Multi-Label Classification

Multilabel classification

Multilabel Classification

0
votes

0
answers

3 views

Can we get (Percentage ratio, eg (detected object) person obtains 60% of the image) from Output of Image segmentation using Tensorflow

I am trying to read what percent of the image the detected object covers . I've already tried Tensorflow object detection and its working great, but due to changing of requirements now I want the ...

tensorflow | object-detection | image-segmentation | tensorflow-datasets | tensorflow-lite

asked 3 mins ago

Karan Gada
15 ● 6

0
votes

0
answers

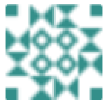
10 views

Getting “Unexpected reserved word” when doing await in export module?

In my export module, I want to use await export default { template : await (await fetch("template.html")).text(); } but this doesn't work. It gets Unexpected reserved word because of await. ...

javascript | import | async-await | export

asked 3 mins ago



omega
10.8k ● 50 ● 137 ● 294

Multilabel Classification

Applications: E-mail

Enron, e-mails messages made public from the Enron corporation.

"a few beers after work?" $\boxed{\text{work}}$ $\boxed{\text{personal}}$ $\boxed{\text{important}}$

For example, the **UC Berkeley Enron Email Analysis Project** multi-labeled 1702 *Enron* e-mails into 53 categories:

Company Business, Strategy, etc.
Purely Personal
Empty Message
Forwarded email(s)
...

Multilabel Classification

Single-label classification: Is this a picture of a beach?

$$\in \{\texttt{yes}, \texttt{no}\}$$

Multi-label classification: Which labels are relevant to this picture?

$$\subseteq \{\texttt{beach}, \texttt{sunset}, \texttt{foliage}, \texttt{field}, \texttt{mountain}, \texttt{urban}\}$$

Multilabel Classification

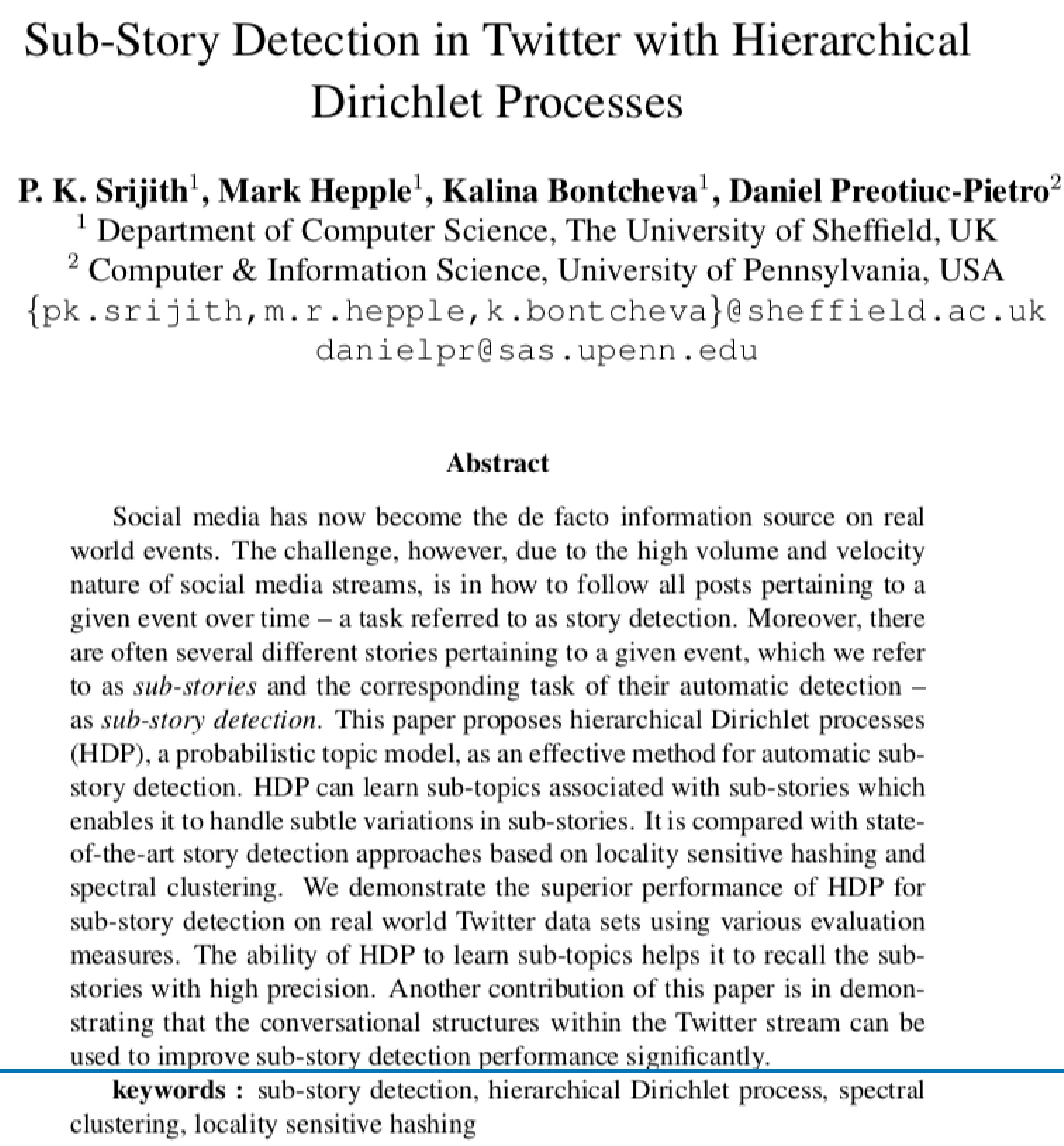
Labelling music/tracks with genres / voices, concepts, etc.

e.g., Emotions dataset, audio tracks labelled with different moods, among: {

- amazed-surprised,
- happy-pleased,
- relaxing-calm,
- quiet-still,
- sad-lonely,
- angry-aggressive

K. Trohidis, G. Tsoumakas, G. Kalliris, I. Vlahavas. "Multilabel Classification of Music into Emotions". Proc. 2008 International Conference on Music Information Retrieval (ISMIR 2008), pp. 325-330, Philadelphia, PA, USA, 2008

Multi-label Classification


Sub-Story Detection in Twitter with Hierarchical Dirichlet Processes

P. K. Srijith[1], **Mark Hepple**[1], **Kalina Bontcheva**[1], **Daniel Preotiuc-Pietro**[2]
[1] Department of Computer Science, The University of Sheffield, UK
[2] Computer & Information Science, University of Pennsylvania, USA
{pk.srijith,m.r.hepple,k.bontcheva}@sheffield.ac.uk
danielpr@sas.upenn.edu

Abstract

Social media has now become the de facto information source on real world events. The challenge, however, due to the high volume and velocity nature of social media streams, is in how to follow all posts pertaining to a given event over time – a task referred to as story detection. Moreover, there are often several different stories pertaining to a given event, which we refer to as *sub-stories* and the corresponding task of their automatic detection – as *sub-story detection*. This paper proposes hierarchical Dirichlet processes (HDP), a probabilistic topic model, as an effective method for automatic sub-story detection. HDP can learn sub-topics associated with sub-stories which enables it to handle subtle variations in sub-stories. It is compared with state-of-the-art story detection approaches based on locality sensitive hashing and spectral clustering. We demonstrate the superior performance of HDP for sub-story detection on real world Twitter data sets using various evaluation measures. The ability of HDP to learn sub-topics helps it to recall the sub-stories with high precision. Another contribution of this paper is in demonstrating that the conversational structures within the Twitter stream can be used to improve sub-story detection performance significantly.

keywords : sub-story detection, hierarchical Dirichlet process, spectral clustering, locality sensitive hashing

Multilabel classification

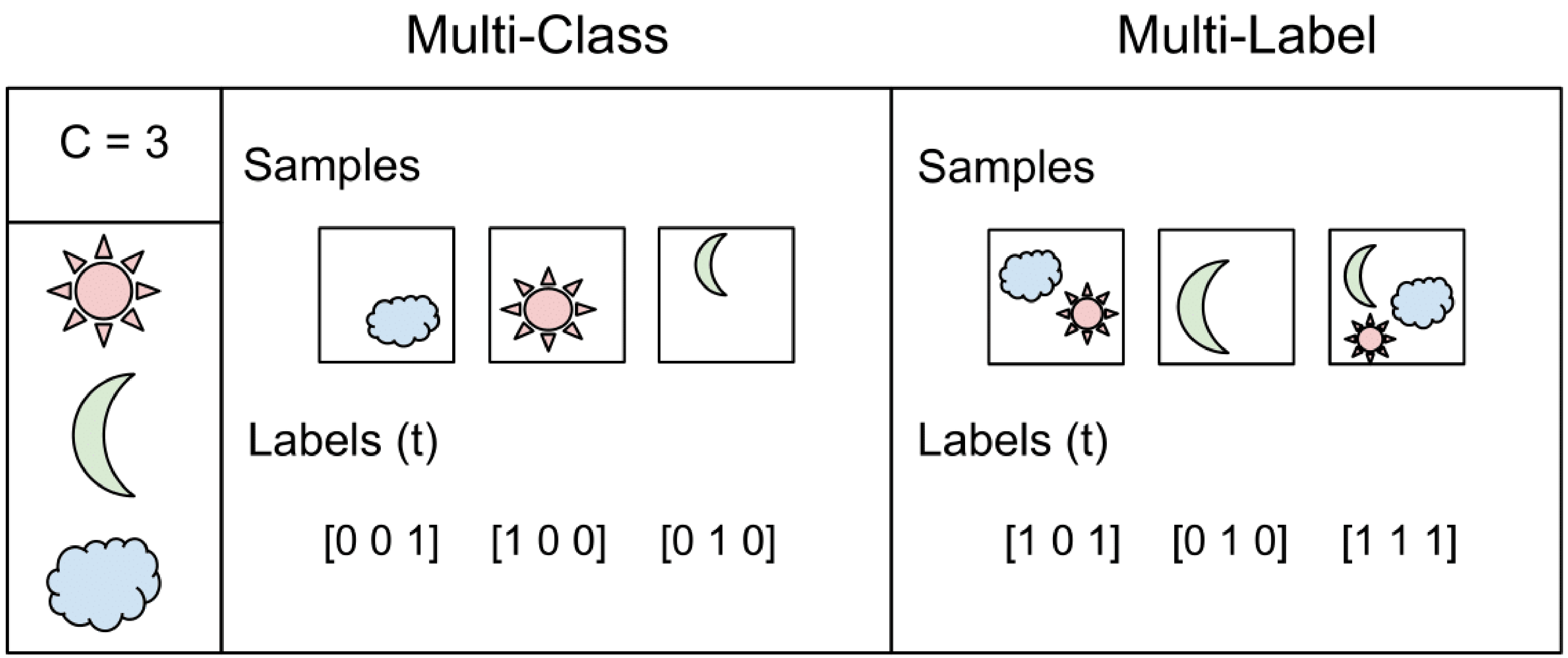


Table : Single-label $Y \in \{0, 1\}$

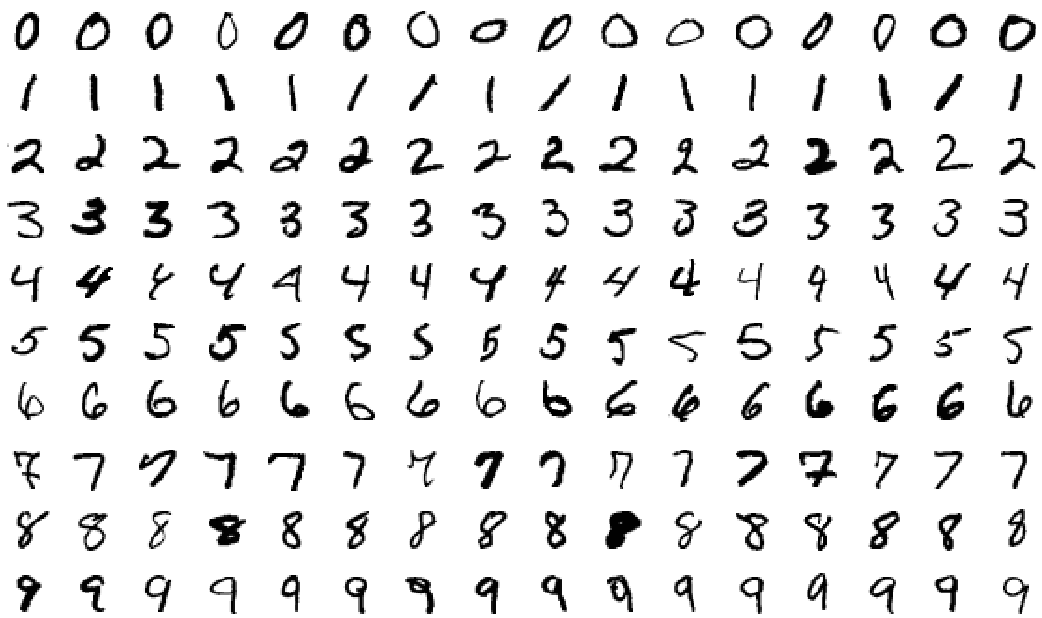
X_1	X_2	X_3	X_4	X_5	Y
1	0.1	3	1	0	0
0	0.9	1	0	1	1
0	0.0	1	1	0	0
1	0.8	2	0	1	1
1	0.0	2	0	1	0
0	0.0	3	1	1	?

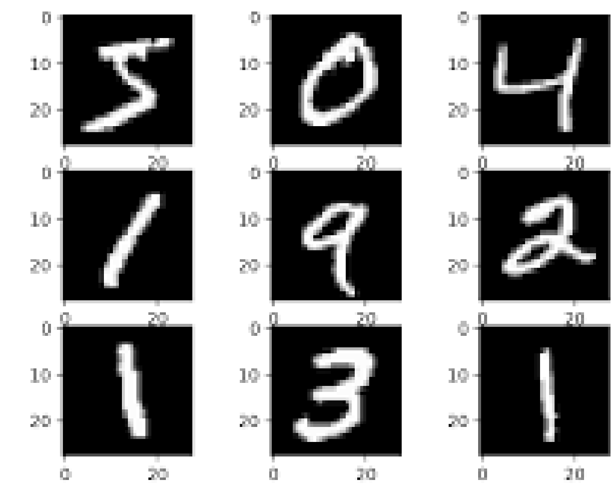
Table : Multi-label $Y_1, \ldots, Y_L \in 2^L$

X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3	Y_4
1	0.1	3	1	0	0	1	1	0
0	0.9	1	0	1	1	0	0	0
0	0.0	1	1	0	0	1	0	0
1	0.8	2	0	1	1	0	0	1
1	0.0	2	0	1	0	0	0	1
0	0.0	3	1	1	?	?	?	?

Slide Credits : Jesse Read, Multi-Label Classification, ML KDD 2013

Multi-class Classification



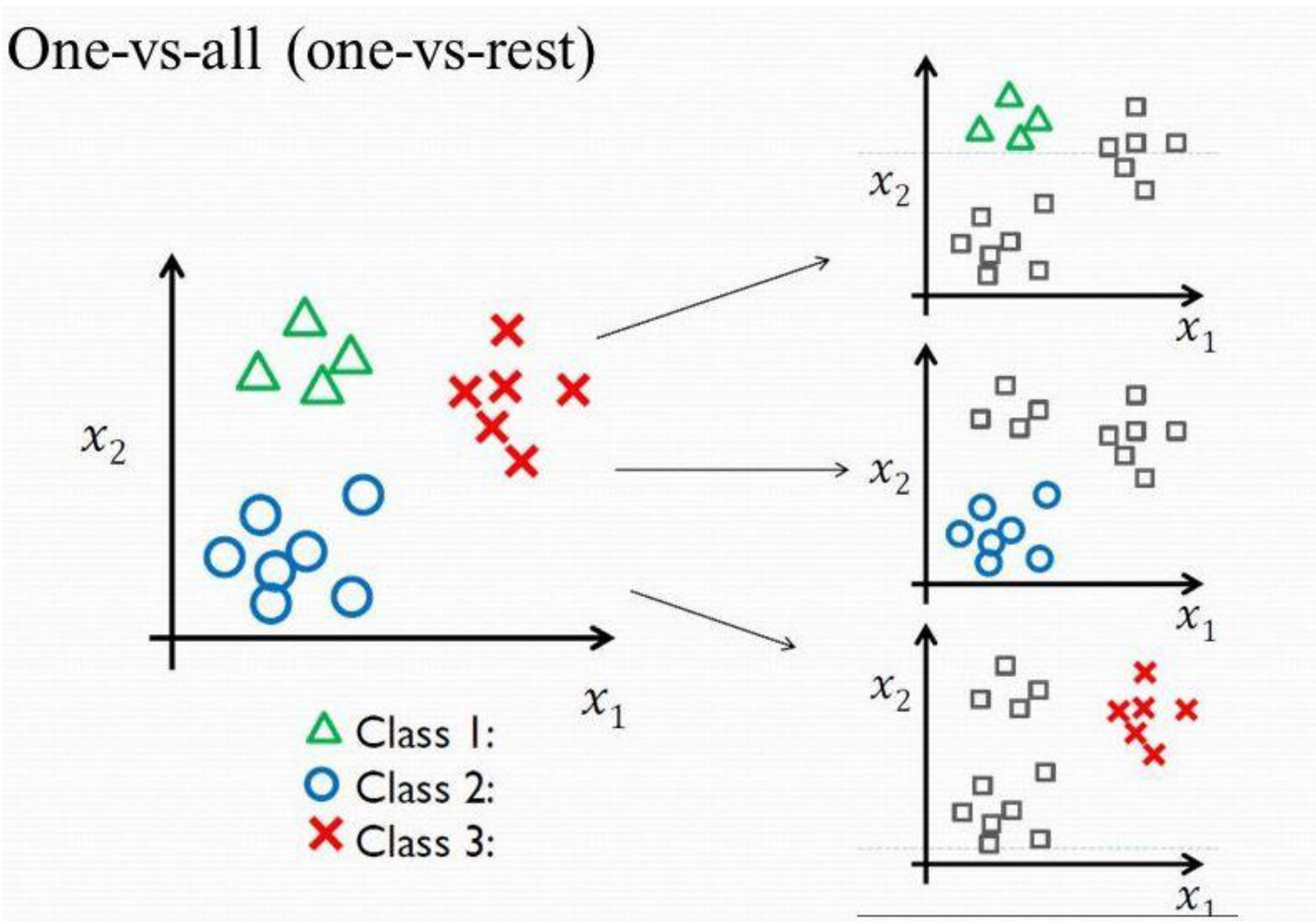


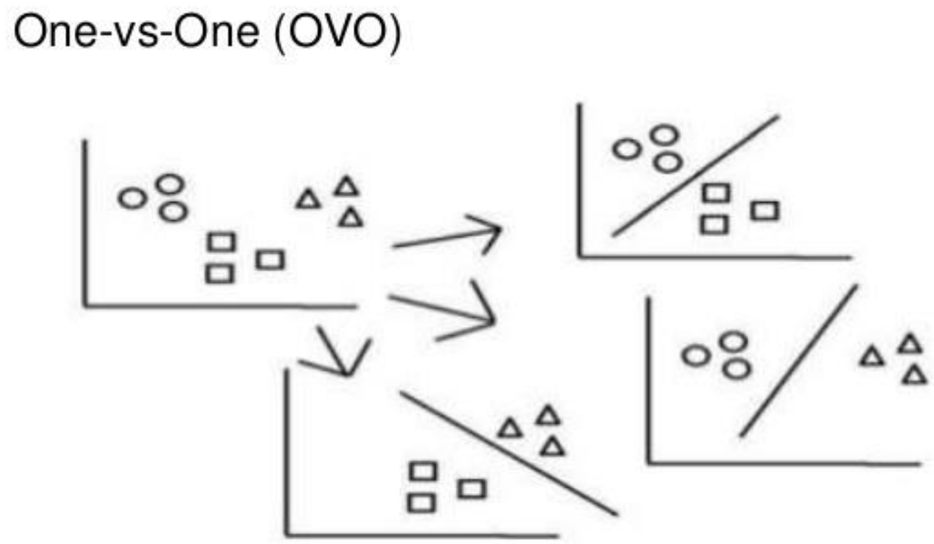
- In multi class classification, a sample belongs to only one class

X1	X2	X3	X4	X5	Y1	Y2	Y3	Y4
1	0.1	3	1	0	0	1	0	0
0	0.9	1	0	1	1	0	0	0
0	0.0	1	1	0	0	1	0	0
1	0.8	2	0	1	1	0	0	0
1	0.0	2	0	1	0	0	0	1
0	0.0	3	1	1	?	?	?	?

Multi-class Classification

- In multi class classification, a sample belongs to only one class
- Logistic Regression with Softmax, K-NN… or using Binary classifiers





Multilabel classification

There are dependencies (i.e., *correlations*, *relationships*, *co-occurences*) among labels

- e.g., $\{\texttt{relaxing-calm}, \texttt{quiet-still}\}$ vs. $\{\texttt{relaxing-calm}, \texttt{angry-aggressive}\}$
- e.g., $\{\texttt{beach}, \texttt{sunset}\}$ vs. $\{\texttt{beach}, \texttt{field}\}$

From the IMDb dataset:

- $P(\texttt{family})P(\texttt{adult}) = 0.068 \cdot 0.015 = 0.001$ ($\approx$ 121 movies)
- $P(\texttt{family}, \texttt{adult}) = 0.0$ (0 movies!)

On most datasets:

- $P(\mathbf{y} = [1, 1, 1, 1, 1, 1]) = 0$

Multilabel classification

There are dependencies (i.e., *correlations*, *relationships*, *co-occurences*) among labels

- e.g., $\{\texttt{relaxing-calm}, \texttt{quiet-still}\}$ vs. $\{\texttt{relaxing-calm}, \texttt{angry-aggressive}\}$
- e.g., $\{\texttt{beach}, \texttt{sunset}\}$ vs. $\{\texttt{beach}, \texttt{field}\}$

From the IMDb dataset:

- $P(\texttt{family})P(\texttt{adult}) = 0.068 \cdot 0.015 = 0.001$ ($\approx$ 121 movies)
- $P(\texttt{family}, \texttt{adult}) = 0.0$ (0 movies!)

On most datasets:

- $P(\mathbf{y} = [1, 1, 1, 1, 1, 1]) = 0$

The main challenges are to

- model label dependencies; and
- do this efficiently.

Methods for Multi-Label Classification

Problem Transformation Methods

- Transforms the multi-label problem into single-label problem(s)
- Use any off-the-shelf single-label classifier to suit requirements
- i.e., **Adapt your data to the algorithm**

Algorithm Adaptation Methods

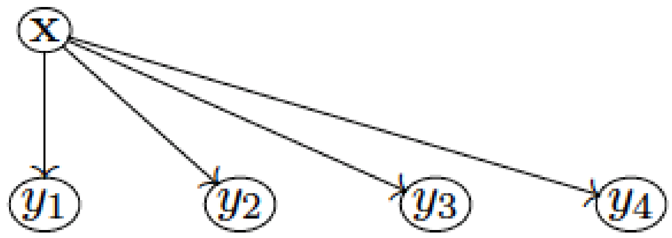
- Adapt a single-label algorithm to produce multi-label outputs
- Benefit from specific classifier advantages (e.g., efficiency)
- i.e., **Adapt your algorithm to the data**

Many methods involve a mix of both approaches.

Binary Relevance

$\mathbf{X}$	Y_1	Y_2	Y_3	Y_4
$\mathbf{x}^{(1)}$	0	1	1	0
$\mathbf{x}^{(2)}$	1	0	0	0
$\mathbf{x}^{(3)}$	0	1	0	0
$\mathbf{x}^{(4)}$	1	0	0	1
$\mathbf{x}^{(5)}$	0	0	0	1

Prediction: $\hat{\mathbf{y}} = [h_1(\tilde{\mathbf{x}}), \ldots, h_L(\tilde{\mathbf{x}})]$



…just make L separate binary problems (one for each label):

$\mathbf{X}$	Y_1
$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	1
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	0

$\mathbf{X}$	Y_2
$\mathbf{x}^{(1)}$	1
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	1
$\mathbf{x}^{(4)}$	0
$\mathbf{x}^{(5)}$	0

$\mathbf{X}$	Y_3
$\mathbf{x}^{(1)}$	1
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	0
$\mathbf{x}^{(5)}$	0

$\mathbf{X}$	Y_4
$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	1

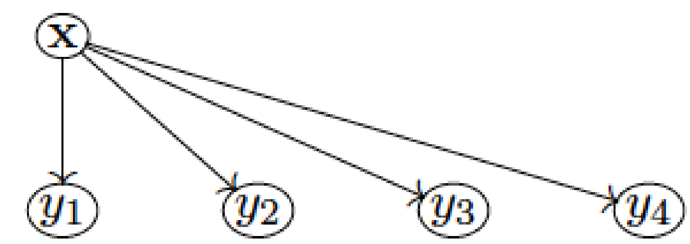
and train with any off-the-shelf binary classifier.

Slide Credits : Jesse Read, Multi-Label Classification, ML KDD 2013

Binary Relevance

$\mathbf{X}$	Y_1	Y_2	Y_3	Y_4
$\mathbf{x}^{(1)}$	0	1	1	0
$\mathbf{x}^{(2)}$	1	0	0	0
$\mathbf{x}^{(3)}$	0	1	0	0
$\mathbf{x}^{(4)}$	1	0	0	1
$\mathbf{x}^{(5)}$	0	0	0	1

Prediction: $\hat{\mathbf{y}} = [h_1(\tilde{\mathbf{x}}), \ldots, h_L(\tilde{\mathbf{x}})]$



...just make L separate binary problems (one for each label):

$\mathbf{X}$	Y_1
$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	1
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	0

$\mathbf{X}$	Y_2
$\mathbf{x}^{(1)}$	1
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	1
$\mathbf{x}^{(4)}$	0
$\mathbf{x}^{(5)}$	0

$\mathbf{X}$	Y_3
$\mathbf{x}^{(1)}$	1
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	0
$\mathbf{x}^{(5)}$	0

$\mathbf{X}$	Y_4
$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	0
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	1

and train with any off-the-shelf binary classifier.

Disadvantages

Does not Model label dependency

Class Imbalance

Slide Credits : Jesse Read, Multi-Label Classification, ML KDD 2013

Stacked Binary Relevance

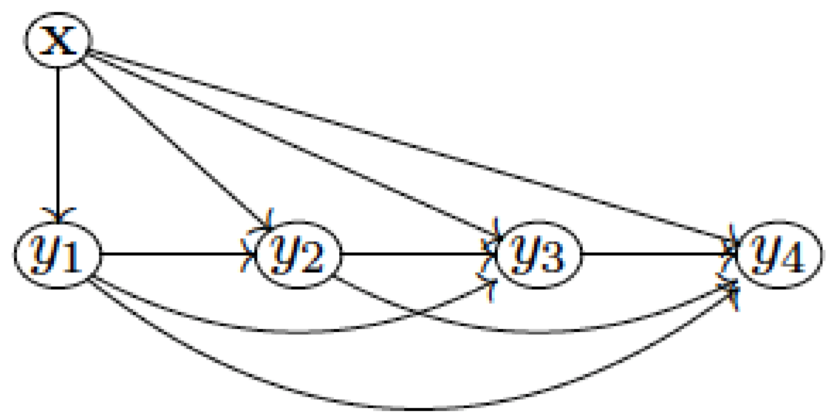
$$\hat{\mathbf{y}} = \mathbf{h}^2(\mathbf{h}^1(\tilde{\mathbf{x}}))$$

For example, given $\tilde{\mathbf{x}}$,

	$\hat{Y}_1$	$\hat{Y}_2$	$\hat{Y}_3$	$\hat{Y}_4$
$\mathbf{h}^1(\tilde{\mathbf{x}})$	1	0	0	1
$\hat{\mathbf{y}} = \mathbf{h}^2(\mathbf{h}^1(\tilde{\mathbf{x}}))$	1	0	0	0

BR stacked with feature outputs. For more information see: Shantanu Godbole, Sunita Sarawagi: Discriminative Methods for Multi-labeled Classification. In: Advances in Knowledge Discovery and Data Mining, 22-30, 2004.

Chain Classifier



Like BR, make L binary problems, but include previous predictions as feature attributes.

$\mathbf{X}$	Y_1
$\mathbf{x}^{(1)}$	0
$\mathbf{x}^{(2)}$	1
$\mathbf{x}^{(3)}$	0
$\mathbf{x}^{(4)}$	1
$\mathbf{x}^{(5)}$	0

$\mathbf{X}$	Y_1	Y_2
$\mathbf{x}^{(1)}$	0	1
$\mathbf{x}^{(2)}$	1	0
$\mathbf{x}^{(3)}$	0	1
$\mathbf{x}^{(4)}$	1	0
$\mathbf{x}^{(5)}$	0	0

$\mathbf{X}$	Y_1	Y_2	Y_3
$\mathbf{x}^{(1)}$	0	1	1
$\mathbf{x}^{(2)}$	1	0	0
$\mathbf{x}^{(3)}$	0	1	0
$\mathbf{x}^{(4)}$	1	0	0
$\mathbf{x}^{(5)}$	0	0	0

$\mathbf{X}$	Y_1	Y_3	Y_3	Y_4
$\mathbf{x}^{(1)}$	0	1	1	0
$\mathbf{x}^{(2)}$	1	0	0	0
$\mathbf{x}^{(3)}$	0	1	0	0
$\mathbf{x}^{(4)}$	1	0	0	1
$\mathbf{x}^{(5)}$	0	0	0	1

Krzysztof Dembczynski, Weiwei Cheng, and Eyke Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *ICML*, volume 10, pages 279–286, 2010.

Label Powerset Method

To model label correlations, we can . . .

$\mathbf{X}$	Y_1	Y_2	Y_3	Y_4
$\mathbf{x}^{(1)}$	0	1	1	0
$\mathbf{x}^{(2)}$	1	0	0	0
$\mathbf{x}^{(3)}$	0	1	1	0
$\mathbf{x}^{(4)}$	1	0	0	1
$\mathbf{x}^{(5)}$	0	0	0	1

. . . make a single multi-*class* problem with 2^L possible class values:

$\mathbf{X}$	$Y \in 2^L$
$\mathbf{x}^{(1)}$	0110
$\mathbf{x}^{(2)}$	1000
$\mathbf{x}^{(3)}$	0110
$\mathbf{x}^{(4)}$	1001
$\mathbf{x}^{(5)}$	0001

and train with any off-the-shelf multi-*class* classifier.

Slide Credits : Jesse Read, Multi-Label Classification, ML KDD 2013

Label Powerset Method

To model label correlations, we can . . .

$\mathbf{X}$	Y_1	Y_2	Y_3	Y_4
$\mathbf{x}^{(1)}$	0	1	1	0
$\mathbf{x}^{(2)}$	1	0	0	0
$\mathbf{x}^{(3)}$	0	1	1	0
$\mathbf{x}^{(4)}$	1	0	0	1
$\mathbf{x}^{(5)}$	0	0	0	1

. . . make a single multi-*class* problem with 2^L possible class values:

$\mathbf{X}$	$Y \in 2^L$
$\mathbf{x}^{(1)}$	0110
$\mathbf{x}^{(2)}$	1000
$\mathbf{x}^{(3)}$	0110
$\mathbf{x}^{(4)}$	1001
$\mathbf{x}^{(5)}$	0001

- complexity: many class labels
- imbalance: not many examples per class label

and train with any off-the-shelf multi-*class* classifier.

Slide Credits : Jesse Read, Multi-Label Classification, ML KDD 2013

Ensembles of Random k-label Subsets (RAKEL)

- Do LP on M subsets $\subset \{\lambda_1, \ldots, \lambda_L\}$ of size k

$\mathbf{X}$	$Y \in 2^k$
$\mathbf{x}^{(1)}$	011
$\mathbf{x}^{(2)}$	100
$\mathbf{x}^{(3)}$	011
$\mathbf{x}^{(4)}$	100
$\mathbf{x}^{(5)}$	000

$\mathbf{X}$	$Y \in 2^k$
$\mathbf{x}^{(1)}$	010
$\mathbf{x}^{(2)}$	100
$\mathbf{x}^{(3)}$	010
$\mathbf{x}^{(4)}$	101
$\mathbf{x}^{(5)}$	001

$\mathbf{X}$	$Y \in 2^k$
$\mathbf{x}^{(1)}$	010
$\mathbf{x}^{(2)}$	100
$\mathbf{x}^{(3)}$	010
$\mathbf{x}^{(4)}$	101
$\mathbf{x}^{(5)}$	001

$\mathbf{X}$	$Y \in 2^k$
$\mathbf{x}^{(1)}$	110
$\mathbf{x}^{(2)}$	000
$\mathbf{x}^{(3)}$	110
$\mathbf{x}^{(4)}$	001
$\mathbf{x}^{(5)}$	001

- 2^k problems much easier to deal with than 2^L (but still models label dependencies)
- use k and M (number of models) to trade-off dependency modelling and scalability

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7): 1079–1089, 2011a.

Ensemble Methods : Prediction

Make Prediction by voting

	$\hat{y}_1$	$\hat{y}_2$	$\hat{y}_3$	$\hat{y}_4$
$\mathbf{h}^1(\tilde{\mathbf{x}})$	1	0	1	
$\mathbf{h}^2(\tilde{\mathbf{x}})$		1	1	0
$\mathbf{h}^3(\tilde{\mathbf{x}})$	1		1	0
$\mathbf{h}^4(\tilde{\mathbf{x}})$	1	0		0
$\mathbf{h}(\tilde{\mathbf{x}})$	3	1	3	0
$\hat{\mathbf{y}}$	1	0	1	0

Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multilabel classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7): 1079–1089, 2011a.

Multi-label Data: Datasets

	$\mathcal{X}$ (data inst.)	$\mathcal{Y}$ (labels)	L	N	D	LC
Music	audio data	emotions	6	593	72	1.87
Scene	image data	scene labels	6	2407	294	1.07
Yeast	genes	biological fns	14	2417	103	4.24
Genbase	genes	biological fns	27	661	1185	1.25
Medical	medical text	diagnoses	45	978	1449	1.25
Enron	e-mails	labels, tags	53	1702	1001	3.38
Reuters	news articles	categories	103	6000	500	1.46
TMC07	textual reports	errors	22	28596	500	2.16
Ohsumed	medical articles	disease cats.	23	13929	1002	1.66
IMDB	plot summaries	genres	28	120919	1001	2.00
20NG	posts	news groups	20	19300	1006	1.03
MediaMill	video data	annotations	101	43907	120	4.38
Del.icio.us	bookmarks	tags	983	16105	500	19.02

- L number of labels
- N number of examples
- D number of input feature attributes
- Label Cardinality (LC) $\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L} y_j^{(i)}$ (Average number of labels per example)

Slide Credits : Jesse Read, Multi-Label Classification, ML KDD 2013

Multi-Label Evaluation

0/1 LOSS

$$= \frac{1}{N}\sum_{i=1}^{N}\mathcal{I}(\hat{\mathbf{y}}^{(i)} \neq \mathbf{y}^{(i)})$$

$$= 0.60$$

	$\mathbf{y}^{(i)}$	$\hat{\mathbf{y}}^{(i)}$
$\tilde{\mathbf{x}}^{(1)}$	[1 0 1 0]	[1 0 0 1]
$\tilde{\mathbf{x}}^{(2)}$	[0 1 0 1]	[0 1 0 1]
$\tilde{\mathbf{x}}^{(3)}$	[1 0 0 1]	[1 0 0 1]
$\tilde{\mathbf{x}}^{(4)}$	[0 1 1 0]	[0 1 0 0]
$\tilde{\mathbf{x}}^{(5)}$	[1 0 0 0]	[1 0 0 1]

HAMMING LOSS

$$= \frac{1}{NL}\sum_{i=1}^{N}\sum_{i=1}^{L}\mathcal{I}[\hat{y}_j^{(i)} \neq y_j^{(i)}]$$

$$= 0.20$$

Multi-Label Evaluation

	$\mathbf{y}^{(i)}$	$\hat{\mathbf{y}}^{(i)}$
$\tilde{\mathbf{x}}^{(1)}$	[1 0 1 0]	[1 0 0 1]
$\tilde{\mathbf{x}}^{(2)}$	[1 0 0 1]	[1 0 0 1]
$\tilde{\mathbf{x}}^{(3)}$	[0 1 1 0]	[0 1 0 0]
$\tilde{\mathbf{x}}^{(4)}$	[1 0 0 0]	[1 0 1 1]
$\tilde{\mathbf{x}}^{(5)}$	[0 1 0 1]	[0 1 0 1]

Multi-Label Evaluation

	$\mathbf{y}^{(i)}$	$\hat{\mathbf{y}}^{(i)}$
$\tilde{\mathbf{x}}^{(1)}$	[1 0 1 0]	[1 0 0 1]
$\tilde{\mathbf{x}}^{(2)}$	[1 0 0 1]	[1 0 0 1]
$\tilde{\mathbf{x}}^{(3)}$	[0 1 1 0]	[0 1 0 0]
$\tilde{\mathbf{x}}^{(4)}$	[1 0 0 0]	[1 0 1 1]
$\tilde{\mathbf{x}}^{(5)}$	[0 1 0 1]	[0 1 0 1]

- Ham. Loss 0.3
- 0/1 Loss 0.6

Multi-Label classification softwares

- MULAN and MEKA based on WEKA provides multi-label classication
- scikit-multilearn: A scikit-based Python environment for performing multi-label classication

Multi-Task Learing

Multiple Tasks

- Examination Scores Prediction[1] (Argyriou et. al.'08)

School 1 - Alverno High School

Student id	Birth year	Previous score	...	School ranking	...
72981	1985	95	...	83%	...

student-dependent school-dependent

→

Exam score
?

...

School 138 - Jefferson Intermediate School

Student id	Birth year	Previous score	...	School ranking	...
31256	1986	87	...	72%	...

student-dependent school-dependent

→

Exam score
?

School 139 - Rosemead High School

Student id	Birth year	Previous score	...	School ranking	...
12381	1986	83	...	77%	...

student-dependent school-dependent

→

Exam score
?

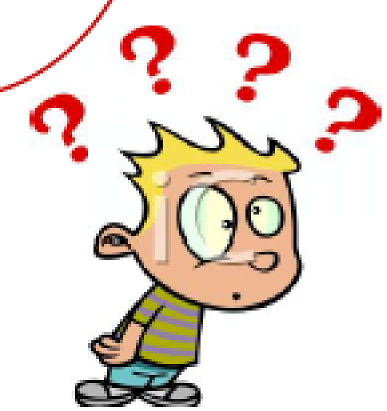
Slide Credits : Jiayu Zhou, Jianhui Chen, Jieping Ye, Multitask learning, SDM Tutorial 2012

Learn Multiple Tasks

- Learning from the pool of all tasks

Student id	Birth year	Previous score	School ranking	…	Exam Score
72981	1985	95	83%	…	?
31256	1986	87	72%	…	?
12381	1987	83	77%	…	?
…	…	…	…	…	…
21901	1986	87	72%	…	?

Students with same Features but different Exam Scores

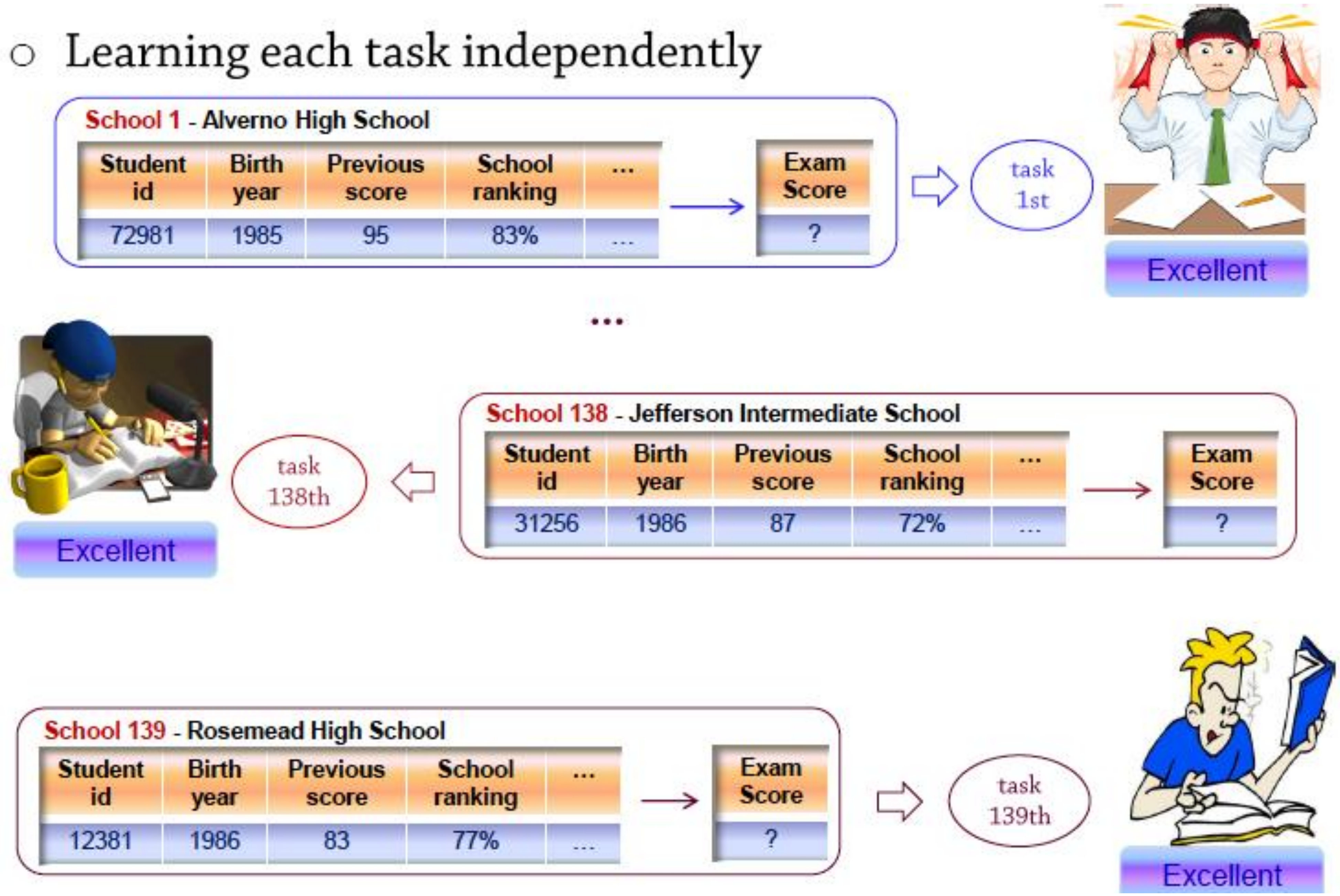


School A

School B

Slide Credits : Jiayu Zhou, Jianhui Chen, Jieping Ye, Multitask learning, SDM Tutorial 2012

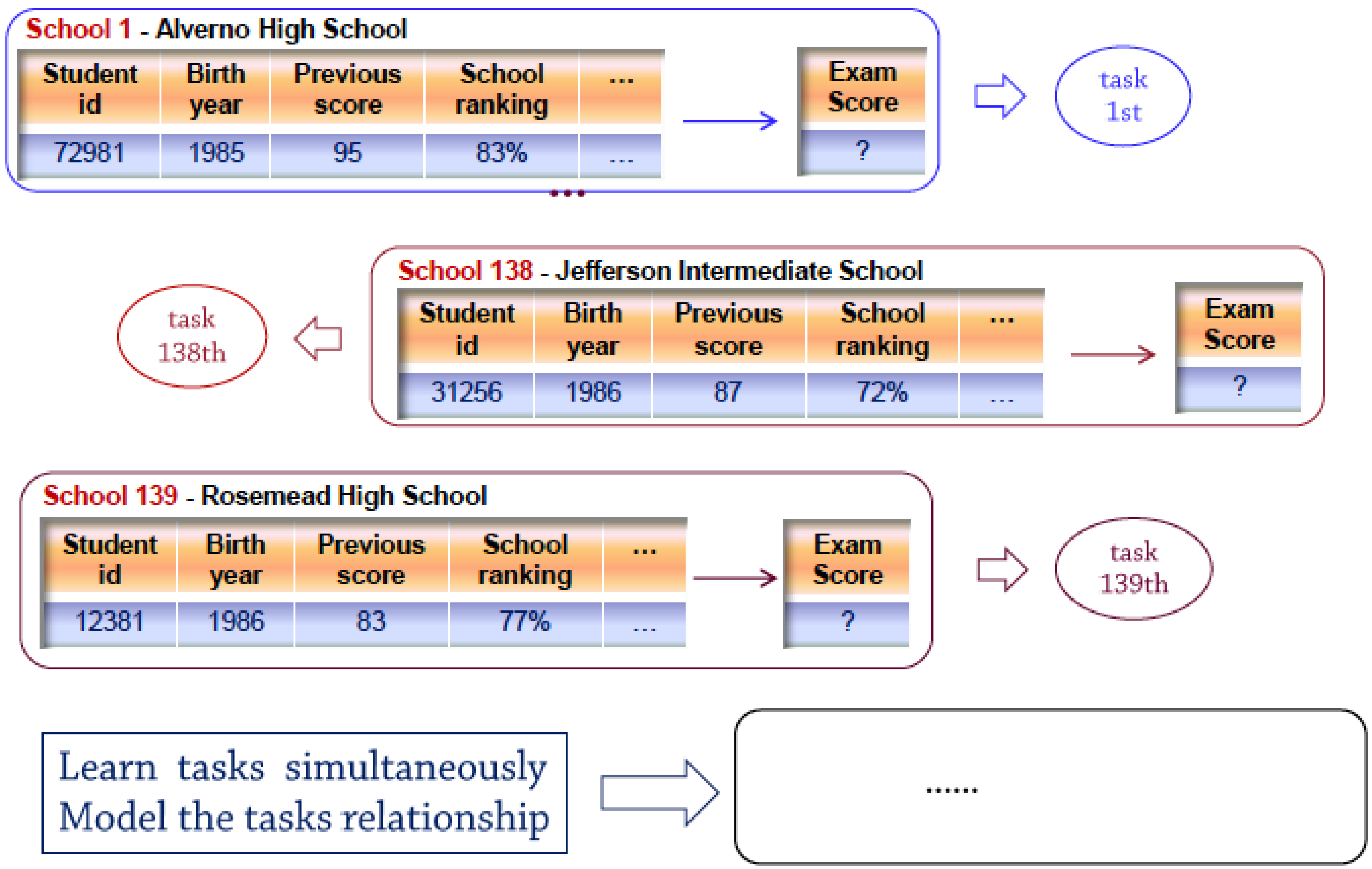
Learning Multiple Tasks



Slide Credits : Jiayu Zhou, Jianhui Chen, Jieping Ye, Multitask learning, SDM Tutorial 2012

Learning Multiple Tasks

- Leaning multiple tasks simultaneously



Slide Credits : Jiayu Zhou, Jianhui Chen, Jieping Ye, Multitask learning, SDM Tutorial 2012

MultiTask Learning

- The preference prediction of each user can be modeled using ordinal regression
- Some users have similar tastes and their predictions may also have similarities

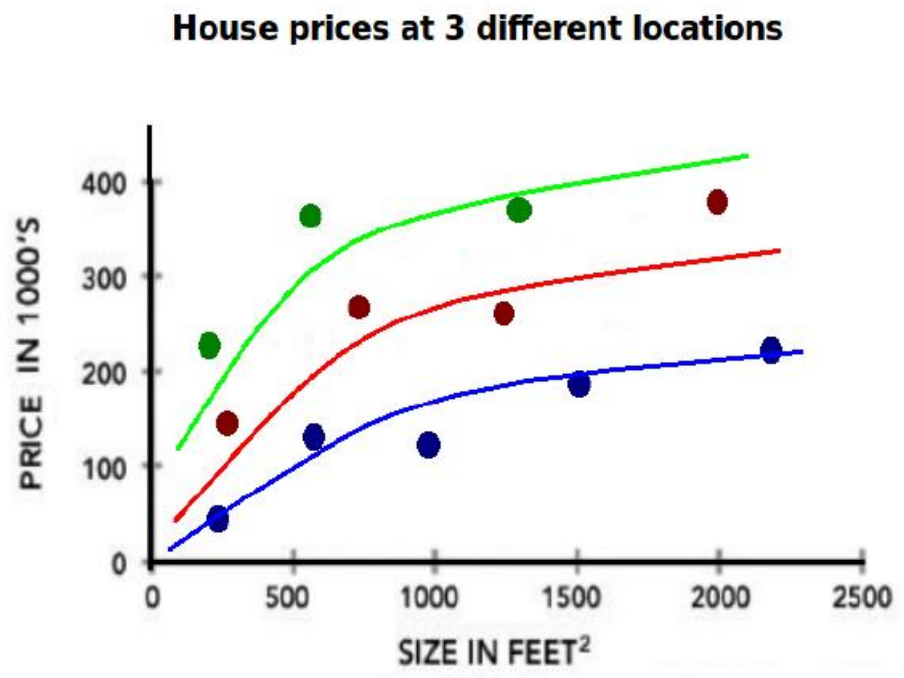
MultiTask Learning

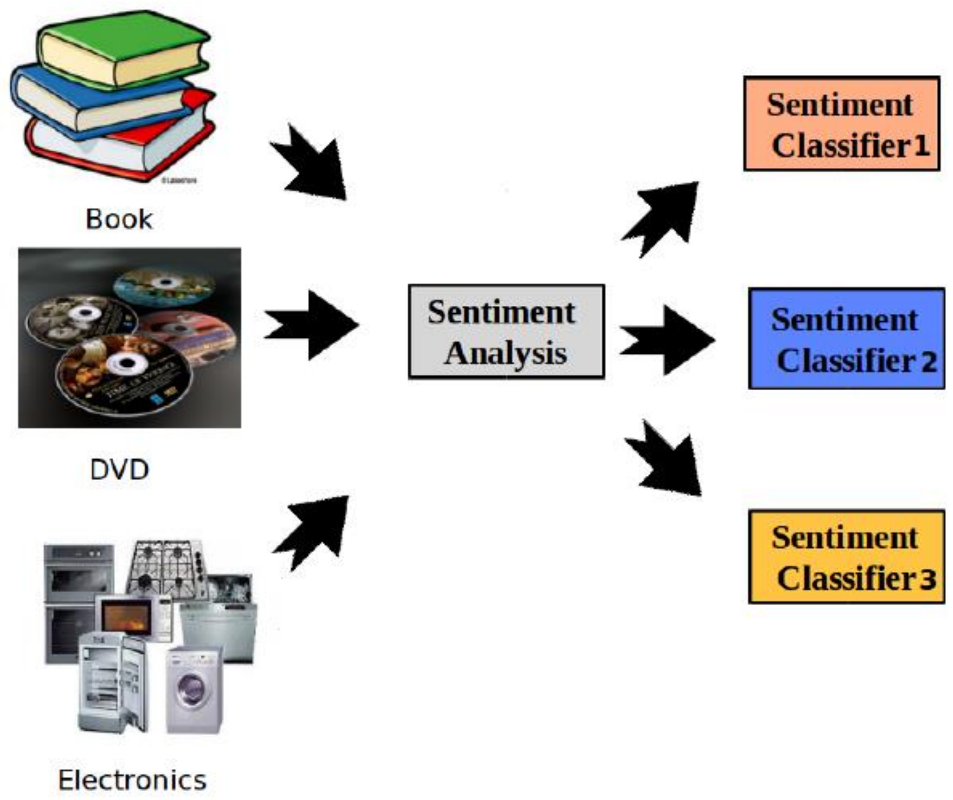
Multi-task Learning (MTL)

- Several related learning problems.
- Each learning problem is associated with limited data
- Models each problem as a task and learns all the tasks simultaneously.

Problem T tasks, $\mathbf{D}^t = (\mathbf{X}^t, \mathbf{y}^t) = \{\mathbf{x}_i^t, y_i^t\}_{i=1}^{N^t}$ $\forall t = 1, \ldots, T$ $y_i^t \in \mathcal{Y}$

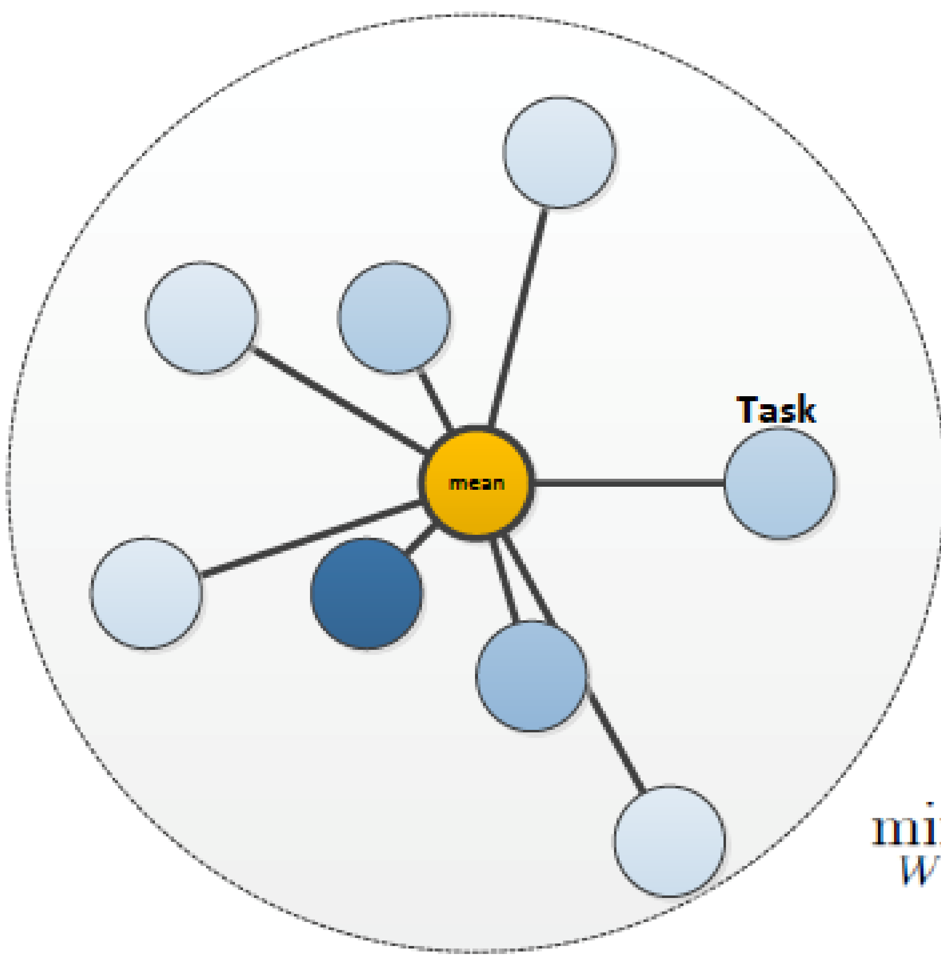
Learn $f^t : \mathcal{X} \rightarrow \mathcal{Y}$ for each task t.





MultiTask Learning - Regularization Based

- Assume all tasks are related in that the models of all tasks come from a particular distribution (Evgeniou & Pontil, KDD 04)

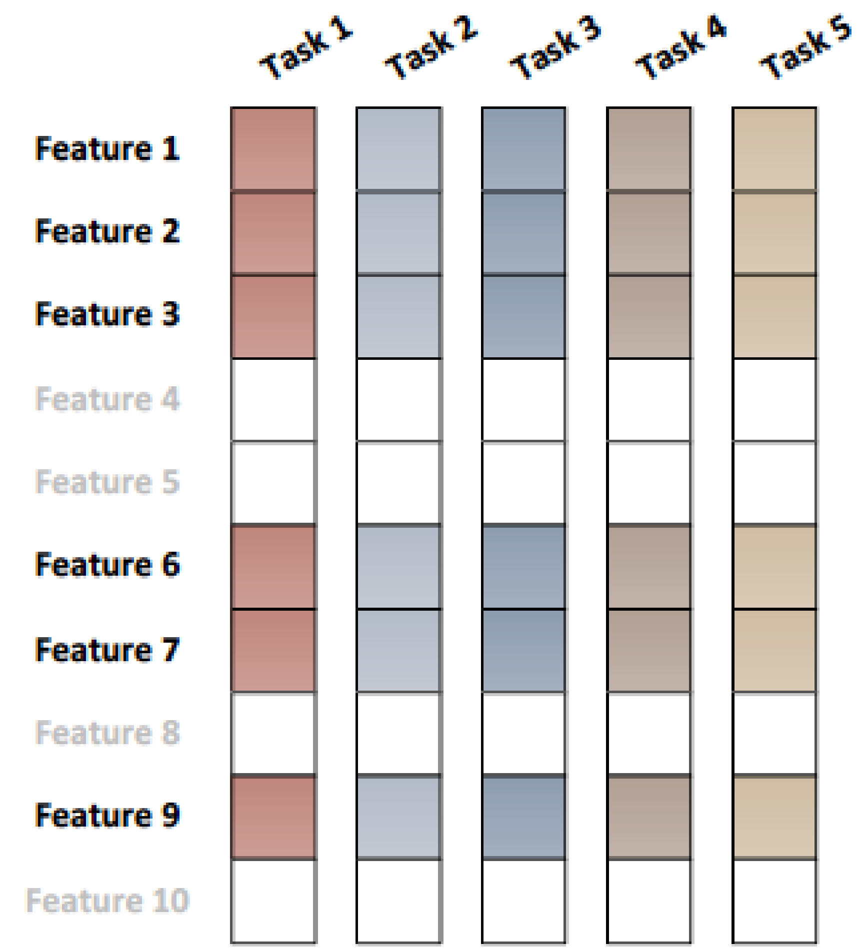


Regularization penalizes the deviation of each task from the mean

$$\min_{W} \mathrm{Loss}(W) + \lambda \sum_{t=1}^{T} \|W_t - \frac{1}{T}\sum_{s=1}^{T} W_s\|$$

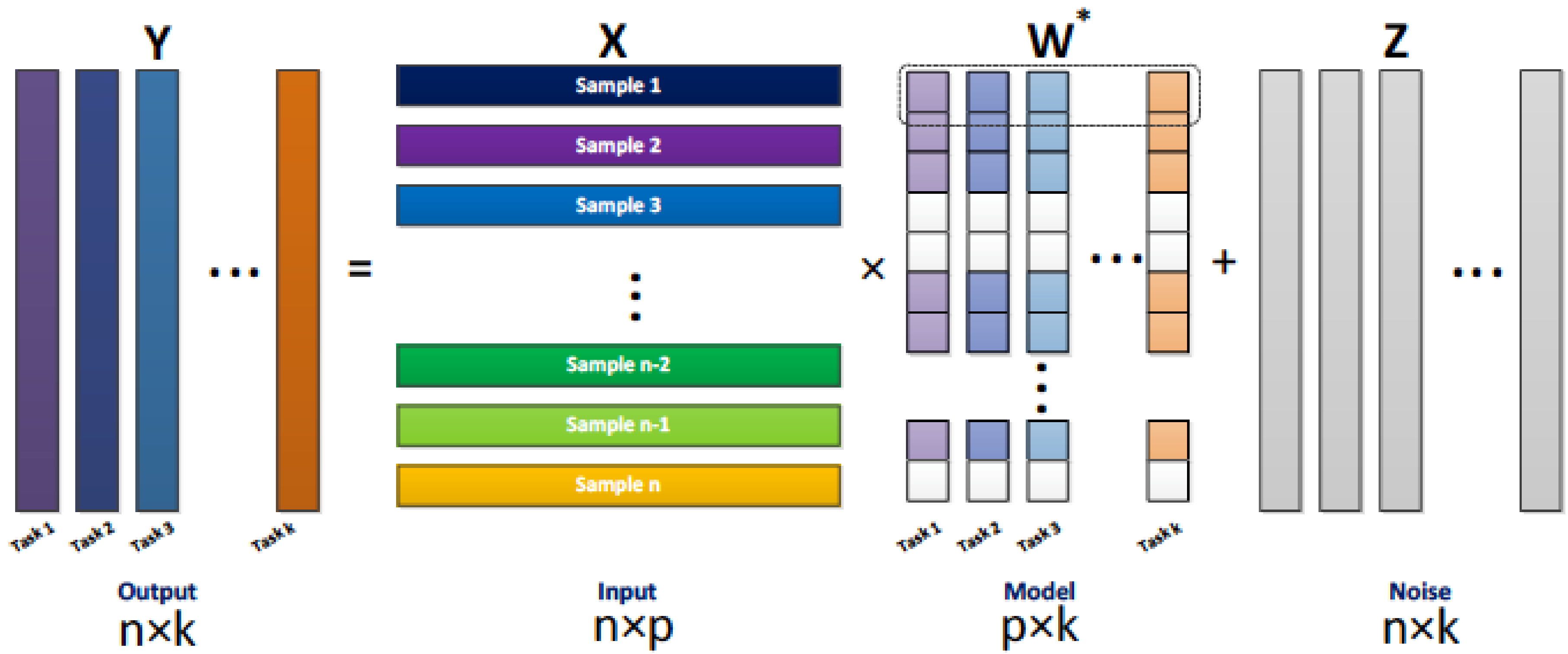
MultiTask Learning - Joint Feature Selection

- One way to capture the task relatedness from multiple related tasks is to constrain all models to share a common set of features.
- For example, in school data, the scores from different schools may be determined by a similar set of features.



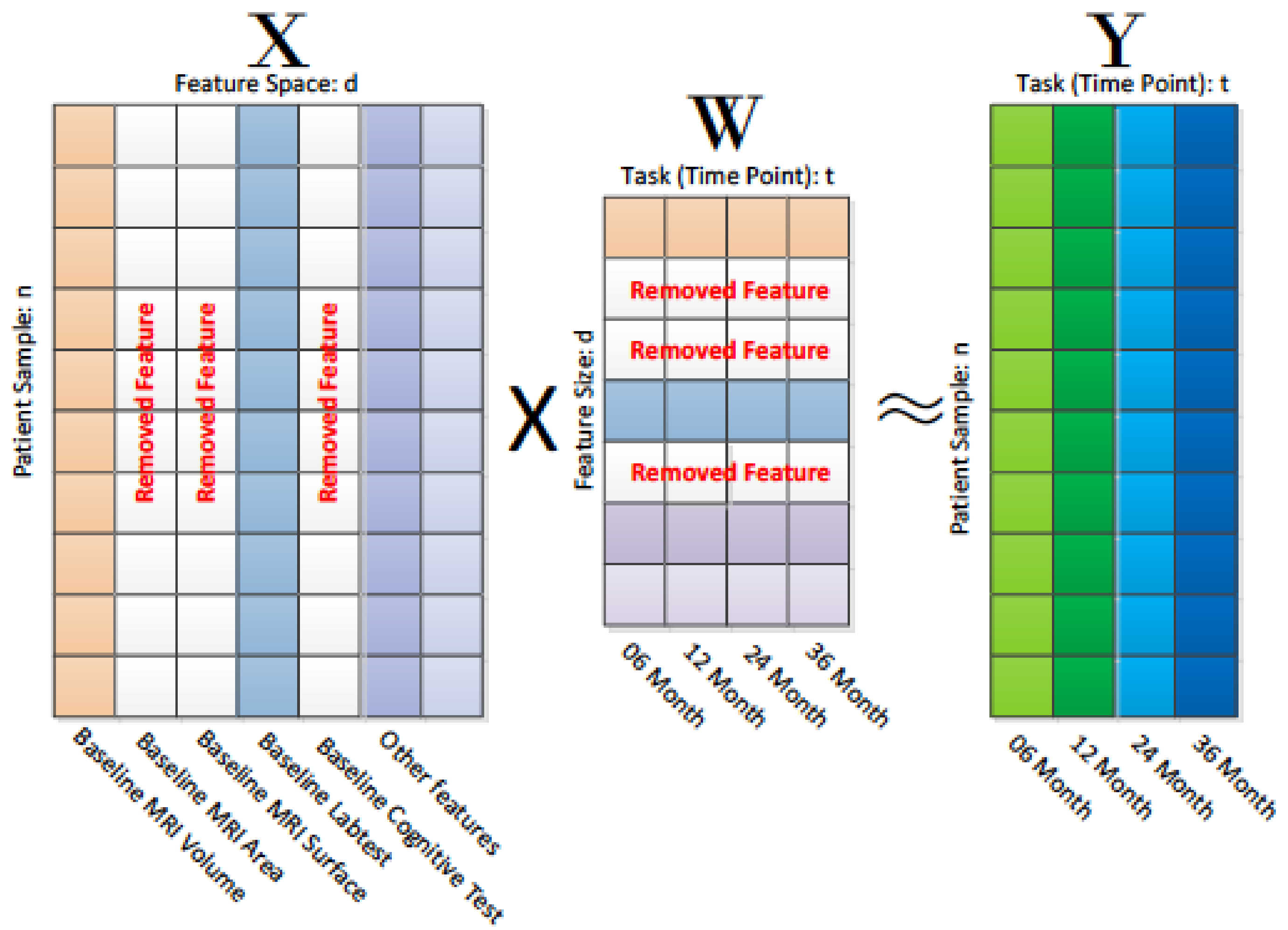
MultiTask Learning - Joint Feature Selection

Using group sparsity: ℓ_1/ℓ_2-norm regularization



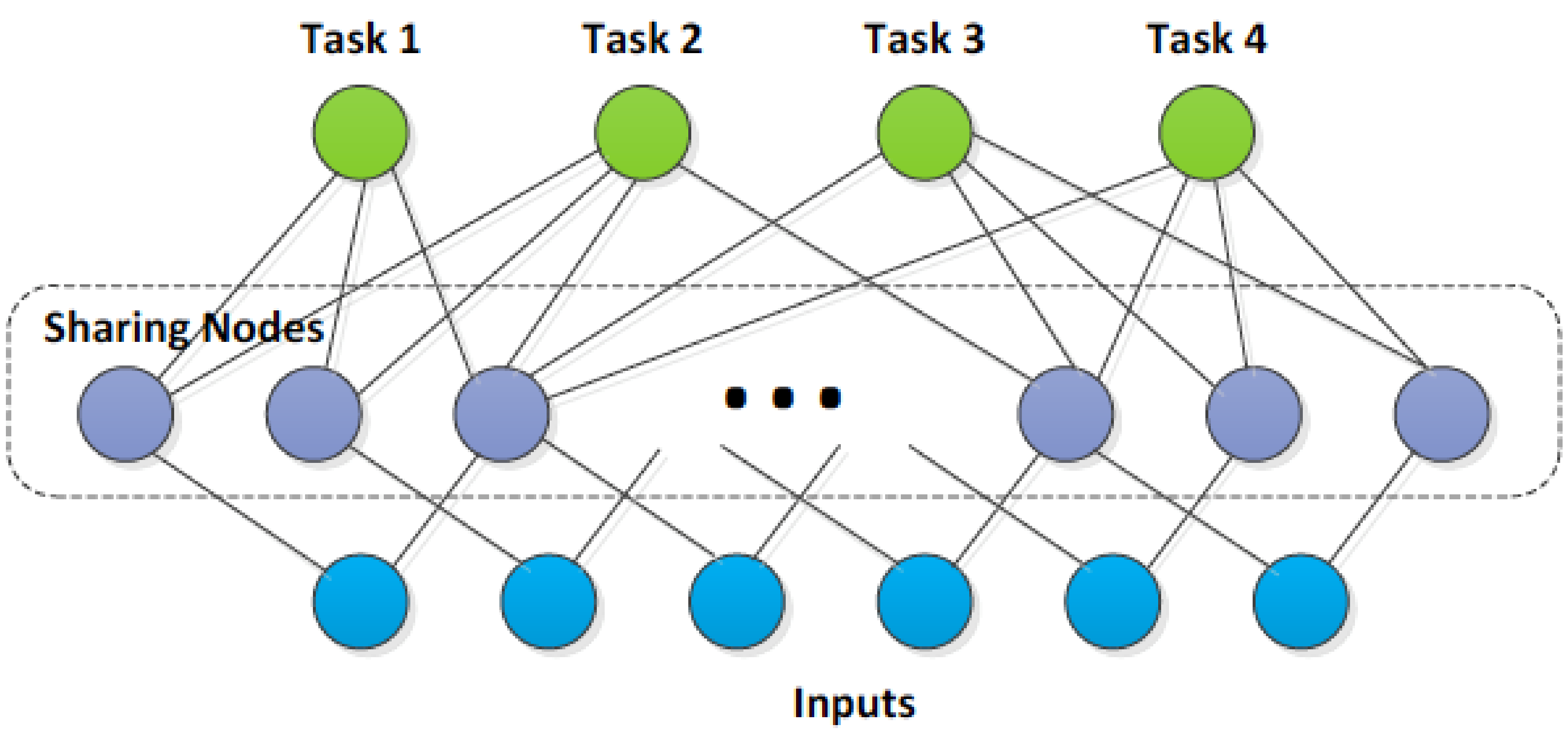
$$\min_{W} \frac{1}{2}\|XW - Y\|_F^2 + \lambda \sum_{i=1}^{p} \|\mathbf{w}_i\|_2$$

MultiTask Learning - Joint Feature Selection

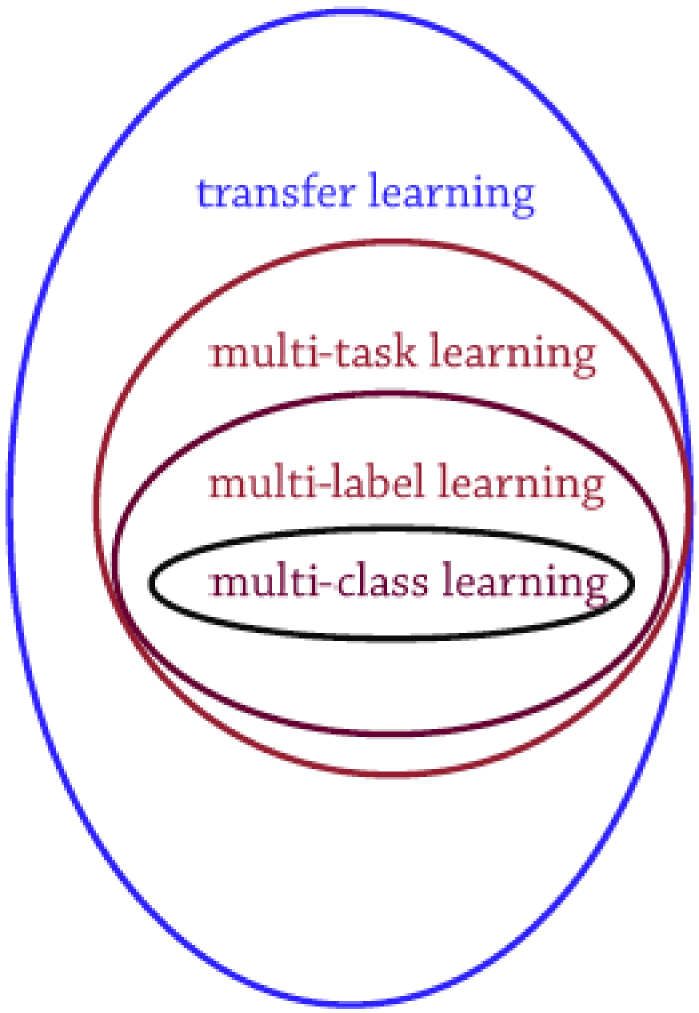


MultiTask Learning in Neural Networks

- Sharing the Hidden Nodes
  - Neural network has been well studied for learning multiple related tasks for improved generalization performance.
  - A set of hidden units are shared among multiple tasks for improved generalization (Caruana ML 97).



Learning Methods



- Transfer Learning
  - Define source & target domains
  - Learn on the source domain
  - Generalize on the target domain
- Multi-task Learning
  - Model the task relatedness
  - Learn all tasks simultaneously
  - Tasks may have different data/features
- Multi-label Learning
  - Model the label relatedness
  - Learn all labels simultaneously
  - Labels share the same data/features
- Multi-class Learning
  - Learn the classes independently
  - All classes are exclusive

MALSAR : Multi-task learning Software