# Generalized test utilities for long-tail performance in extreme multi-label classification

Erik Schultheis<sup>1</sup> Marek Wydmuch<sup>2</sup> Wojtek Kotłowski<sup>2</sup> Rohit Babbar<sup>1,3</sup> Krzysztof Dembczyński<sup>2,4</sup>

<sup>1</sup>Aalto University, Helsinki, Finland <sup>3</sup>University of Bath,UK <sup>4</sup>Yahoo! Research, New York, USA









NeurIPS '23, December 10-16, 2023, New Orleans, USA

#### Multi-label classification:

$$oldsymbol{x} \in \mathcal{X} o oldsymbol{y} \in \mathcal{Y} := \{0,1\}^m$$

e.g.: 
$$\frac{y_1 \quad y_2 \quad y_3 \quad \dots \quad y_m}{y = 0 \quad 1 \quad 1 \quad \dots \quad 0}$$

#### Multi-label classification:

$$\boldsymbol{x} \in \mathcal{X} 
ightarrow \boldsymbol{y} \in \mathcal{Y} := \{0,1\}^m$$

- a large number of labels  $m \ (\geq 10^5)$ ,
- a label vector  $\boldsymbol{y}$  is very sparse,  $\|\boldsymbol{y}\|_1 \ll m$ ,
- many problems are naturally budgeted at k (requirement for a prediction ŷ: ||ŷ||<sub>1</sub> = k),

18201 @ Polent Inter Search	Unacen			Q	=0.	Herita, styr in Accessed & Links -	& Criters	¥.
li Yoday'ı Deals Hegistry Gusterner Servic	sittem se							
a Abanadiarit bochissa batieler	& Hore Anazor-Book Clube Childhort Bo	in Technic Technik hat	is but built of the Hersti-	Institution of 2015 So Far	teur Congeny B	ookhuf		
UNDERSTANDING MACHINE LEARNING	Understanding Mi Algorithms 1st Edi by Sale Sale-Savet Without S 4 # # # # # # # # # # # # # # # # # # #	tion at the best during search search search	: From Theory	to (	6) (m)	Buy new: List Pric S No Import Fres Dep Shipping to Foland Oxform Friday, New Order milden 12 hrs	\$68.00 et \$66.00 Deta ever \$6.56 O.P ost & \$15.25 Cotada ~ sember 24. 48 mine	0 6 2
Contraction of the second	Madeumbour Nex and Distance of the second se	102.01 Planchine State     102.01 Planchine State     102.01 Planchine State     102.01 Planchine State     102.01 Planchine     1	alence, with far-reacting ap adaptes it offers, in a principal machine learning and the air searcation of the basics of 6 strebols. These include a di- longeoment along literate parts Publisher	Elizations. The airs of this ad ways. The book periodics is the or Faid, the book covers as a reaction of the source structure targets including structures for PLODODICS data	n I I I	or tabled answryt November 21 D tablet in Stock Dy 1 V Add to Royt Unjo han Jameses	Cart Cart	
foll over image to most in	10	<b>H</b> U	Bo	83	5	billy Amouna Natures Related	ani Lantill Ann FA 20	
Read sample	110900132 876-0	101967-05 54	Conferinge Referencing Press	May 10, 2014		Payment Securit to Add a gift receip returns	t for easy	
Del Deler     Del Deler     Del Del     Del Del     Del Del     Del						Bay sont 515.21 AMTIN UM		
Frequently bought togethe						New & Used (25) ht	m 1991	,
MACHINE	1076	AREA TO THE CONSTRUCTION OF THE CONSTRUCT OF THE CONSTRU	Tetal price \$172.51			amagonbo	ok clubs	
Carlos Carlos	· ·		AM #L3 N-CHR			See Ci	els .	
and the second sec						Not in a club?	Learn more	
This litere Understanding Hashine Learning men Theory in Appendixis 1600° 1624° Similar iterns that may deliver to you	enis of Soloblod Holewales Data Mining, Inference Color, Second Editory.	for Madrine Learning		74	pe 1 of T		lana	red Ø
			Linca Randon Recipionedada Randon Randon	CONSTRUCTO	>			
Machine Learning Campula Harchese Estavolt Machine	non and Statistical Learning Da Assuing sected Printing, Information, and	a Machine Learning (Information Science a	Optimization for Machine Learning A	Leaving An Introduction (Magdaw				
**************************************	Mass Prediction, Second	- Christopher 14. Bishop A A A A A A 7728	-Own C. Ingernal ******	Hards F. Harphy				
142.99 Handcaver 113.71 dragang CELECE	In Finday Hotology	160.56	10-thow 135.84	190.08				
Counting 195,02	162.49 11111.domina	Cell & as soon as Perday, Marr	Det it is soon is Pédig, Nov M	BALLED BOTH PERMIT				
\$15.25 un	ging	\$11.05-shipping	\$13.45 shipping	\$18.34 shoping				

#### Multi-label classification:

$$\boldsymbol{x} \in \mathcal{X} 
ightarrow \boldsymbol{y} \in \mathcal{Y} := \{0,1\}^m$$

- a large number of labels  $m \ (\geq 10^5)$ ,
- a label vector  $\boldsymbol{y}$  is very sparse,  $\|\boldsymbol{y}\|_1 \ll m$ ,
- many problems are naturally budgeted at k (requirement for a prediction ŷ: ||ŷ||<sub>1</sub> = k),

Advantation Registry Constraint Service	A New Address door of the	Dittor hair	in Declarat Process	her built of the same	but had at 100 ****	. The fact	too to		
Province party with fickants . But beings	a Para Anazar Sook Outra	CONTRACTORIES	ren and the field of	mus muse of the Month.	man. monto (f. 2010 50 8	- Congen			
Back to wood to									
the lines there and the low load	Understand	ling Machine	Learning:	From Theory	to	ė.	Buy new:	\$68.00	c
UNIDEDETANDING	Algorithms	1st Edition					List F	rice: \$66.89 Det	1
UNDERSTANDING	by Shai Shaley-Shwarts	Chuthors, Shel Ben-David	(Author)				the bound family	month & Society	
MACHINE		and sharing			See all formats and	editors	Shipping to Fola	d Cetails	
LEARNING	#horDosh \$26.11 - \$44.65	Randoover \$33.15 - \$63.65	Faperback \$55.08 - \$55.15				Onder within 191	n 40 mins	
HEN DRON TO A OPTIMS	Red without hiss app	25 Chef Furn \$25.24					Or faster, deliver	Tasaday	
a colut in							O Calverts False		
2/1/3/202	Machine losening is one textbook is to introduce	r of the Easteral growing an e machine learning, and th	reas of scengular sil he algorithmic pasa	ience, with far-reaching ag dignes it offers, in a princip	plications. The sits of 6 and way. The book pro-	66 dos.on	In Stack		
1000	extensive theoretical ac transform these princip	court of the fundamental fealings practical alcority	d ideas underlying a ms. Following a pre	nuchine learning and the s sentation of the basics of	nathematical derivation the Fald, the book cover	that ta with	(0y 1 v)		
· · · ·	array of central tapics to	hat have not been address	med by previous tax	theoks. These include a di	scussion of the somputs	tional		ia Cart	
	v Feed more							These	í
1010	Caller M	1999-13	100m	nonhy.	NAME AND ADDRESS OF		Distant Analy		1
fed continuents recently	19			Bo	63		billy seas	Laam ENGLANDE AND PTL D	
Real secole	1109067154	106-1101007-005		Combridge	150 15, 2014		Payment Secure	runa Nor	
				menorally Press			Add a gift reo	light for easy	
Failure the authors									
Dal Paler Felor							Bay might		
Dal Raley Indox							Bay worth \$19.21		
Builder     Number     Number     Number     Number     Number     Number     Num							Rey such SIS 22		
Dat Valor     Note     Note     Date     Det							Add to U.A		
Builder     Traine     Builder     Traine     Builder     Traine  Frequently bought togethe							And to Unit	haars 1997 1	
Frequently bought togethe		MARKATE IN	8				Add to U.M.	haan 1997 I	
Frequently bought togethe			0	Testprice \$172.51			And the Los And the Los New & Sheet Difference areagons to	haan 199/1 cock chabs	
Frequently bought togethe		V		Tatal price \$172.51 Add all 3 to Care	_		New A lived DSV Area to a constraint of the second	haan 19971 OOK Chidas Chida	
Frequently bought togethe		Ŷ	•	Teral price \$172.51 Add all 3 to Care	-		New A lived (23)	haan, 1997 1 Ook chabs an fy annes Clains Clains	
But Uter     Tetre     But Uter     Tetre     But Uter     Tetre     Te	title the second		D	Tent price \$172.51 Add at 3 to Can	-		Add to Uar Add to Uar New & Uned U20 Add to Uar See See to a char	haan 199/1 Ook chidas anty annes Chita Chita	
Mal Value         Nature           March State         Nature           Proguently bought togethe         Proguently bought togethe           This state states and planters         Proguently bought togethe	Total Management	Notice for Madvest	E	Terra price \$172.51 Add at 3 to Care	-		Ney costs 199-22 Addres Lite New & Uned (20) agesgoon See Notice a dat	haan <sup>1</sup> 99 <sup>1</sup> Ook clubs sefy asses Clubs	
Note: Section 1 and 1 an	The second secon	Notice the Madride	E	Tenar price \$172.51 Add all 3 to Can	-		Ney costs 199-22 Addres Lite New & Uned (20) agesgoon See Notice a dat	haan 1997 1 OOK Clubs Stafy assess Clubs 21 Laare record	
Water         Index           Water         Index           Water         Index           Water         Index           Precuently bought togethe         Index           Water         Index         Index           The metastration in the statemetastration in the statemetastratestration in the statemetastration in the statemetas	With the second	Subsection for Machine 1		Tent price \$172.51 Ant at 3 to Call	-		Ney worth 199 27 Addres Lin Pere & Uned DN: arrangent See in a dur	haan 1997 I cook clubs as for anom Clubs I Learn more	
National States and St	which is full factorial and the full factoria	indicensations for Machine In Aggin	Carenteg (	Test ploc \$172.51 AM at 3 vi Cat	-	Page 1 of T	Ny ani 19522 Additio Lin Piere & Used D20 agragont Section a dur	hours 1991 Ook Chilos C	
The second secon	Extension Sector of Statistical Sector of S	Scherosoftes for Madvant 40 <sup>19</sup>	2 Lawing	Tessi price \$172.51 AM at 3 N Car	- -	Page 1 of T	Buy and 199.27 Add to Lie Reve & Uned DSC assigned Section a due	hann 1991 Ook Chilos Ch	
The second secon	write of Statistical Data Statistical EQUAL TO A Statistical Columnia St	Scharter for Machine 1 40 <sup>29</sup>	2 Learning	These price \$172.51 And all 3 to Can.	4	Page10FT	Buy and 199 27 Addres List Pierr & Uned List Reve & Uned List Sectors a char Not be a char	hun 1997 OC clubs ne (clubs Clubs Clubs Clubs Spece	
Balance     Balance	were of Statistical Costs Holing Information Costs Holing Information C	Northeast for Machine of Machineo	C	Test plot \$172.51 And at 3 to Call		Pape 1 of 7	Buy unit 199 27 Add to Lie Pere & Uned 199 Ber & Uned 199 Get in a cho	hun 1991 an cubs an cu	
Barbar Terrer Barbar Terrer Frequently bought together Frequently bought together	exclusion factors in the second secon	West of the West o	2 Learning	Perception & \$772.51 And its sector And its sector	A Distant	Page107	Ney and 199.23 Add to LM New & Une (U) acceptor Sol to a du	hean 1997 1 cold childs may annot Chile 1 Lann more Speed	
Berner i rere     Berner i rere i rere     Berner i rere i rere     Bernere i rere     Bernere i reree i rere     Berner i rere     Berner	estellar est	The second	transfer	Tend paice \$172.51 Add 314 Carl	a Distant	Page 1 of 7	New week Address List New & Uned List New & Uned List Sectors a char Notice	han 1997 Old Calos Calo Class Class Class	
Barrier Terrent (Construction)     Barrier Terrent (Construction)     Barrier Terrent (Construction)     Barrier Terrent (Construction)     Barrier (Co	estation of the second se	Hard States		Tend pick \$172.51 And all success and all success the all succ	A Deserve and A	Papilat	Buy and 100 J 2 Add to Lint Free & Unid J 22 Boy in a dia Statistical and Add Statistical and Add Statistic	hann 1997 f con chuide s an tha ann an Chile of Lanne more Spree	
Beneric Construction     Constructi	eneral data data data data data data data da	The second secon	e Landy	Teorgeous \$772.51 And 19 to on	A Constant of the Second	Page 1 of 7	Buy and 100 - 24 Add to List Reve & Uned D20 Sector & data Sector & data	hyan 1997 I Book Chidas Chida	
Barlier Terrent (Construction)     Barlier Ter	event stated thorse were thank of the thorse were thank of the thorse were thorse were the thorse were the thore were the thorse were the thorse were the thorse were the thor		B Landy File Color File Color Fil	Ped pick \$172.51 And 3 to core	A construction of the second s	Rape 1 of T	Buy and 109.23 Addres List New & United (10) Berlin and (10) See Section and (10) Section and (1	hunn 199/1 COAC Claids and y annou Clain O Learn more Speed	
Beneric I of the second s	ested and the second s	And and a second s	2 samba free 65 states to 100 states to 100	Produce \$722.51 Area to compare to the second secon	A construction	Pape 1 of 7	By and 1997 24 Addra Lba Piere & Lbeid US essagent Soch in a cho	hour 199/1 COA Chairs and provide Calor Calor Speed	

#### Multi-label classification:

$$oldsymbol{x} \in \mathcal{X} 
ightarrow oldsymbol{y} \in \mathcal{Y} := \{0,1\}^m$$

- a large number of labels  $m \ (\geq 10^5)$ ,
- a label vector  $\boldsymbol{y}$  is very sparse,  $\|\boldsymbol{y}\|_1 \ll m$ ,
- many problems are naturally budgeted at k (requirement for a prediction ŷ: ||ŷ||<sub>1</sub> = k),



#### Multi-label classification:

$$\boldsymbol{x} \in \mathcal{X} 
ightarrow \boldsymbol{y} \in \mathcal{Y} := \{0,1\}^m$$

- a large number of labels  $m \ (\geq 10^5)$ ,
- a label vector  $\boldsymbol{y}$  is very sparse,  $\|\boldsymbol{y}\|_1 \ll m$ ,
- many problems are naturally budgeted at k (requirement for a prediction ŷ: ||ŷ||<sub>1</sub> = k),



#### Multi-label classification:

$$\boldsymbol{x} \in \mathcal{X} \rightarrow \boldsymbol{y} \in \mathcal{Y} := \{0, 1\}^m$$

- a large number of labels  $m \ (\geq 10^5)$ ,
- a label vector  $\boldsymbol{y}$  is very sparse,  $\|\boldsymbol{y}\|_1 \ll m$ ,
- many problems are naturally budgeted at k (requirement for a prediction ŷ: ||ŷ||<sub>1</sub> = k),
- long-tail distribution of labels.



Problem with long tail and common metrics budgeted at  $\boldsymbol{k}$ 

Standard instance-wise metrics, e.g.:

$$\mathsf{Precision}@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k} \sum_{j=1}^{m} y_{ij} \hat{y}_{ij}$$

#### Problem with long tail and common metrics budgeted at k

Standard instance-wise metrics, e.g.:

$$\mathsf{Precision}@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k} \sum_{j=1}^{m} y_{ij} \hat{y}_{ij}$$

Table: Performance measures (%) on AmazonCat-13k of a classifier trained on the full set of labels and a classifier trained with only 1k head (most frequent) labels.

Metric	+	full labels	5		head labels	
	@1	@3	@5	@1 (diff.)	@3 (diff.)	@5 (diff.)
Precision	93.03	78.51	63.74	93.08 (+0.05%)	76.42 (-2.66%)	58.21 (-8.67%)
nDCG	93.03	87.25	85.35	93.08 (+0.05%)	85.75 (-1.71%)	80.91 (-5.19%)
PS-Precision	49.76	62.63	70.35	49.07 (-1.39%)	57.71 (-7.84%)	57.41 (-18.40%)

#### Problem with long tail and common metrics budgeted at k

Standard instance-wise metrics, e.g.:

Metrics that linearly decompose over labels:

$$\mathsf{Precision}@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k} \sum_{j=1}^{m} y_{ij} \hat{y}_{ij}$$

$$\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{j=1}^{m} \psi^{j}(\boldsymbol{y}_{:j}, \hat{\boldsymbol{y}}_{:j})$$

Table: Performance measures (%) on AmazonCat-13k of a classifier trained on the full set of labels and a classifier trained with only 1k head (most frequent) labels.

Metric	+	full labels	5		head labels	
	@1	@3	@5	@1 (diff.)	@3 (diff.)	@5 (diff.)
Precision	93.03	78.51	63.74	93.08 (+0.05%)	76.42 (-2.66%)	58.21 (-8.67%)
nDCG	93.03	87.25	85.35	93.08 (+0.05%)	85.75 (-1.71%)	80.91 (-5.19%)
PS-Precision	49.76	62.63	70.35	49.07 (-1.39%)	57.71 (-7.84%)	57.41 (-18.40%)

#### Problem with long tail and common metrics budgeted at k

Standard instance-wise metrics, e.g.:

Metrics that linearly decompose over labels:

$$\mathsf{Precision}@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k} \sum_{j=1}^{m} y_{ij} \hat{y}_{ij}$$

$$\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{j=1}^{m} \psi^{j}(\boldsymbol{y}_{:j}, \hat{\boldsymbol{y}}_{:j})$$

Table: Performance measures (%) on AmazonCat-13k of a classifier trained on the full set of labels and a classifier trained with only 1k head (most frequent) labels.

Metric	full labels			head labels				
	@1	@3	@5	@1 (diff.)	@3 (diff.)	@5 (diff.)		
Precision	93.03	78.51	63.74	93.08 (+0.05%)	76.42 (-2.66%)	58.21 (-8.67%)		
nDCG	93.03	87.25	85.35	93.08 (+0.05%)	85.75 (-1.71%)	80.91 (-5.19%)		
<b>PS-Precision</b>	49.76	62.63	70.35	49.07 (-1.39%)	57.71 (-7.84%)	57.41 (-18.40%)		
Macro-Precision	13.28	32.65	44.16	4.31 (-67.54%)	5.28 (-83.82%)	4.32 (-90.21%)		
Macro-Recall	1.38	11.06	30.57	0.47 ( <del>-65.61%</del> )	2.69 (-75.71%)	4.10 (-86.59%)		
Macro-F1	2.26	14.67	32.84	0.74 (-67.37%)	3.10 (-78.88%)	3.77 (-88.51%)		
Coverage	15.19	40.53	60.88	5.11 (-66.32%)	7.37 (-81.82%)	7.52 (-87.65%)		

## Our contributions

• We analyze the problem of **optimization** of general family of metrics linearly decomposeable over labels calculated at k under **expected test utility framework (ETU)** 

$$\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{j=1}^{m} \psi^{j}(\boldsymbol{y}_{:j}, \hat{\boldsymbol{y}}_{:j}), \quad \hat{\boldsymbol{Y}}^{\star} = \operatorname*{argmax}_{\hat{\boldsymbol{Y}} \in \mathcal{Y}_{k}^{n}} \mathbb{E}_{\boldsymbol{Y} \mid \boldsymbol{X}}[\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}})].$$

## Our contributions

• We analyze the problem of optimization of general family of metrics linearly decomposeable over labels calculated at k under expected test utility framework (ETU)

$$\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{j=1}^{m} \psi^{j}(\boldsymbol{y}_{:j}, \hat{\boldsymbol{y}}_{:j}), \quad \hat{\boldsymbol{Y}}^{\star} = \operatorname*{argmax}_{\hat{\boldsymbol{Y}} \in \mathcal{Y}_{k}^{n}} \mathbb{E}_{\boldsymbol{Y} \mid \boldsymbol{X}}[\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}})].$$

• Our framework only requires the probability estimates of individual labels for each instance  $\eta(x) = (\eta_1(x), \dots, \eta_m(x)) \coloneqq \mathbb{E}_{y|x}[y] \to \text{easy to apply on-top of existing classifiers}$ 

## Our contributions

• We analyze the problem of optimization of general family of metrics linearly decomposeable over labels calculated at k under expected test utility framework (ETU)

$$\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{j=1}^{m} \psi^{j}(\boldsymbol{y}_{:j}, \hat{\boldsymbol{y}}_{:j}), \quad \hat{\boldsymbol{Y}}^{\star} = \operatorname*{argmax}_{\hat{\boldsymbol{Y}} \in \mathcal{Y}_{k}^{n}} \mathbb{E}_{\boldsymbol{Y} \mid \boldsymbol{X}}[\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}})].$$

- Our framework only requires the probability estimates of individual labels for each instance  $\eta(x) = (\eta_1(x), \dots, \eta_m(x)) \coloneqq \mathbb{E}_{y|x}[y] \to \text{easy to apply on-top of existing classifiers}$
- We provide:
  - optimal prediction rules,
  - efficient approximations with guarantees,
  - ▶ regret bounds quantifying influence of label probability estimation error,
  - ▶ general algorithm, based on block coordinate ascent, that scales to XMLC problems.

# Thank you for your attention

Poster: Thursday, December 14, Poster Session 5, #1025 Paper: https://arxiv.org/pdf/2311.05081.pdf