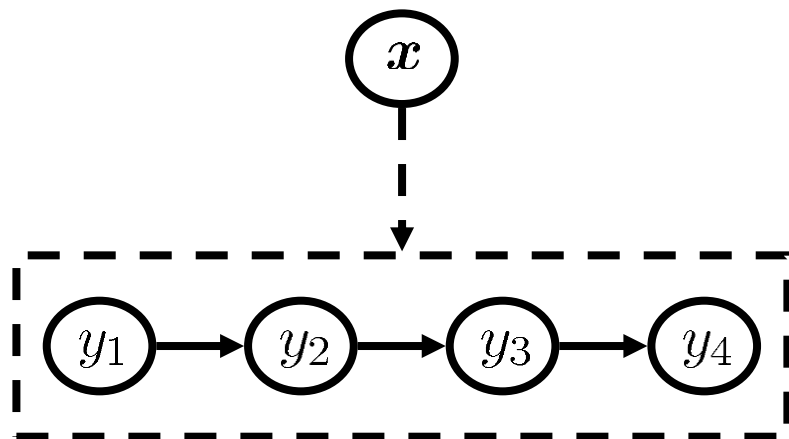

$$h(\mathbf{x}) = \arg \min_{\mathbf{y}} p(\mathbf{y} \mid \mathbf{x})$$

Encode $p(\mathbf{y} \mid \mathbf{x})$ via Bayesian network structure:

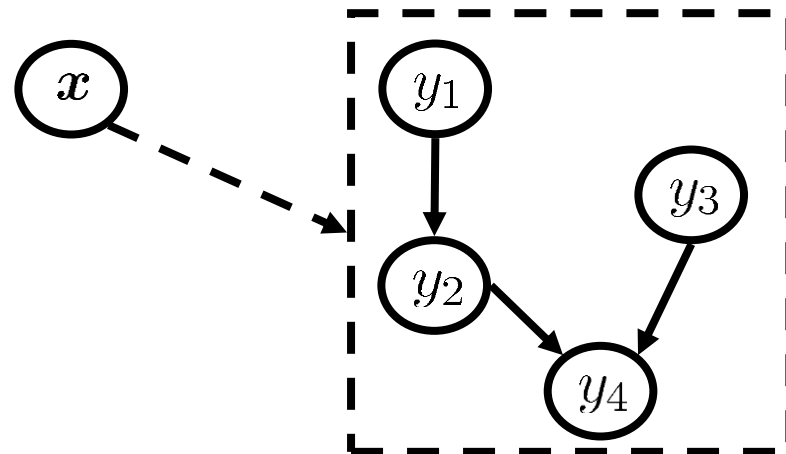
$$p(\mathbf{y} \mid \mathbf{x}) = \prod_{j=1}^q p(y_j \mid \mathbf{pa}_j, \mathbf{x})$$

LEAD - Cont.

Examples of modeling $p(\mathbf{y}|\mathbf{x})$ via BN structure



$$\begin{aligned} p(\mathbf{y} \mid \mathbf{x}) &= p(y_1 \mid \mathbf{x}) \times \\ &\quad p(y_2 \mid y_1, \mathbf{x}) \times \\ &\quad p(y_3 \mid y_2, \mathbf{x}) \times \\ &\quad p(y_4 \mid y_3, \mathbf{x}) \end{aligned}$$



$$\begin{aligned} p(\mathbf{y} \mid \mathbf{x}) &= p(y_1 \mid \mathbf{x}) \times \\ &\quad p(y_2 \mid y_1, \mathbf{x}) \times \\ &\quad p(y_3 \mid \mathbf{x}) \times \\ &\quad p(y_4 \mid y_2, y_3, \mathbf{x}) \end{aligned}$$

LEAD - Cont.

Difficulties in modeling $p(\mathbf{y}|\mathbf{x})$:

Existing BN learning techniques not directly applicable

- ❑ **Mixed type of variables**: the labels are discrete while the features are continuous
- ❑ **Prohibitively high complexity**: the input dimensionality, i.e. the number of features, may be too large (e.g. text data)

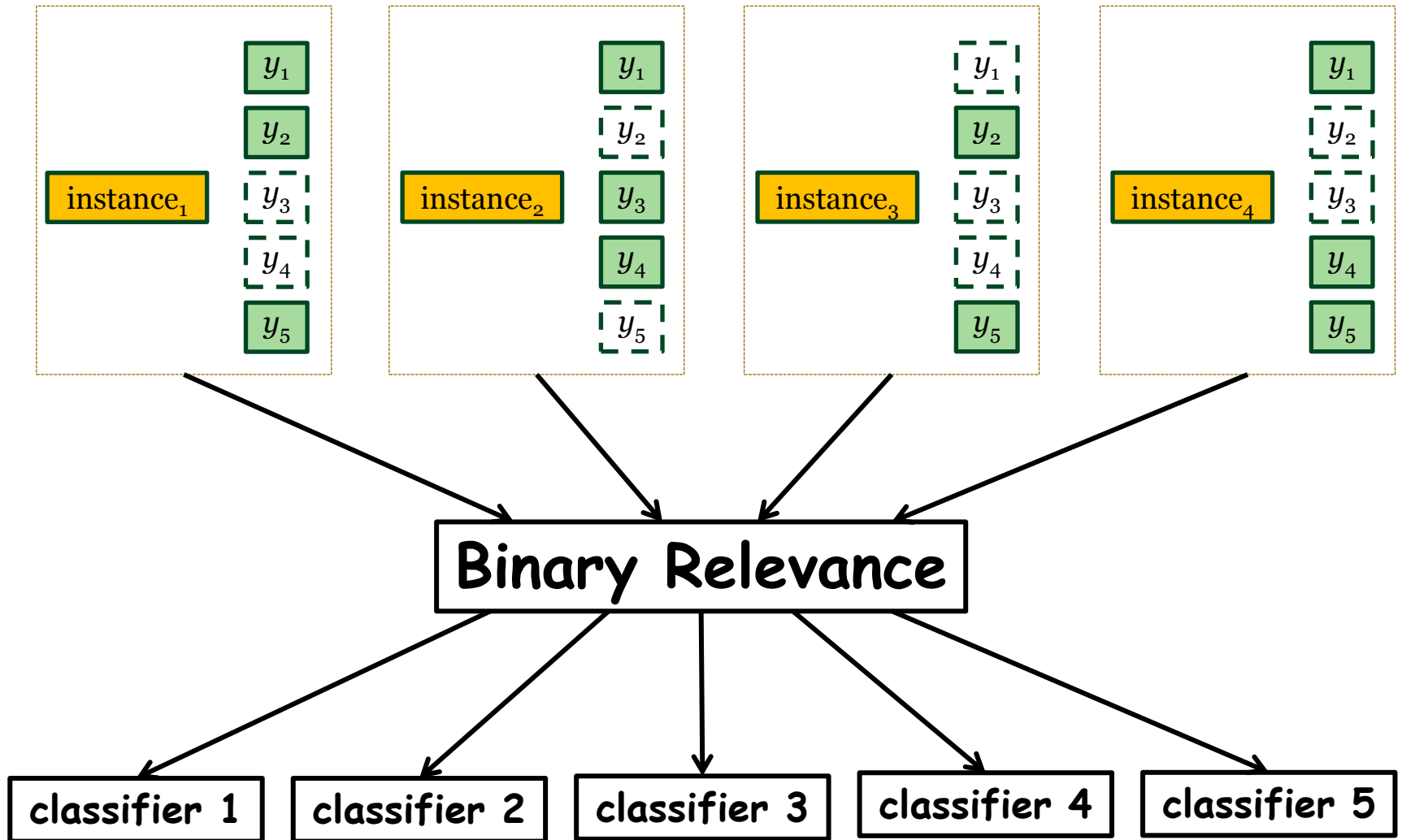
Solution:

“Eliminate” the effects of the feature set \mathbf{x} on all labels

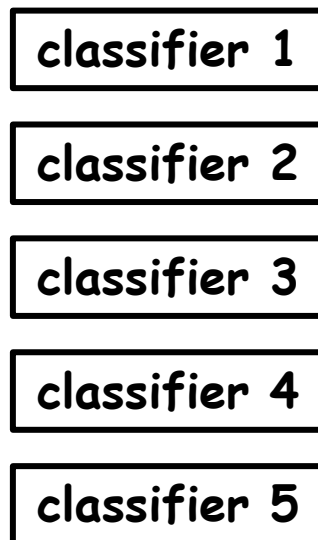
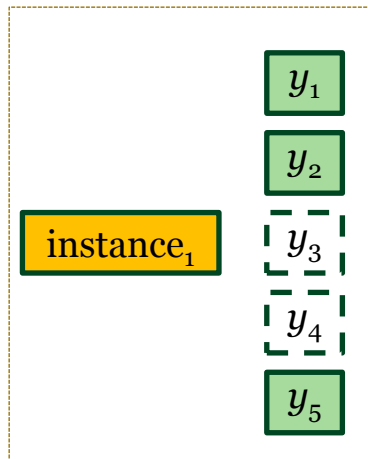
$$e_j = y_j - f_j(\mathbf{x}) \quad (1 \leq j \leq q)$$



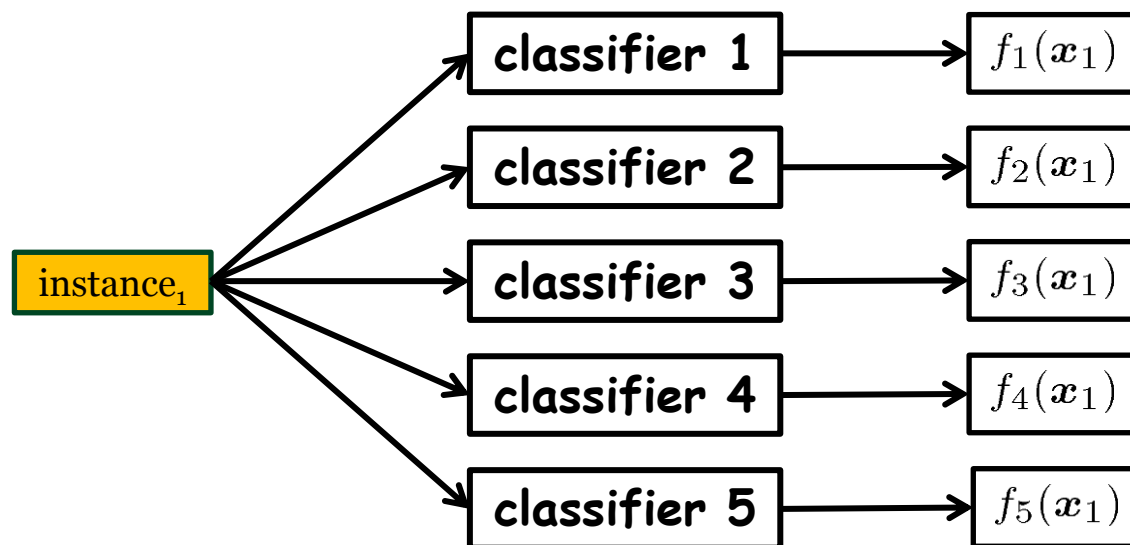
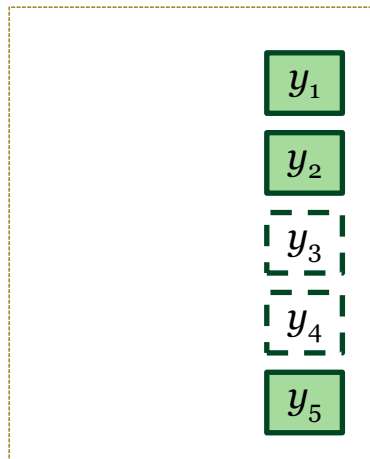
LEAD - Cont.



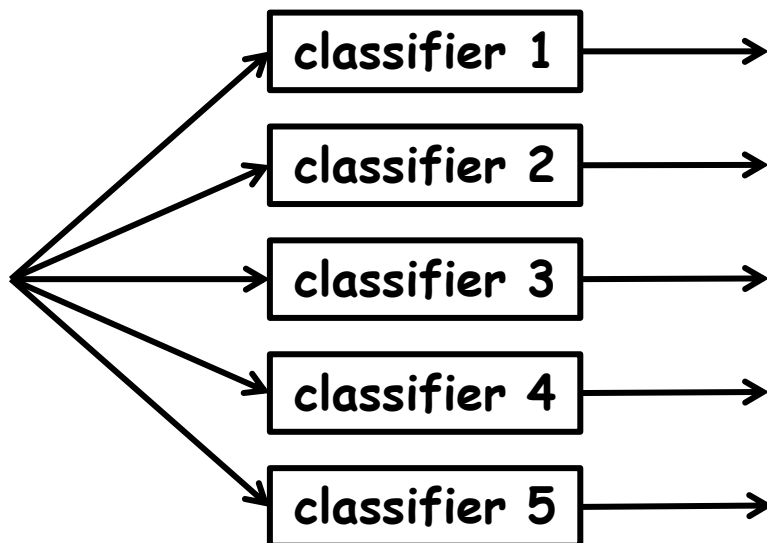
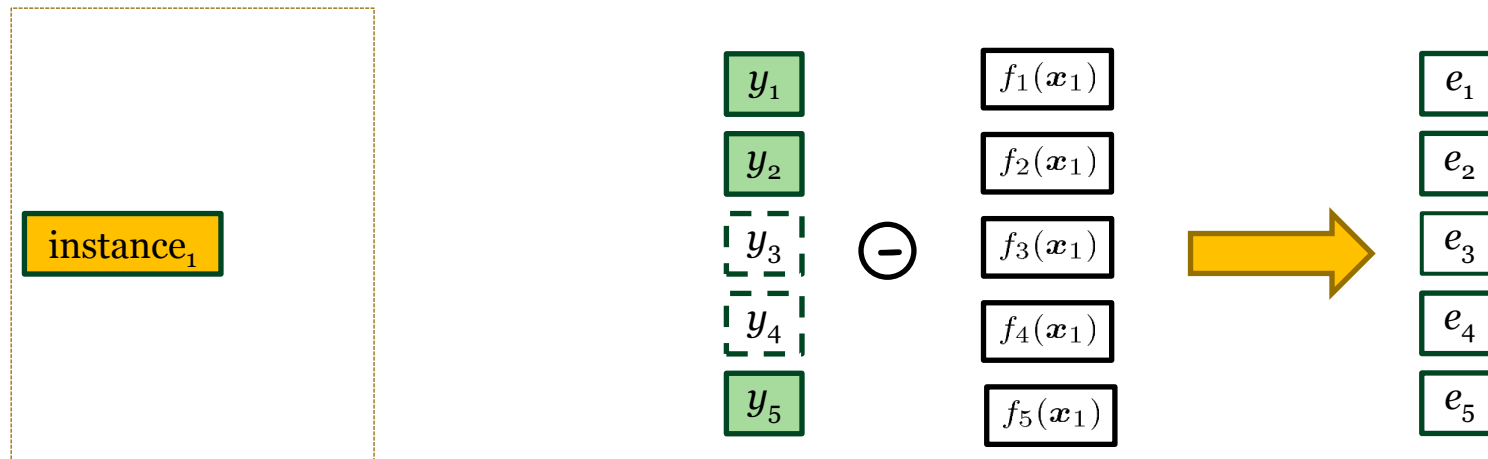
LEAD - Cont.



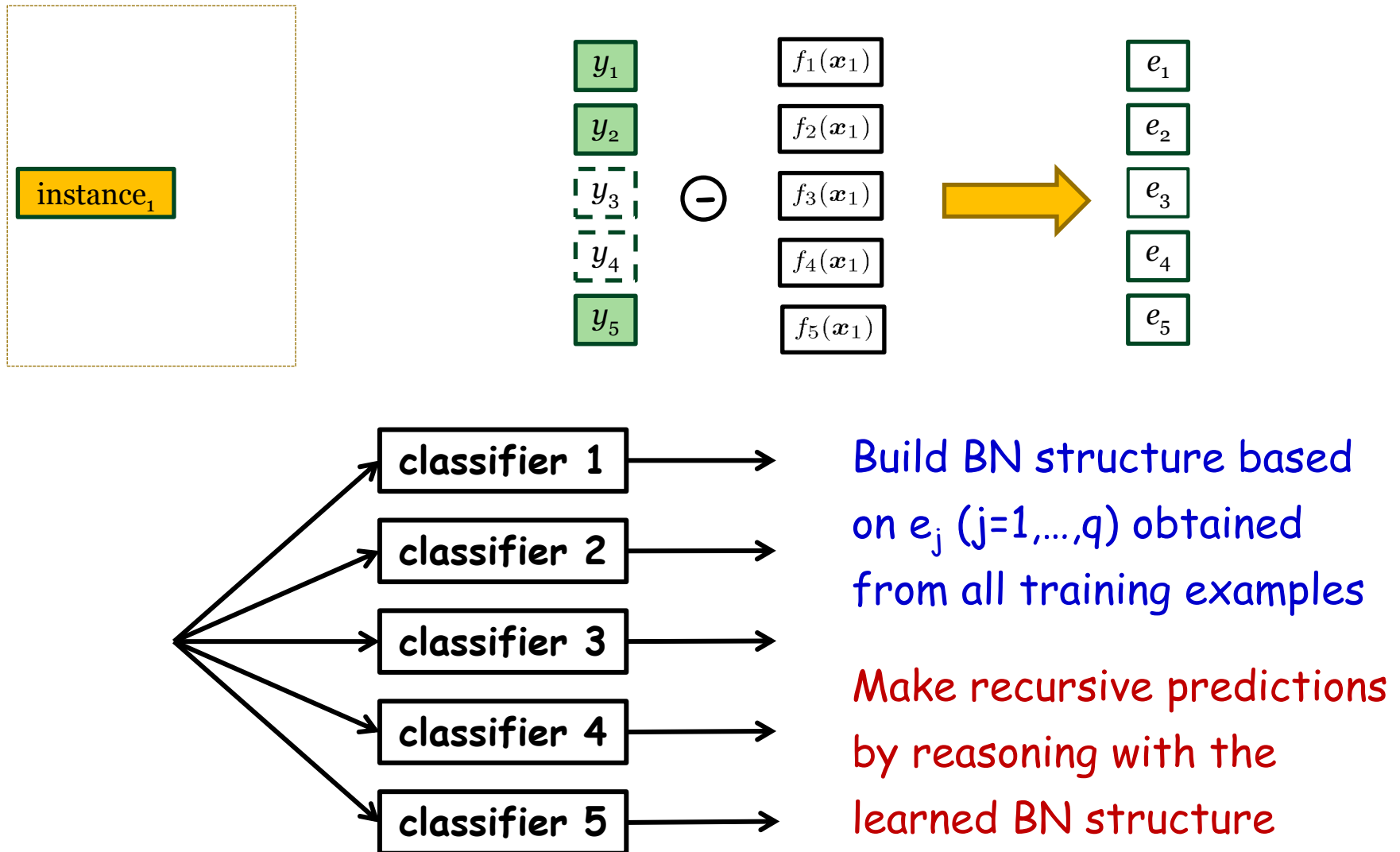
LEAD - Cont.



LEAD - Cont.



LEAD - Cont.



Other Algorithm Adaptation Style Methods

■ First-order

- [McCallum, AAAI'99w; Clare & King, PKDD'01; Spyromitros et al., SETN'08; Wang et al., AAAI'10;]

■ Second-order

- [Ueda & Saito, NIPS'03; Ghamrawi & McCallum, CIKM'05; Zhu et al., SIGIR'05; Zhang & Zhou, TKDE06; Brinker & Hüllermeier, IJCAI'07; Qi et al., ACM MM'07;]

■ High-order

- [Yan et al., KDD'07; Cheng & Hüllermeier, ECML'09;]



Outline

- Multi-Label Learning (MLL)
- Learning Algorithms
- Advanced Topics
 - Noisy/Weak Label
 - Unlabeled Data
 - Dimensionality Reduction
 - Large-Scale
- Resources



Advanced Topics I:

Noisy/Weak Label

Noisy Label [Jin & Ghahramani, NIPS'03; Ozonat & Young, KDD'09]

The set of labels assigned to each example may not be fully valid, e.g. some labels may be wrongly assigned due to mistakes of human labellers

Weak Label [Sun et al., AAAI'10]

The absence of some labels do not necessarily mean they are invalid for the example, e.g. only a “partial” set of proper labels is assigned by the human labeller



Advanced Topics II:

Unlabeled Data

Active MLL [Brinker, GFKL'05; Qi et al., TPAMI'08; Yang et al., KDD'09; Esuli & Sebastiani, ECIR'09; Singh et al., TechRep09]

Exploit unlabeled data with human intervention

Semi-supervised/Transductive MLL [Liu et al., AAAI'06; Chen et al., SDM'08]

Exploit unlabeled data by the learner automatically, without human intervention



Advanced Topics III:

Dimensionality Reduction

Filter Style [Yu et al., SIGIR'05; Zhang & Zhou, AAAI'08, TKDD10]

Conduct dimensionality reduction without resorting to specific MLL process

Wrapper Style [Ji & Ye, IJCAI'09; Qian & Davidson, AAAI'10]

Conduct dimensionality reduction and MLL simultaneously

Filter + Wrapper Style [Zhang et al., INS09]

Conduct unsupervised feature extraction, followed by supervised feature subset selection



Advanced Topics IV:

Large-Scale

Large Number of Labels [Tsoumakas et al., ECML/PKDD'08w; Hsu et al., NIPS'09; Zhang et al., SDM'10]

Most existing MLL algorithms will fail when the label space is large, e.g. $q > 50$, especially for the second-order and high-order approaches.

The labeling sparsity should be exploited.

Large Number of Examples [Hariharan et al., ICML'10]

Is it necessary to consider label correlations given large # of examples, e.g. $m > 10k$? Would the simple binary relevance strategy suffice?



Outline

- Multi-Label Learning (MLL)
- Learning Algorithms
- Advanced Topics
- Resources
 - Events
 - Active Groups
 - Data Sets & Software
 - Online Bibliography



Resources I:

Events

ECML/PKDD 2009 Tutorial

<http://www.ecmlpkdd2009.net/program/tutorials/learning-from-multi-label-data/>

Workshops

MLD'09 - <http://lpis.csd.auth.gr/workshops/mld09/>

in conjunction with *ECML/PKDD 2009*

MLD'10 - <http://cse.seu.edu.cn/conf/MLD10/>

in conjunction with *ICML/COLT 2010*

[Machine Learning Journal Special Issue](#)

<http://mlkd.csd.auth.gr/events/ml2010si.html>



Resources II:

Active Groups

- ❑ LAMDA Group at Nanjing University (led by Prof. Zhi-Hua Zhou)
- ❑ MLKD Group at Aristotle University of Thessaloniki, Greece (key MLL researcher: Dr. Grigorios Tsoumakas)
- ❑ KEBI Lab at Philipps-Universität Marburg, Germany (led by Prof. Eyke Hüllermeier)
- ❑ KE Group at Technische Universität Darmstadt, Germany (led by Prof. Johannes Fürnkranz)
- ❑ AI Lab at Arizona State University (key MLL researcher: Prof. Jieping Ye)
- ❑ ML Group at Michigan State University (led by Prof. Rong Jin)
- ❑ Media Computing Group at MSRA (led by Dr. Xian-Sheng Hua)
- ❑



Resources III:

Data Sets & Software

Data Sets

- ❑ [Multi-label data sets from MULAN project](#)
- ❑ [Multi-label data sets from LIBSVM](#)
- ❑ [Multi-label data sets from sourceforge.net](#)
- ❑

Software

- ❑ [The MULAN Library](#) (built upon [Weka](#))
- ❑ [Matlab Codes for Some MLL Algorithms](#)
- ❑



Resources IV:

Online Bibliography

Maintained at **citeulike** 

<http://www.citeulike.org/group/7105/tag/multilabel>

Currently 100+ papers on MLL

You can...

- ✓ Grab BibTex and RIS records
- ✓ View abstract and follow links to papers' full pdf
- ✓ Subscribe to the corresponding RSS feed
- ✓



Acknowledgements

Special Thanks to:

Prof. Zhi-Hua Zhou (Nanjing University)

Thanks to Collaborators:

Dr. Grigorios Tsoumakas (AUTH, Greece)

Dr. Kun Zhang (MPI, Germany)

Profs. José M. Peña & Victor Robles (UPM, Spain)

Prof. Zhi-Jian Wang (Hohai University)

.....

