

Probabilistic Label Trees in Vowpal Wabbit

Kalina Jasinska-Kobus¹ Marek Wydmuch¹ Mikhail Kuznetsov²
Róbert Busa-Fekete³ Krzysztof Dembczyński^{1,2}

¹ *Institute of Computing Science, Poznan University of Technology, Poland*

² *Yahoo! Research, New York, USA*

³ *Google Research, New York, USA*



The NeurIPS 2020 Vowpal Wabbit Workshop

Agenda

- 1 Extreme multi-label classification (XMLC)
- 2 Formal setting
- 3 Probabilistic label trees (PLTs)
- 4 PLT in Vowpal Wabbit

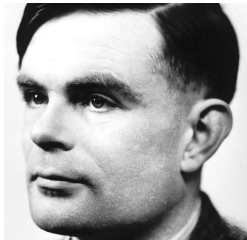
Agenda

- 1 Extreme multi-label classification (XMLC)
- 2 Formal setting
- 3 Probabilistic label trees (PLTs)
- 4 PLT in Vowpal Wabbit

Extreme multi-label classification (XMLC) is a problem of **labeling** an item with a **small** set of labels out of an **extremely large** number of potential labels.

Document tagging

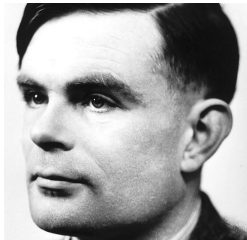
WIKIPEDIA
The Free Encyclopedia



Alan Turing

Document tagging

WIKIPEDIA
The Free Encyclopedia



Alan Turing

Alan Turing | 1912 births | 1954 deaths
20th-century mathematicians | 20th-century philosophers
Academics of the University of Manchester Institute of Science and Technology
Alumni of King's College | Cambridge Artificial intelligence researchers
Atheist philosophers | Bayesian statisticians | British cryptographers | British logicians
British anti-fascists | British long-distance runners | British people of World War II
Computability theorists | Computer designers | English atheists
English computer scientists | English inventors | English logicians
English long-distance runners | English mathematicians
English people of Scottish descent | English philosophers | Former Protestants
Fellows of the Royal Society | Gay men | Gay academics | GCHQ people
Government Communications Headquarters people | History of artificial intelligence
Inventors who committed suicide | Male long-distance runners
Mathematicians who committed suicide | Officers of the Order of the British Empire
Bletchley Park people | People educated at Sherborne School
People from Maida Vale | People from Wilmslow
People prosecuted under anti-homosexuality laws | Princeton University alumni
Programmers who committed suicide | People who have received posthumous pardons
Recipients of British royal pardons | Academics of the University of Manchester
Suicides by cyanide poisoning | Suicides in England | Theoretical computer scientists

Document tagging

WIKIPEDIA
The Free Encyclopedia



text article
 \Rightarrow features x

Alan Turing

Alan Turing | 1912 births | 1954 deaths
20th-century mathematicians | 20th-century philosophers
Academics of the University of Manchester Institute of Science and Technology
Alumni of King's College | Cambridge Artificial intelligence researchers
Atheist philosophers | Bayesian statisticians | British cryptographers | British logicians
British anti-fascists | British long-distance runners | British people of World War II
Computability theorists | Computer designers | English atheists
English computer scientists | English long-distance runners
English people of Scottish descent | English philosophers | Former Protestants
Fellows of the Royal Society | Gay men | Gay academics | GCHQ people
Government Communications Headquarters people | History of artificial intelligence
Inventors who committed suicide | Male long-distance runners
Mathematicians who committed suicide | Officers of the Order of the British Empire
Bletchley Park people | People educated at Sherborne School
People from Maida Vale | People from Wilmslow
People prosecuted under anti-homosexuality laws | Princeton University alumni
Programmers who committed suicide | People who have received posthumous pardons
Recipients of British royal pardons | Academics of the University of Manchester
Suicides by cyanide poisoning | Suicides in England | Theoretical computer scientists

categories \Rightarrow labels y

Information retrieval for search and advertising

Tagging ads with relevant queries:¹

Ad:



Do you know how much you could save on car insurance by switching to GEICO?

Get an online quote: Enter ZIP [GO](#)

Get a quote by phone: [1-800-842-1500](#)

Find a local agent: Enter ZIP [GO](#)

Need other insurance?

ATV	Homeowners	Overseas
Boat	ID Theft Protection	Renters
Commercial Auto	Life	RV
Chevy/Co-op	Mobile Home	Umbrella
Flood	Motorcycle	

Already a policyholder? [Log in.](#)

GEICO

© 1999-2012 GEICO

¹ Rahul Agrawal, Archit Gupta, Yashoteja Prabhu, and Manik Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *WWW*, 2013

Information retrieval for search and advertising

Tagging ads with relevant queries:¹

Ad:



Do you know how much you could save on car insurance by switching to GEICO?

Get an online quote: Enter ZIP [GO](#)

Get a quote by phone: [1-800-841-3500](#)

Find a local agent: Enter ZIP [GO](#)

Need other insurance?

ATV	Homeowners	Overseas
Boat	ID Theft Protection	Renters
Commercial Auto	Life	RV
Chevy/Co-op	Mobile Home	Umbrella
Flood	Motorcycle	

Already a policyholder? [Log in.](#)

GEICO

© 1999-2012 GEICO

Queries:

"geico car insurance"

"geico auto insurance"

"geico insurance"

"www geico com"

"care geicos"

"geico com"

"need cheap auto insurance"

"wisconsin car insurance quotes"

"cheap auto insurance florida"

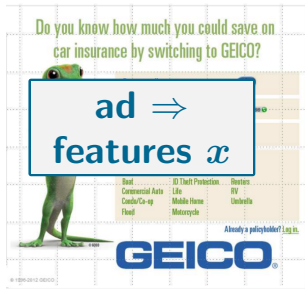
"all state auto insurance coupon code"

¹ Rahul Agrawal, Archit Gupta, Yashoteja Prabhu, and Manik Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *WWW*, 2013

Information retrieval for search and advertising

Tagging ads with relevant queries:¹

Ad:



Queries:

"geico car insurance"

"geico auto insurance"

"geico insurance"

"www geico com"

"care geicos"

"geico com"

"need cheep auto insurance"

"wisconsin car insurance quotes"

"cheap auto insurance florida"

"all state auto insurance coupon code"

queries =>
labels y

¹ Rahul Agrawal, Archit Gupta, Yashoteja Prabhu, and Manik Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *WWW*, 2013

Information retrieval for search and advertising

Predicting relevant products for queries:²

The screenshot shows the Amazon India homepage with the search bar containing 'headphones'. The search results page displays 1-16 of over 20,000 results. On the left, there are filters for 'Avg. Customer Review' (4 stars & up, 3 stars & up, 2 stars & up, 1 star & up), 'Department' (Audio Headphones, Over-Ear Headphones, On-Ear Headphones, Earbud & In-Ear Headphones, Music Recording Equipment, DJ Equipment, DJ Headphones, and a link to 'See All 24 Departments'), 'Brand' (Sony, Mpow, Cowin, Bose, OneOdio, Apple, LETSCOM, and a link to 'See more'), and 'Price'. The main content area is titled 'Shop by feature' and includes buttons for DJ Style, Foldable, Lightweight, Microphone, Noise-Cancelling, Noise-Isolating, Phone Control, Sports & Exercise, Tangle-Free Cord, and Volume Control. Below this, two product listings are shown. The first is 'Amazon's Choice' for 'Sony MDRZX110/BLK ZX Series Stereo Headphones (Black)' with a 4-star rating, 23,799 reviews, and a price of \$59.99. The second is 'COWIN E7 Active Noise Cancelling Headphones Bluetooth Headphones with Microphone Deep Bass Wireless Headphones Over Ear, Comfortable Protein Earpads, 30 Hours Playtime for Travel/Work, Black' with a 4-star rating, 41,338 reviews, and a price of \$59.99 with a 20% discount coupon.

amazon

Deliver to Poland

Today's Deals Customer Service Gift Cards Registry Sell

1-16 of over 20,000 results for "headphones"

Sort by: Featured

Avg. Customer Review

- ★★★★★ & Up
- ★★★★☆ & Up
- ★★★☆☆ & Up
- ★★☆☆☆ & Up

Department

- Audio Headphones
- Over-Ear Headphones
- On-Ear Headphones
- Earbud & In-Ear Headphones
- Music Recording Equipment
- DJ Equipment
- DJ Headphones
- [See All 24 Departments](#)

Brand

- ☐ Sony
- ☐ Mpow
- ☐ cowin
- ☐ Bose
- ☐ OneOdio
- ☐ Apple
- ☐ LETSCOM
- [See more](#)

Price

Shop by feature

- DJ Style
- Foldable
- Lightweight
- Microphone
- Noise-Cancelling
- Noise-Isolating
- Phone Control
- Sports & Exercise
- Tangle-Free Cord
- Volume Control

Amazon's Choice

Sony MDRZX110/BLK ZX Series Stereo Headphones (Black)

★★★★☆ ~ 23,799

Ships to Poland

COWIN E7 Active Noise Cancelling Headphones Bluetooth Headphones with Microphone Deep Bass Wireless Headphones Over Ear, Comfortable Protein Earpads, 30 Hours Playtime for Travel/Work, Black

★★★★☆ ~ 41,338

\$59.99

Save 20% with coupon

² Tharun Kumar Reddy Medini, Qixuan Huang, Yiqiu Wang, Vijai Mohan, and Anshumali Shrivastava. Extreme classification in log memory using count-min sketch: A case study of amazon search with 50m products. In *NeurIPS*, 2019

Information retrieval for search and advertising

Predicting relevant products for queries:²

query + user information \Rightarrow features x

The screenshot shows an Amazon search results page for the query "headphones". The page includes a navigation bar at the top with links like "Deliver to Poland", "Today's Deals", "Customer Service", "Gift Cards", "Registry", and "Sell". Below the navigation bar, the search results are displayed. On the left, there are filters for "Avg. Customer Review" (showing star ratings and "4 & Up"), "Department" (listing various headphone types), and "Brand" (listing brands like Sony, Mpow, Cowin, Bose, OneOdio, Apple, and LETSCOM). The main content area shows a list of products. The first product is "Sony MDRZX110/BLK ZX Series Stereo Headphones (Black)", which is marked as "Amazon's Choice". It has a 4.5-star rating and 23,799 reviews. The second product is "COWIN E7 Active Noise Cancelling Wireless Headphones Over Ear, Co", which has a 4.3-star rating and 41,338 reviews. A price tag of \$59.99 is shown for the second product, with a "Save 20% with coupon" label. A text box on the right side of the screenshot contains the text "clicked/bought items \Rightarrow labels y ".

² Tharun Kumar Reddy Medini, Qixuan Huang, Yiqiu Wang, Vijai Mohan, and Anshumali Shrivastava. Extreme classification in log memory using count-min sketch: A case study of amazon search with 50m products. In *NeurIPS*, 2019

Recommendation systems

User-to-item and item-to-item recommendations:³

Your recently viewed items and featured recommendations

Recommendations & Popular Items

Page 1 of 3



Anker Power Port II
★★★★☆ 37
EUR 39.99 ✓prime



USB C to DisplayPort Cable – 4 K @ 60Hz, CHOETECH (4ft/1.2 m) USB 3.1 Type C...
★★★★☆ 157
EUR 13.99 ✓prime



Mini Display Port to Display Port Cable Posugear Gold Plated Thunderbolt Mini DP to...
★★★★☆ 23
EUR 8.99 ✓prime



Apple Lightning USB Cable 1 m, MQGJZM/A
★★★★☆ 3
EUR 21.00 ✓prime



Anker USB C Charger PowerPort PD 2 Wall Charger 30W Dual Port with 18W Power...
★★★★☆ 5
EUR 18.99 ✓prime



Tech Mat Pencil Lightning Cable Charge Adapter for Apple iPad Pro Female to Female (.75 Inch)
★★★★☆ 45
EUR 7.99 ✓prime

Sponsored products related to this item

Page 1 of 27



Wihlasure Bluetooth Kopfhörer Kabellos V5.0 Touch Bluetooth Headset Sport Ohrhörer ...
★★★★☆ 310
EUR 51.99 ✓prime



TriLink USB Typ C auf 3,5mm Audio Kopfhörer Jack Adapter(Hi-Resolution Audio & DAC ...
★★★★☆ 6
EUR 12.99 ✓prime



AUSDOM M06 Bluetooth Kopfhörer On Ear mit Mikrofon Over-Ear HiFi Kopfhörer, Bluetooth...
★★★★☆ 23
EUR 31.99 ✓prime



Kopfhörer auf Ohr, VOGEL Falzbare kabelgebundene on Ear Kopfhörer mit Verbessertem ...
★★★★☆ 73
EUR 13.99 ✓prime



Teufel BAMSTER Schwarz Bluetooth NFC Android ios spotify aptx Bluetooth Lautsprecher...
★★★★☆ 28
EUR 125,00 ✓prime



Bluetooth Kopfhörer on Ear Wireless HiFi Stereo Headset mit Mikrofon Freisprechlein...
★★★★☆ 14
EUR 23.99 ✓prime

Ad feedback

³ Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. Learning tree-based deep model for recommender systems. In *KDD*, 2018

Recommendation systems

User-to-item and item-to-item recommendations:³

Your recently viewed items and featured recommendations

Recommendations & Popular Items Page 1 of 3

user's history and features
 \Rightarrow **features x**

Anker Power Port II

★★★★☆ 37

EUR 39.99 ✓prime

'USB C to DisplayPort Cable – 4 K @ 60Hz), CHOETECH (4ft/1.2 m) USB 3.1 Type C...

★★★★☆ 157

EUR 13.99 ✓prime

Mini Display Port to Display Port Cable Posugear Gold Plated Thunderbolt Mini DP to...

★★★★☆ 23

EUR 8.99 ✓prime

Apple Lightning USB Cable 1 m, MQGJZM/A

★★★★☆ 3

EUR 21.00 ✓prime

Anker USB C Charger PowerPort PD 2 Wall Charger 30W Dual Port with 18W Power...

★★★★☆ 5

EUR 18.99 ✓prime

Tech Mat Pencil Lightning Cable Charge Adapter for Apple iPad Pro Female to Female (.75 Inch)

★★★★☆ 45

EUR 7.99 ✓prime

Sponsored products related to this item Page 1 of 27

item \Rightarrow features x

Wihcure Bluetooth Kopfhörer Kabellos V5.0 Touch Bluetooth Headset Sport Ohrhörer ...

★★★★☆ 310

EUR 51.99 ✓prime

TriLink USB Typ C auf 3,5mm Audio Kopfhörer Jack Adapter(Hi-Resolution Audio & DAC ...

★★★★☆ 6

EUR 12.99 ✓prime

AUSDOM M06 Bluetooth Kopfhörer On Ear mit Mikrophon Over-Ear HIFI Kopfhörer, Bluetooth...

★★★★☆ 23

EUR 31.99 ✓prime

Kopfhörer auf Ohr Faltbare kabelgebundene on Ear Kopfhörer mit Verbessertem ...

★★★★☆ 73

EUR 13.99 ✓prime

clicked/bought items \Rightarrow labels y

Bluetooth NFC Android ios spotify aptx Bluetooth Lautsprecher...

★★★★☆ 28

EUR 125,00 ✓prime

Ear Wireless HIFI Stereo Headset mit Mikrophon Freisprechein...

★★★★☆ 14

EUR 23,99 ✓prime

³ Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. Learning tree-based deep model for recommender systems. In *KDD*, 2018

7 / 28

Agenda

- ① Extreme multi-label classification (XMLC)
- ② Formal setting
- ③ Probabilistic label trees (PLTs)
- ④ PLT in Vowpal Wabbit

Formal setting

- Multi-label classification:

$$\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d \xrightarrow{h(\mathbf{x})} \mathbf{y} = (y_1, y_2, \dots, y_m) \in \{0, 1\}^m$$

x_1	x_2	\dots	x_d	y_1	y_2	\dots	y_m
4.0	2.5		-1.5	1	1		0

Formal setting

- **Multi-label classification:**

$$\mathbf{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d \xrightarrow{h(\mathbf{x})} \mathbf{y} = (y_1, y_2, \dots, y_m) \in \{0, 1\}^m$$

x_1	x_2	\dots	x_d	y_1	y_2	\dots	y_m
4.0	2.5		-1.5	1	1		0

- **Goal:** find a classifier $h(\mathbf{x}) : \mathcal{X} \rightarrow \mathcal{R}^m$ with small **expected loss**:

$$L_\ell(\mathbf{h}) = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim P(\mathbf{x}, \mathbf{y})}(\ell(\mathbf{y}, \mathbf{h}(\mathbf{x})))$$

Formal setting

- The **regret** of a classifier \mathbf{h} with respect to ℓ

$$\text{reg}_\ell(\mathbf{h}) = L_\ell(\mathbf{h}) - L_\ell(\mathbf{h}_\ell^*) = L_\ell(\mathbf{h}) - L_\ell^*$$

quantifies the suboptimality of \mathbf{h} compared to the optimal (**Bayes**) classifier:

$$\mathbf{h}_\ell^* = \arg \min_{\mathbf{h}} L_\ell(\mathbf{h})$$

Conditional marginal probability

- **Conditional marginal probability** of a label:

$$\eta_j(\mathbf{x}) = P(y_j = 1|\mathbf{x}) = \sum_{\mathbf{y}:y_j=1} P(\mathbf{y}|\mathbf{x})$$

Conditional marginal probability

- **Conditional marginal probability** of a label:

$$\eta_j(\mathbf{x}) = P(y_j = 1|\mathbf{x}) = \sum_{\mathbf{y}:y_j=1} P(\mathbf{y}|\mathbf{x})$$

- Bayes classifiers for popular MLC losses, such as:
 - ▶ Hamming loss,
 - ▶ micro and macro-F1 measure,
 - ▶ precision@ k ,
 - ▶ DCG@ k .

are directly expressed via conditional label probabilities.

Conditional marginal probability

- **Conditional marginal probability** of a label:

$$\eta_j(\mathbf{x}) = P(y_j = 1|\mathbf{x}) = \sum_{\mathbf{y}:y_j=1} P(\mathbf{y}|\mathbf{x})$$

- Bayes classifiers for popular MLC losses, such as:
 - ▶ Hamming loss,
 - ▶ micro and macro-F1 measure,
 - ▶ precision@ k ,
 - ▶ DCG@ k .

are directly expressed via conditional label probabilities.

- Hence accurate estimation of $\eta_j(\mathbf{x})$ is crucial for solving XMLC problems.

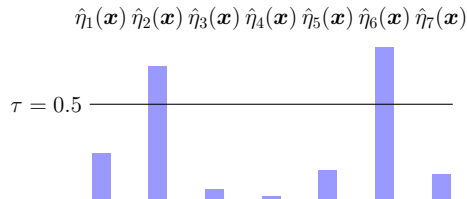
MLC under hamming loss

- **Hamming loss:**

$$\ell_H(\mathbf{y}, \mathbf{h}(\mathbf{x})) = \frac{1}{m} \sum_{j=1}^m \mathbb{I}[y_j \neq h_j(\mathbf{x})]$$

- Sparse labels \Rightarrow Hamming loss of an **all-zero** classifier close to **0**
- **The optimal strategy:**⁴

$$h_j^*(\mathbf{x}) = \mathbb{I}[\eta_j(\mathbf{x}) > 0.5]$$



⁴ Krzysztof Dembczynski, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88, 2012

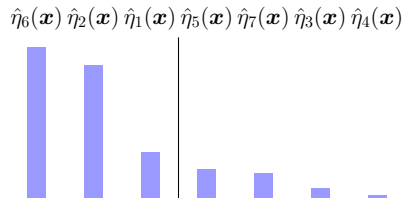
MLC under precision@ k

- **Precision at position k (p@ k):**

$$\text{p@}k(\mathbf{y}, \mathbf{h}, \mathbf{x}) = \frac{1}{k} \sum_{j \in \hat{\mathcal{Y}}_k} \mathbb{I}[y_j = 1],$$

where $\hat{\mathcal{Y}}_k$ is a set of k labels predicted by \mathbf{h} .

- **The optimal strategy:**⁵
select top k labels according to $\eta_j(\mathbf{x})$.



⁵ Marek Wydmuch, Kalina Jasinska, Mikhail Kuznetsov, Róbert Busa-Fekete, and Krzysztof Dembczynski. A no-regret generalization of hierarchical softmax to extreme multi-label classification. In *NeurIPS*, 2018

Computational challenge

Extreme classification \Rightarrow a **large** number of **labels** $m (\geq 10^5)$

\Rightarrow a **large** number of **features** $d (\geq 10^6)$

\Rightarrow a **large** number of **examples** $n (\geq 10^6)$

Computational challenge

- Naive approach: One-vs-All with linear models, $\hat{y} = \mathbf{W}^\top \mathbf{x}$

Problem size:

$$n \geq 10^6, d \geq 10^6, m \geq 10^5 \Rightarrow$$

Complexity:

$$\text{training time} \geq 10^{17}$$

$$\text{space} \geq 10^{11}$$

$$\text{predict time} \geq 10^{11}$$

Computational challenge

- Naive approach: One-vs-All with linear models, $\hat{y} = \mathbf{W}^\top \mathbf{x}$

Problem size:

$$n \geq 10^6, d \geq 10^6, m \geq 10^5 \Rightarrow$$

Complexity:

$$\text{training time} \geq 10^{17}$$

$$\text{space} \geq 10^{11}$$

$$\text{predict time} \geq 10^{11}$$

- Is it really that hard?

Computational challenge

- Naive approach: One-vs-All with linear models, $\hat{y} = \mathbf{W}^\top \mathbf{x}$

Problem size:

$$n \geq 10^6, d \geq 10^6, m \geq 10^5 \Rightarrow$$

Complexity:

$$\text{training time} \geq 10^{17}$$

$$\text{space} \geq 10^{11}$$

$$\text{predict time} \geq 10^{11}$$

- Is it really that hard?**
 - High performance computing resources available

Computational challenge

- Naive approach: One-vs-All with linear models, $\hat{y} = \mathbf{W}^\top \mathbf{x}$

Problem size:

$$n \geq 10^6, d \geq 10^6, m \geq 10^5 \Rightarrow$$

Complexity:

$$\text{training time} \geq 10^{17}$$

$$\text{space} \geq 10^{11}$$

$$\text{predict time} \geq 10^{11}$$

- Is it really that hard?**
 - ▶ High performance computing resources available
 - ▶ Large data \longrightarrow sparse data (sparse features and labels)

Computational challenge

- Naive approach: One-vs-All with linear models, $\hat{y} = \mathbf{W}^\top \mathbf{x}$

Problem size:

$$n \geq 10^6, d \geq 10^6, m \geq 10^5 \Rightarrow$$

Complexity:

$$\text{training time} \geq 10^{17}$$

$$\text{space} \geq 10^{11}$$

$$\text{predict time} \geq 10^{11}$$

- **Is it really that hard?**
 - ▶ High performance computing resources available
 - ▶ Large data \longrightarrow sparse data (sparse features and labels)
 - ▶ Vowpal Wabbit – library for fast online learning exists

Computational challenge

- Naive approach: One-vs-All with linear models, $\hat{y} = \mathbf{W}^\top \mathbf{x}$

Problem size:

$$n \geq 10^6, d \geq 10^6, m \geq 10^5 \Rightarrow$$

Complexity:

$$\text{training time} \geq 10^{17}$$

$$\text{space} \geq 10^{11}$$

$$\text{predict time} \geq 10^{11}$$

- **Is it really that hard?**
 - ▶ High performance computing resources available
 - ▶ Large data \rightarrow sparse data (sparse features and labels)
 - ▶ Vowpal Wabbit – library for fast online learning exists
 - ▶ Learning time for binary model and $n = 10^6$: $\sim 2 - 3s$

Computational challenge

- Naive approach: One-vs-All with linear models, $\hat{y} = \mathbf{W}^\top \mathbf{x}$

Problem size:

$$n \geq 10^6, d \geq 10^6, m \geq 10^5 \Rightarrow$$

Complexity:

$$\text{training time} \geq 10^{17}$$

$$\text{space} \geq 10^{11}$$

$$\text{predict time} \geq 10^{11}$$

- Is it really that hard?**
 - ▶ High performance computing resources available
 - ▶ Large data \rightarrow sparse data (sparse features and labels)
 - ▶ Vowpal Wabbit – library for fast online learning exists
 - ▶ Learning time for binary model and $n = 10^6$: $\sim 2 - 3s$
 - ▶ Learning time $m = 10^5$: $10^5 \times 2s > 2 \text{ day}$ – **not great, not terrible**

Agenda

- 1 Extreme multi-label classification (XMLC)
- 2 Formal setting
- 3 Probabilistic label trees (PLTs)
- 4 PLT in Vowpal Wabbit

Probabilistic label trees (PLTs)⁶

- **PLT**s follows the learning reductions framework: the original problem is decomposed to a **set of binary problems** organized in a tree structure.

⁶ Kalina Jasinska, Krzysztof Dembczynski, Róbert Busa-Fekete, Karlson Pfannschmidt, Timo Klerx, and Eyke Hüllermeier. Extreme F-measure maximization using sparse probability estimates. In *ICML*, 2016

Probabilistic label trees (PLTs)⁶

- **PLT**s follows the learning reductions framework: the original problem is decomposed to a **set of binary problems** organized in a tree structure.
- Path from the root to a leaf corresponds to one and only one label.

⁶ Kalina Jasinska, Krzysztof Dembczynski, Róbert Busa-Fekete, Karlson Pfannschmidt, Timo Klerx, and Eyke Hüllermeier. Extreme F-measure maximization using sparse probability estimates. In *ICML*, 2016

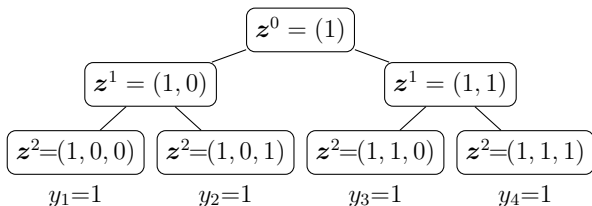
Probabilistic label trees (PLTs)⁶

- **PLT**s follows the learning reductions framework: the original problem is decomposed to a **set of binary problems** organized in a tree structure.
- Path from the root to a leaf corresponds to one and only one label.
- Each label/path is **coded** by vector $\mathbf{z} = (1, z_1, \dots, z_l) \in \mathcal{C}$.

⁶ Kalina Jasinska, Krzysztof Dembczynski, Róbert Busa-Fekete, Karlson Pfannschmidt, Timo Klerx, and Eyke Hüllermeier. Extreme F-measure maximization using sparse probability estimates. In *ICML*, 2016

Probabilistic label trees (PLTs)⁶

- **PLT**s follows the learning reductions framework: the original problem is decomposed to a **set of binary problems** organized in a tree structure.
- Path from the root to a leaf corresponds to one and only one label.
- Each label/path is **coded** by vector $\mathbf{z} = (1, z_1, \dots, z_l) \in \mathcal{C}$.
- An internal node is identified by a **partial** code $\mathbf{z}^j = (z_1, \dots, z_j)$.
- The code does **not** have to be binary.

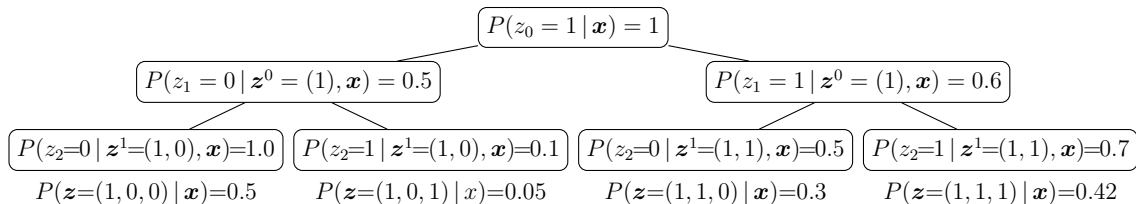


⁶ Kalina Jasinska, Krzysztof Dembczynski, Róbert Busa-Fekete, Karlson Pfannschmidt, Timo Klerx, and Eyke Hüllermeier. Extreme F-measure maximization using sparse probability estimates. In *ICML*, 2016

Probabilistic label trees (PLTs)

- **Factorization** of the conditional marginal probability:

$$\eta_j(\mathbf{x}) = P(\mathbf{z} | \mathbf{x}) = \prod_{i=0}^l P(z_i | \mathbf{z}^{i-1}, \mathbf{x}).$$



Probabilistic label trees (PLTs)

- Efficient learning, requires updating at most $\|\mathbf{y}\|_1$ paths per training example.

Probabilistic label trees (PLTs)

- Efficient learning, requires updating at most $\|\mathbf{y}\|_1$ paths per training example.
- Different tree-search algorithms can be applied for prediction:

Probabilistic label trees (PLTs)

- Efficient learning, requires updating at most $\|\mathbf{y}\|_1$ paths per training example.
- Different tree-search algorithms can be applied for prediction:
 - ▶ beam search (approx. top- k prediction, $\log m$ complexity for balanced trees)

Probabilistic label trees (PLTs)

- Efficient learning, requires updating at most $\|\mathbf{y}\|_1$ paths per training example.
- Different tree-search algorithms can be applied for prediction:
 - ▶ beam search (approx. top- k prediction, $\log m$ complexity for balanced trees)
 - ▶ uniform-cost search (exact top- k prediction, $\log m$ complexity under additional assumptions)

Probabilistic label trees (PLTs)

- Efficient learning, requires updating at most $\|\mathbf{y}\|_1$ paths per training example.
- Different tree-search algorithms can be applied for prediction:
 - ▶ beam search (approx. top- k prediction, $\log m$ complexity for balanced trees)
 - ▶ uniform-cost search (exact top- k prediction, $\log m$ complexity under additional assumptions)
 - ▶ threshold-based search (threshold-based prediction, $\log m$ complexity under additional assumptions)

PLTs and other label tree approaches

- **PLT**s can be treated as multi-label generalization of:
 - ▶ Hierarchical softmax⁷
 - ▶ Conditional probabilistic estimation trees⁸
 - ▶ Nested dichotomies⁹
- Other (non-probabilistic) label tree approaches:
 - ▶ Label embedding trees¹⁰
 - ▶ Filter tree¹¹
 - ▶ HOMER¹²

⁷ Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *AISTATS*, 2005

⁸ Alina Beygelzimer, John Langford, Yury Lifshits, Gregory B. Sorkin, and Alexander L. Strehl. Conditional probability tree estimation analysis and algorithms. In *UAI*, 2009

⁹ John Fox. *Applied regression analysis, linear models, and related methods*. Sage, 1997

¹⁰ Samy Bengio, Jason Weston, and David Grangier. Label embedding trees for large multi-class tasks. In *NIPS*, 2010

¹¹ Alina Beygelzimer, John Langford, and Pradeep Ravikumar. Error-correcting tournaments. In *ALT*, 2009

¹² Grigorios Tsoumakas and Ioannis Katakis I Vlahavas. Effective and efficient multilabel classification in domains with large number of labels. In *ECML/PKDD 2008 Workshop on Mining Multidimensional Data*, 2008

PLTs and other label tree approaches

- The **PLT** model has been used in several XMLC packages:
 - ▶ Parabel¹³
 - ▶ extremeText¹⁴
 - ▶ Bonsai¹⁵
 - ▶ AttentionXML¹⁶
 - ▶ napkinXC¹⁷
- **It is now also available in Vowpal Wabbit!**

¹³Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *WWW*, 2018

¹⁴Marek Wydmuch, Kalina Jasinska, Mikhail Kuznetsov, Róbert Busa-Fekete, and Krzysztof Dembczynski. A no-regret generalization of hierarchical softmax to extreme multi-label classification. In *NeurIPS*, 2018

¹⁵Sujay Khandagale, Han Xiao, and Rohit Babbar. Bonsai - diverse and shallow trees for extreme multi-label classification, 2019

¹⁶Ronghui You, Zihan Zhang, Ziyi Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In *NeurIPS*. 2019

¹⁷Kalina Jasinska-Kobus, Marek Wydmuch, Krzysztof Dembczynski, Mikhail Kuznetsov, and Robert Busa-Fekete. Probabilistic label trees for extreme multi-label classification, 2020

Theoretical guarantees

- **Theorem:**¹⁸ For any distribution P and internal node classifiers f_{z^i} the following holds:

$$|\eta_j(\mathbf{x}) - \hat{\eta}_j(\mathbf{x})| \leq \sum_{i=0}^l \sqrt{\frac{2}{\lambda}} \sqrt{\text{reg}_\ell(f_{z^i} | \mathbf{x})},$$

where $\text{reg}_\ell(f_{z^i} | \mathbf{x})$ is binary classification regret for a strongly proper composite loss ℓ (e.g., logistic loss) and λ is a constant specific for loss ℓ .

¹⁸ Kalina Jasinska-Kobus, Marek Wydmuch, Krzysztof Dembczynski, Mikhail Kuznetsov, and Robert Busa-Fekete. Probabilistic label trees for extreme multi-label classification, 2020

Theoretical guarantees

- **Theorem:**¹⁸ For any distribution P and internal node classifiers f_{z^i} the following holds:

$$|\eta_j(\mathbf{x}) - \hat{\eta}_j(\mathbf{x})| \leq \sum_{i=0}^l \sqrt{\frac{2}{\lambda}} \sqrt{\text{reg}_\ell(f_{z^i} | \mathbf{x})},$$

where $\text{reg}_\ell(f_{z^i} | \mathbf{x})$ is binary classification regret for a strongly proper composite loss ℓ (e.g., logistic loss) and λ is a constant specific for loss ℓ .

- This theorem leads to guarantees for such metrics as Hamming loss, generalized performance metrics, and precision@ k .¹⁸

¹⁸ Kalina Jasinska-Kobus, Marek Wydmuch, Krzysztof Dembczynski, Mikhail Kuznetsov, and Robert Busa-Fekete. Probabilistic label trees for extreme multi-label classification, 2020

Agenda

- ① Extreme multi-label classification (XMLC)
- ② Formal setting
- ③ Probabilistic label trees (PLTs)
- ④ PLT in Vowpal Wabbit

PLT in Vowpal Wabbit

```
$ vw <dataset> --plt <m> --kary_tree <k>
```

- The option `--plt <m>`, where `<m>` is the number of distinct labels, directs vw to perform multi-label classification using a PLT.

PLT in Vowpal Wabbit

```
$ vw <dataset> --plt <m> --kary_tree <k>
```

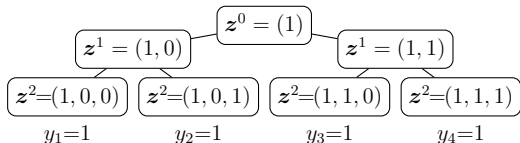
- The option `--plt <m>`, where `<m>` is the number of distinct labels, directs vw to perform multi-label classification using a PLT.
- Labels must be natural numbers from set $\{1, \dots, \text{<m>}\}$, with `<m>` being the maximum label value.

PLT in Vowpal Wabbit

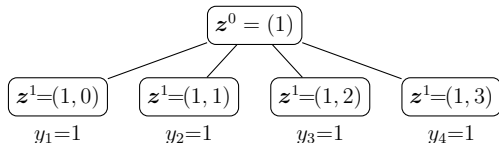
```
$ vw <dataset> --plt <m> --kary_tree <k>
```

- The option `--plt <m>`, where `<m>` is the number of distinct labels, directs vw to perform multi-label classification using a PLT.
- Labels must be natural numbers from set $\{1, \dots, \text{<m>}\}$, with `<m>` being the maximum label value.
- The option `--kary_tree <k>` controls tree node arity.

`--plt 4 --kary_tree 2`



`--plt 4 --kary_tree 4`



PLT in Vowpal Wabbit

```
$ vw <dataset> -i <plt_model> --threshold <thr>/--top_k <k>
```

- The `--threshold <thr>` option indicates the use of the threshold-based prediction, i.e., labels with $\hat{\eta}_j(\mathbf{x})$ above threshold `<thr>` are predicted.
- The `--top_k <k>` option indicates the use of the top- k prediction instead.

	d ($\dim \mathcal{X}$)	m ($\dim \mathcal{Y}$)	n_{train}	n_{test}	avg. $\ \mathbf{y}\ _1$
AmazonCat-13K ¹⁹	203882	13330	1186239	306782	5.04

```
$ wget http://www.cs.put.poznan.pl/mwydmuch/data/amazonCat_train.vw
```

```
$ wget http://www.cs.put.poznan.pl/mwydmuch/data/amazonCat_test.vw
```

¹⁹K. Bhatia, K. Dahiya, H. Jain, A. Mittal, Y. Prabhu, and M. Varma. The extreme classification repository: Multi-label datasets and code, 2016

²⁰https://github.com/VowpalWabbit/vowpal_wabbit/tree/master/demo/plr

Demo

```
$ time vw amazonCat_train.vw -c --multilabel_oaa 13330 \  
  --passes 3 -b 31 -f ovr_weights --holdout_off ⇒ 339.11 mins
```

```
$ time vw amazonCat_test.vw -i ovr_weights ⇒ 23.82 mins
```

Demo

```
$ time vw amazonCat_train.vw -c --multilabel_oaa 13330 \  
  --passes 3 -b 31 -f ovr_weights --holdout_off ⇒ 339.11 mins
```

```
$ time vw amazonCat_test.vw -i ovr_weights ⇒ 23.82 mins
```

```
$ time vw amazonCat_train.vw -c --plt 13330 --kary_tree 16 \  
  --passes 3 -b 31 -f plt_weights --holdout_off ⇒ 24.07 mins
```

```
$ time vw amazonCat_test.vw -i plt_weights --top_k 5 ⇒ 38.24 secs
```

Demo

```
$ time vw amazonCat_train.vw -c --multilabel_oaa 13330 \  
  --passes 3 -b 31 -f ovr_weights --holdout_off ⇒ 339.11 mins
```

```
$ time vw amazonCat_test.vw -i ovr_weights ⇒ 23.82 mins
```

```
$ time vw amazonCat_train.vw -c --plt 13330 --kary_tree 16 \  
  --passes 3 -b 31 -f plt_weights --holdout_off ⇒ 24.07 mins
```

```
$ time vw amazonCat_test.vw -i plt_weights --top_k 5 ⇒ 38.24 secs
```

	$p@1$	$p@3$	$p@5$
vw --multilabel_oaa	91.11	76.40	61.41
vw --plt	91.46	75.01	59.68

Thank you for your attention

Try:

```
$ vw <dataset> --plt <m>
```

Read more:

Probabilistic Label Trees for Extreme Multi-label Classification

<https://arxiv.org/pdf/2009.11218.pdf>