Multi-Output Learning with Chaining

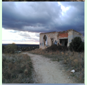
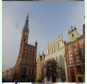
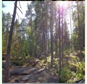
Jesse Read

14 November 2019
Wrocław

Outline

1. Multi-Label Classification and Related Applications
2. Classifier Chains
3. View 1: Probabilistic Classifier Chains
4. View 2: Classifier Chains as a Neural Network
5. The Question of Chain Order
6. Regressor Chains
7. Summary and Perspectives

Multi-Label Classification

Input	Beach	Sunset	Foliage	Urban
	1	0	1	0
	0	1	0	0
	0	1	0	1
	0	1	1	0
	0	0	1	1
	?	?	?	?

Given an instance $\boldsymbol{x}$, we obtain predictions $\widehat{\boldsymbol{y}} = h(\boldsymbol{x})$.

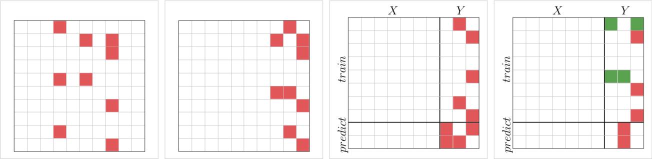
Missing Value Imputation

X_2	X_4	X_1	X_3	X_5
0	0	1	1	0
1	1	?	0	?
0	0	1	0	0
1	1	?	0	1
0	0	0	?	?
1	0	?	1	?

Missing Value Imputation

X_2	X_4	X_1	X_3	X_5
0	0	1	1	0
1	1	?	0	?
0	0	1	0	0
1	1	?	0	1
0	0	0	?	?

X_2	X_4	X_1	X_3	X_5
1	0	?	1	?

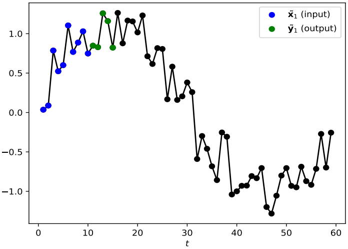


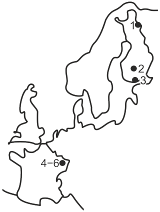
Also applicable to recommender systems.

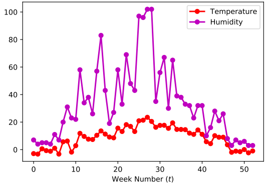
Time Series Forecasting

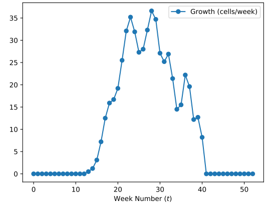
e.g., series $\{19, 21, 24, 23, 20, 17, 15, 12, 13, \ldots, 7, 9, 10, ?, ?, ?, \ldots\}$:

X_{t-3}	X_{t-2}	X_{t-1}	X_t	X_{t+1}	X_{t+2}
19	21	24	23	20	17
21	24	23	20	17	15
24	23	20	17	15	12
23	20	17	15	12	13
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
7	9	10	?	?	?

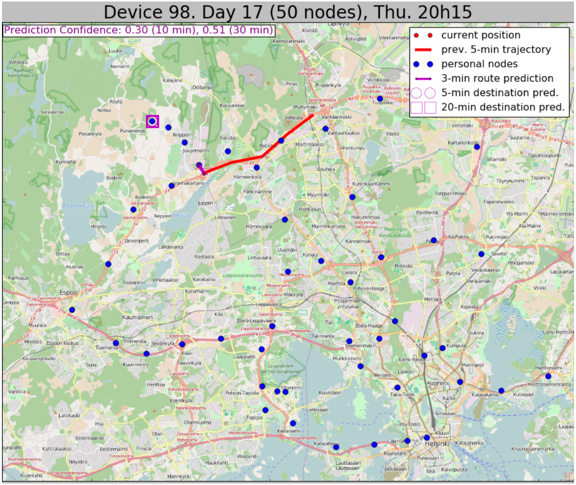








Predicting Celular Growth in Scots Pine across 6 Sites

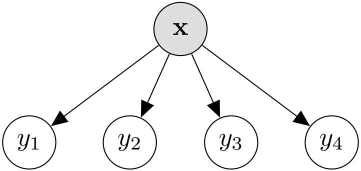

Device 98. Day 17 (50 nodes), Thu. 20h15
Prediction Confidence: 0.30 (10 min), 0.51 (30 min)
● ● current position
prev. 5-min trajectory
● ● personal nodes
3-min route prediction
○ ○ 5-min destination pred.
□ □ 20-min destination pred.
Espoo
Kauniainen
Helsinki

Other topics (a selection, found in Google Scholar citing MEKA):

- *[. . .] Multi-label Sentiment Classification of Health Forums*
- *Using Multi-Label Classification for Improved Question Answering*
- *Predictive Skill Based Call Routing [. . .]*
- *[. . .] Methods for Prediagnosis of Cervical Cancer*
- *[. . .] Expert Systems for Reasoning in Clinical Depressive Disorders*
- *Multi-label classification for intelligent health risk prediction*
- *Deep learning based multi-label classification for surgical tool presence detection in laparoscopic videos*
- *Spectral features for audio based vehicle and engine classification*
- *Ensemble-Based Location Tracking Using Passive RFID*
- *[. . .] big data streams analysis: The case of object trajectory prediction*
- *Multi-task network embedding*
- *Multi-Target Classification and Regression in Wineinformatics*

Binary Relevance: The Baseline

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



The binary relevance method = *one binary classifier trained for each label*, i.e., independent models.

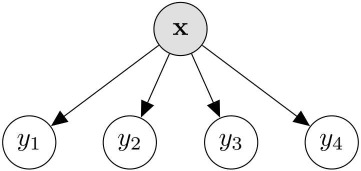
Binary Relevance: The Baseline

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

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

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

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



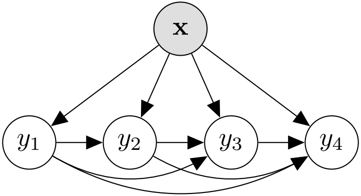
The binary relevance method = *one binary classifier trained for each label*, i.e., independent models.

Outline

1. Multi-Label Classification and Related Applications
2. Classifier Chains
3. View 1: Probabilistic Classifier Chains
4. View 2: Classifier Chains as a Neural Network
5. The Question of Chain Order
6. Regressor Chains
7. Summary and Perspectives

Classifier Chains

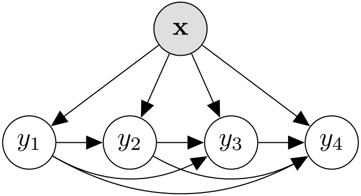
A chain of classifiers:



where the output of each classifier becomes an additional feature for all following classifiers.

- A model of label dependence
- A transformation method (*base classifier* as a hyperparameter)

Read et al., ECML-PKDD 2009; Kajdanowicz and Kazienko, CCI-SSM 2009



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

$\tilde{\boldsymbol{x}}$	$\widehat{y}_1$

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

$\tilde{\boldsymbol{x}}$	$\widehat{y}_1$	$\widehat{y}_2$

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

$\tilde{\boldsymbol{x}}$	$\widehat{y}_1$	$\widehat{y}_2$	$\widehat{y}_3$

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

$\tilde{\boldsymbol{x}}$	$\widehat{y}_1$	$\widehat{y}_2$	$\widehat{y}_3$	$\widehat{y}_4$

- Widely applicable in many domains, with
- Off-the-shelf binary classifiers
- State-of-the-art predictive performance
- Similar running time as independent classifiers in practice

- Widely applicable in many domains, with
- Off-the-shelf binary classifiers
- State-of-the-art predictive performance
- Similar running time as independent classifiers in practice

But how does it work?

What is it optimising?

Can we get a better chain?

Does it work with continuous outputs?

Is it still relevant?

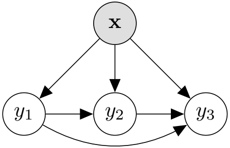
Outline

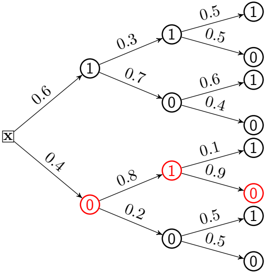
1. Multi-Label Classification and Related Applications
2. Classifier Chains
3. View 1: Probabilistic Classifier Chains
4. View 2: Classifier Chains as a Neural Network
5. The Question of Chain Order
6. Regressor Chains
7. Summary and Perspectives

View 1: Classifier Chains as a Probabilistic Model

$$\widehat{y}_j = h_j(\boldsymbol{x}) = \underset{y_j \in \{0,1\}}{\operatorname{argmax}} P(y_j | \boldsymbol{x}, y_1, \ldots, y_{j-1})$$

e.g., logistic regression, then:





$$\widehat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \{0,1\}^L}{\operatorname{argmax}} P(y_1 | \boldsymbol{x}) \prod_{j=2}^{L} P(y_j | \boldsymbol{x}, y_1, \ldots, y_{j-1})$$

as proposed in probabilistic classifier chains[1].

[1] Dembczyński, Cheng, and Hüllermeier, ICML 2010; and followup work

$$\boxed{\widehat{\boldsymbol{y}} = \underset{\boldsymbol{y}\in\{0,1\}^L}{\operatorname{argmax}}\, P(y_1|\boldsymbol{x}) \prod_{j=2}^{L} P(y_j|\boldsymbol{x}, y_1, \ldots, y_{j-1})}$$

- It's a MAP estimate, optimising subset 0/1 loss,

$$\ell(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = 1_{\boldsymbol{y}\neq\widehat{\boldsymbol{y}}}$$

- Inference is a search
  - standard classifier chain = greedy search.
  - exhaustive search: try all 2^L combinations/paths
  - much room for trade-off[2]

[2] As surveyed in Mena et al., Wiley Int. Rev. 2016

$$\boxed{\widehat{\boldsymbol{y}} = \underset{\boldsymbol{y} \in \{0,1\}^L}{\text{argmax}}\, P(y_1|\boldsymbol{x}) \prod_{j=2}^{L} P(y_j|\boldsymbol{x}, y_1, \ldots, y_{j-1})}$$

- It's a MAP estimate, optimising subset 0/1 loss,

$$\ell(\boldsymbol{y}, \widehat{\boldsymbol{y}}) = 1_{\boldsymbol{y} \neq \widehat{\boldsymbol{y}}}$$

- Inference is a search
  - standard classifier chain = greedy search.
  - exhaustive search: try all 2^L combinations/paths
  - much room for trade-off[2]

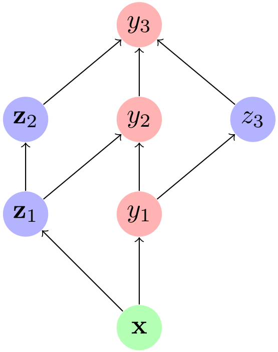
Empirical observation: Classifier chains also outperforms baseline methods on Hamming loss.

[2] As surveyed in Mena et al., Wiley Int. Rev. 2016

Outline

1. Multi-Label Classification and Related Applications
2. Classifier Chains
3. View 1: Probabilistic Classifier Chains
4. View 2: Classifier Chains as a Neural Network
5. The Question of Chain Order
6. Regressor Chains
7. Summary and Perspectives

View 2: Classifier Chains as a Deep Neural Network

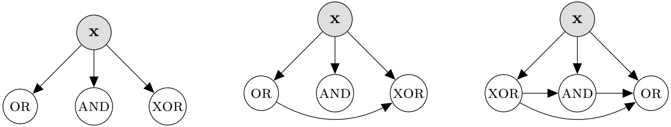


with delay nodes $z = f(x) = x$

- Forward propagation = greedy inference
- It's deep in the label space!
- labels = feature space; "hidden nodes" for free

Read and Hollmén, IDA 2014; Cisse, Al-Shedivat, and Bengio, ICML 2016

Consider, where $\boldsymbol{x} \in \{0,1\}^2$, and labels are logical operations:



- Labels are conditionally independent, given a good choice of base classifier
- Only one of these models works with ‘default parameters’ (linear SVM, greedy inference)

Outline

1. Multi-Label Classification and Related Applications
2. Classifier Chains
3. View 1: Probabilistic Classifier Chains
4. View 2: Classifier Chains as a Neural Network
5. The Question of Chain Order
6. Regressor Chains
7. Summary and Perspectives

Chain Order/Structure

An important question for *accuracy* (good structure), *scalability* (sparse structure), and *interpretability*.

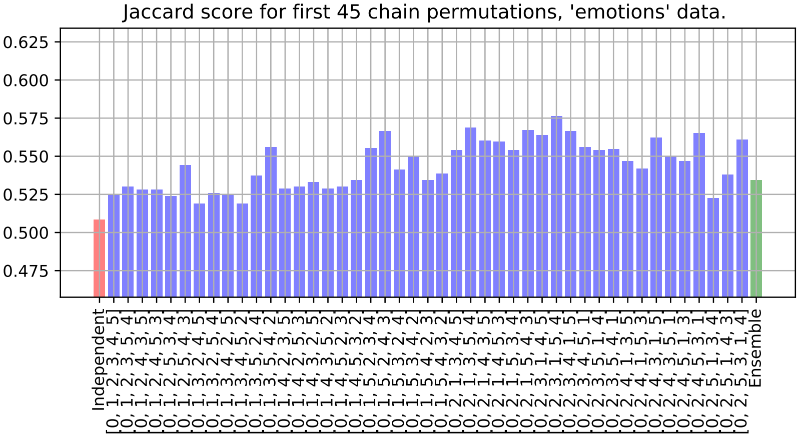
The literature proposes:

1. Random (ensembles). Effective but boring (and large)
2. Use an existing hierarchy. Not worth the effort in parsing (in terms of accuracy)[3]
3. Based on label dependence[4]. It depends (recall toy example!)
4. Based on predictive power of individual classifiers. Still, it depends!
5. Trial and error (Search the label-structure space[5]): Slow!

[3] Puurula, Read, and Bifet, Kaggle 2014 won Kaggle LSHTC14 (large scale *hierarchical* text classification), *ignoring the hierarchy*!

[4] Zaragoza et al., IJCAI 2011, Kajdanowicz and Kazienko, FQAS 2013; and others

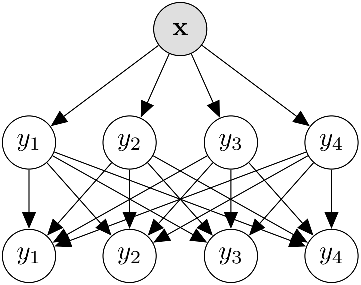
[5] Kumar et al., ECML-PKDD 2012; Read, Martino, and Luengo, Pat. Rec. 2014; Gasse, U. Lyon 2017; etc.



- Note:
  - large difference chain $[\ldots, 3, 4]$ vs $[\ldots, 4, 3]$
  - small difference chain $[\ldots, 1, 5, 2, 4, 3]$ vs $[\ldots, 2, 5, 3, 1, 4]$
- Many local maxima; hill-climbing search can work
- No need to discard suboptimal models: use for dynamic chains

Alternatives to Chaining: Stacking and Undirected Nets

Stacking:



Strong connection with chaining: $\widehat{y}_j$ used to predict y_k

Outline

1. Multi-Label Classification and Related Applications
2. Classifier Chains
3. View 1: Probabilistic Classifier Chains
4. View 2: Classifier Chains as a Neural Network
5. The Question of Chain Order
6. Regressor Chains
7. Summary and Perspectives

What about Regressor Chains?

X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3
2.12	1.217	-0.675	-0.451	0.342	37.00	25	0.88
-0.717	-0.826	0.064	-0.259	-0.717	-22.88	22	0.22
1.374	0.95	0.175	-0.006	-0.522	19.21	12	0.25
1.392	-0.496	-2.441	-1.012	0.268	88.23	11	0.77
1.591	0.208	0.17	-0.207	1.686	?	?	?

Another easy off-the-shelf application of chaining?

- base learner – linear regression?

What about Regressor Chains?

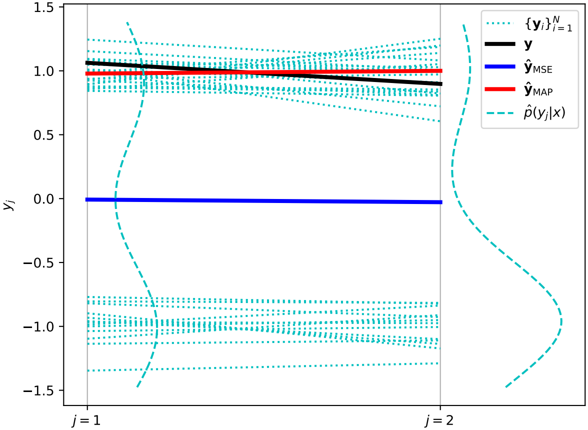
X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3
2.12	1.217	-0.675	-0.451	0.342	37.00	25	0.88
-0.717	-0.826	0.064	-0.259	-0.717	-22.88	22	0.22
1.374	0.95	0.175	-0.006	-0.522	19.21	12	0.25
1.392	-0.496	-2.441	-1.012	0.268	88.23	11	0.77
1.591	0.208	0.17	-0.207	1.686	?	?	?

Another easy off-the-shelf application of chaining?

- base learner – linear regression?

Compared to individual classifiers, you get no improvement from classifier chains in the *best case*, and potentially catastrophic results otherwise (error propagation in $\mathbb{R}^L$!)

- Linear regression = chain collapses into $\widehat{y}_j = \boldsymbol{x}\boldsymbol{w}_j$
- Minimizer of squared error = $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{x}] = \boldsymbol{x}\boldsymbol{W}$



- We can use non-linear classifiers[6], but
- can't do tree search for MAP estimate (there's no tree!)

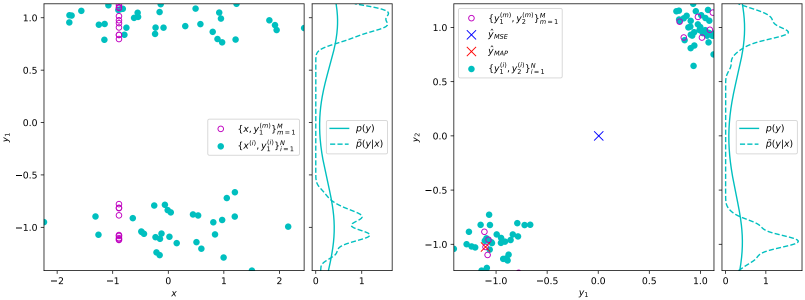
[6] Spyromitros-Xioufis et al., Mach. Learn. 2016

Probabilistic Regressor Chains

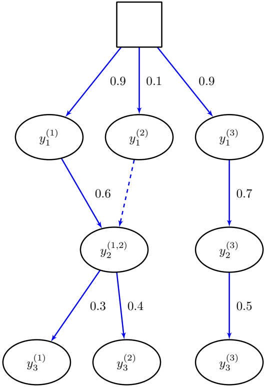
For each $\boldsymbol{x}$, we can build a tree by taking samples at each label/step

$$\{y_j^{(m)}\}_{m=1}^{M} \sim p(y|\boldsymbol{x}, y_1^{(m)}, \ldots, y_{j-1}^{(m)})$$

(given suitable p).



Read and Martino, ArXiv 2019: Probabilistic Regressor Chains with Monte Carlo Methods



$L = 3$ labels/steps. A probability tree built on $M = 3$ samples per step

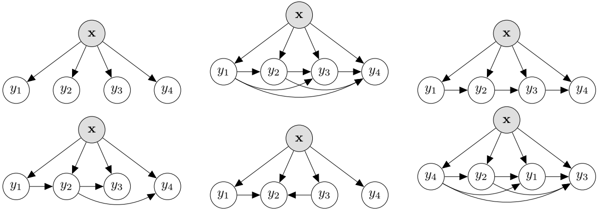
- We build a tree from the samples
- For an approx. MAP estimate: take the path of highest payoff.

Outline

1. Multi-Label Classification and Related Applications
2. Classifier Chains
3. View 1: Probabilistic Classifier Chains
4. View 2: Classifier Chains as a Neural Network
5. The Question of Chain Order
6. Regressor Chains
7. Summary and Perspectives

Summary so far

Chaining methods are flexible, widely applicable, competitive.



Classifier chains . . .

- Have a probabilistic interpretation,
- are mode seekers, via probability tree search
- but also provide representation power via non-linearity.

Regressor chains . . .

- have no natural non-linearity, not great off-the-shelf,
- but probabilistic chains help find modes; interpretable.

There are clear connections with many other methods.

Issues worth mentioning

- Scalability (features quadratic wrt number of labels).
- Overlap/competitiveness vs deep neural network architectures

Is ‘chaining’ still relevant vs deep learning?

- Interpretability – what can the chain tell us about the data?
- Other issues (that affect multi-label learning in general):
    - Class imbalance
    - Weak labels
    - . . .
    - What metrics should we be using?

Title recipe of many recent multi-label papers:

> "*Deep X for Extreme Multi-label [Text] Classification*"

where X $\subset$ {Neural Networks, Convolution, Attention, LSTM, Seq2Seq, Adversarial, Sparse, Autoencoder, Latent, . . . }.

Title recipe of many recent multi-label papers:

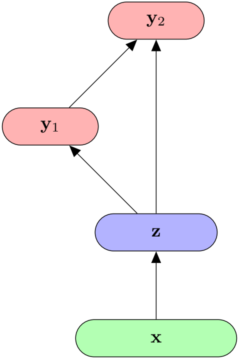
“*Deep X for Extreme Multi-label [Text] Classification*”

where X $\subset$ {Neural Networks, Convolution, Attention, LSTM, Seq2Seq, Adversarial, Sparse, Autoencoder, Latent, . . . }.

It’s difficult to justify chaining in this context, but Chains

- are still competitive useful for ‘un-extreme’ learning
- off interpretability
- a method of transfer learning
- But why contrast? Recall: chaining *is* a kind of deep network.

Chain-Inspired Deep Architectures

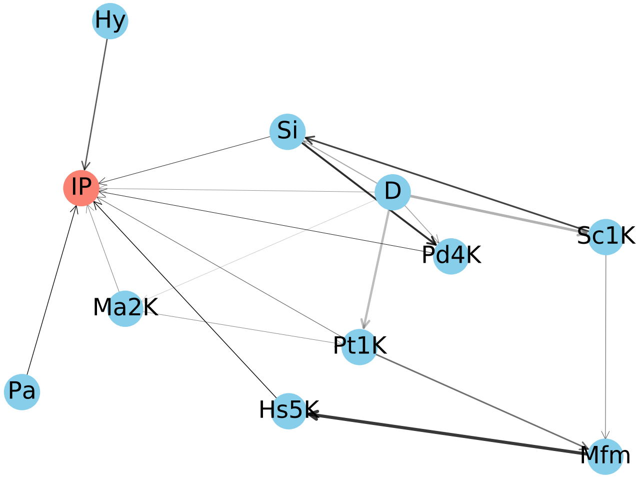


where $\boldsymbol{y}_h$ are based on subsets of labels $S_h \subseteq \mathcal{L}$, *possibly overlapping*.

Combining the advantages of chaining, probabilistic interpretation, stacking, deep learning frameworks.

Read and Hollmén, IDA 2014; Cisse, Al-Shedivat, and Bengio, ICML 2016

Exploration into Interpretation



‘Feature chains’ for predicting ‘paradoxical insomnia’ (IP)

Manuscript in preparation

Multi-Output Learning with Chaining

Jesse Read

Thank You!