Multi-label Learning

Trees, Embeddings, and much more!

SIGML

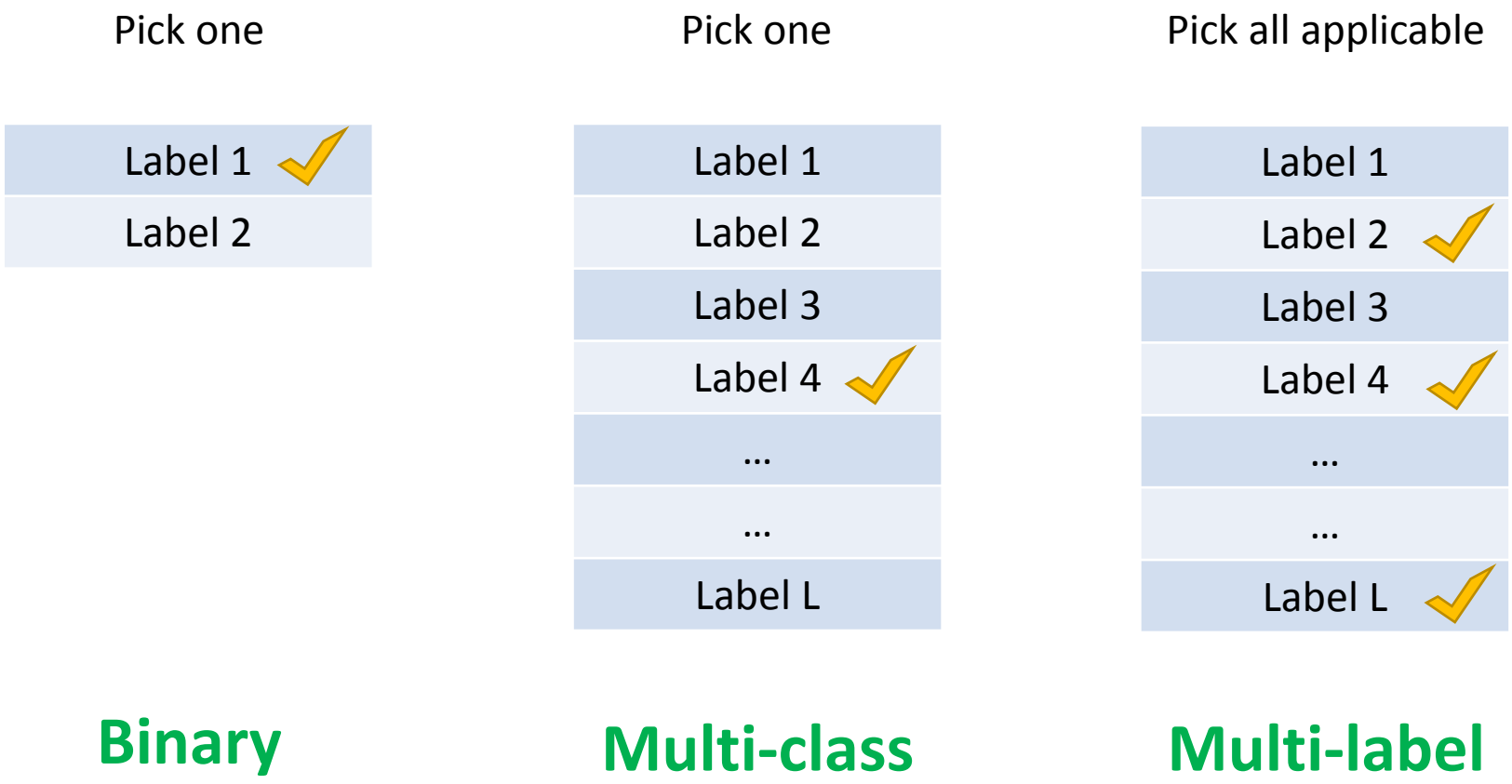
Special Interest Group in Machine Learning

Purushottam Kar

Department of CSE

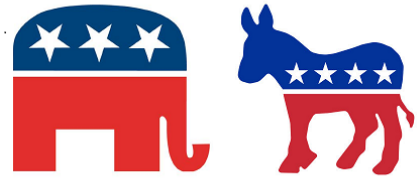
IIT Kanpur

Classification Paradigms



Classification Paradigms

Pick one

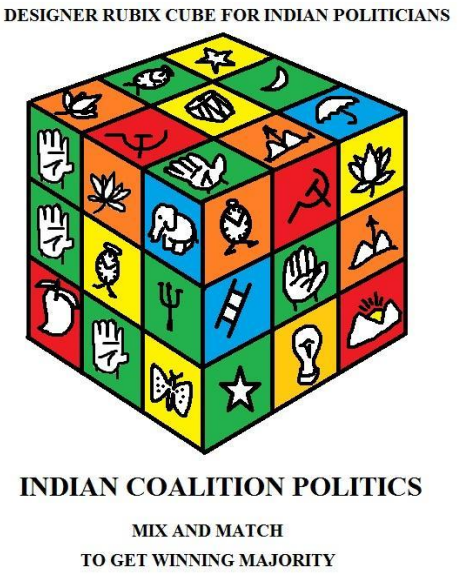


Binary

Pick one

Multi-class

Pick all applicable



Multi-label

Examples

eXtreme Multi-label Classification

What all items would this user buy?

$f\left(\right.$  $\left.\right)$

$f : \mathcal{X} \longrightarrow 2^{\mathcal{Y}}$

$\mathcal{X}$: Users

$\mathcal{Y}$: Items

eXtreme Multi-label Classification

Who all are present in this selfie?

eXtreme Multi-label Classification

Create account Log in

Article Talk | Read Edit View history | Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here

Bharata Natyam

From Wikipedia, the free encyclopedia

Bharathanatyam (Tamil: பரதநாட்டியம்) is a form of Indian classical dance that originated in the temples of Tamil Nadu.[1][2][3][4][5] It was described in the treatise *Natya Shastra* by Bharata around the beginning of the common era. Bharata Natyam is known for its grace, purity, tenderness, expression and sculpturesque poses. Lord Shiva is considered the God of this dance form. Today, it is one of the most popular and widely performed dance styles and is practiced by male and female dancers all over the world, although it is more commonly danced by women.[6]

Contents [hide]

1 Etymology
2 Dance tradition
3 Essential ideas
3.1 Spiritual symbolism
4 Medieval decline
5 Modern rebirth

Bharathanatyam

Dances by name, Indian culture, Performing arts in India, South India, Tamil culture

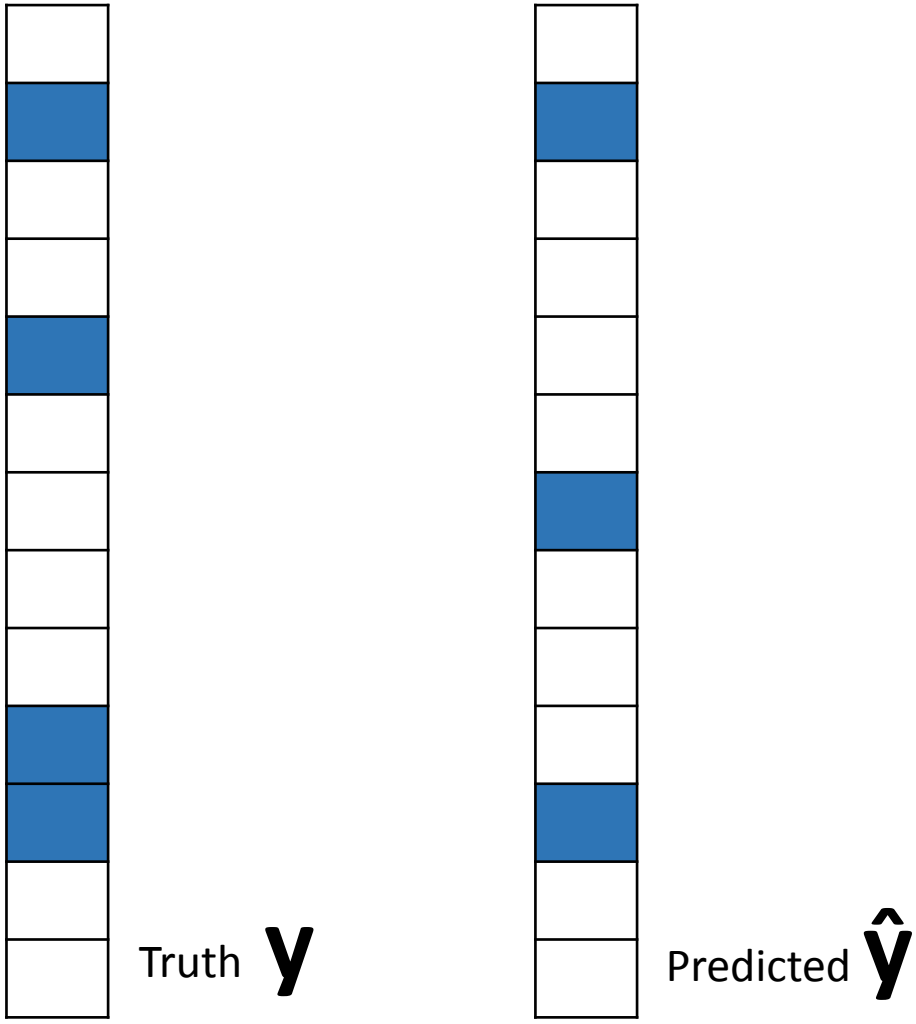
Challenges and Opportunities in Multi-label learning

- Exploit label correlations
  - Problem not as large as it seems
- Missing labels in training and test set
  - Appropriate training and evaluation?
- Novelty and Diversity in predicted set of labels?
  - Useful in recommendation and tagging tasks

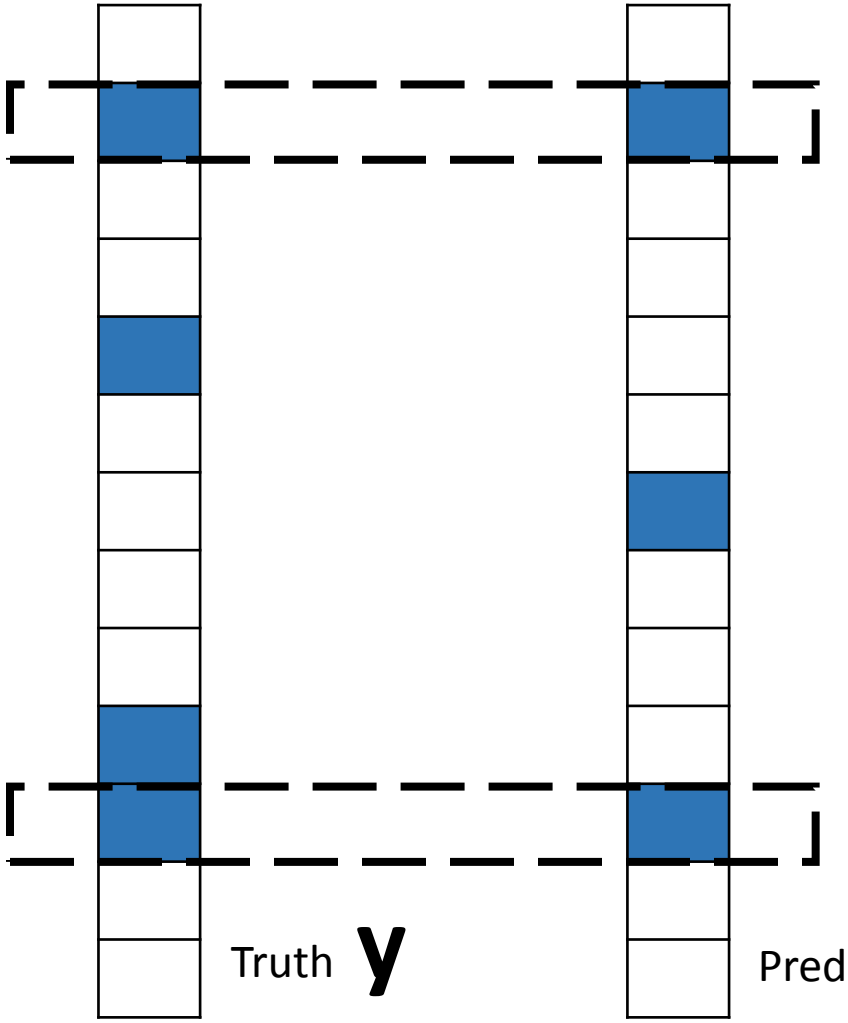
Evaluation Techniques

An Invitation to Optimization Connoisseurs

Classification Metrics

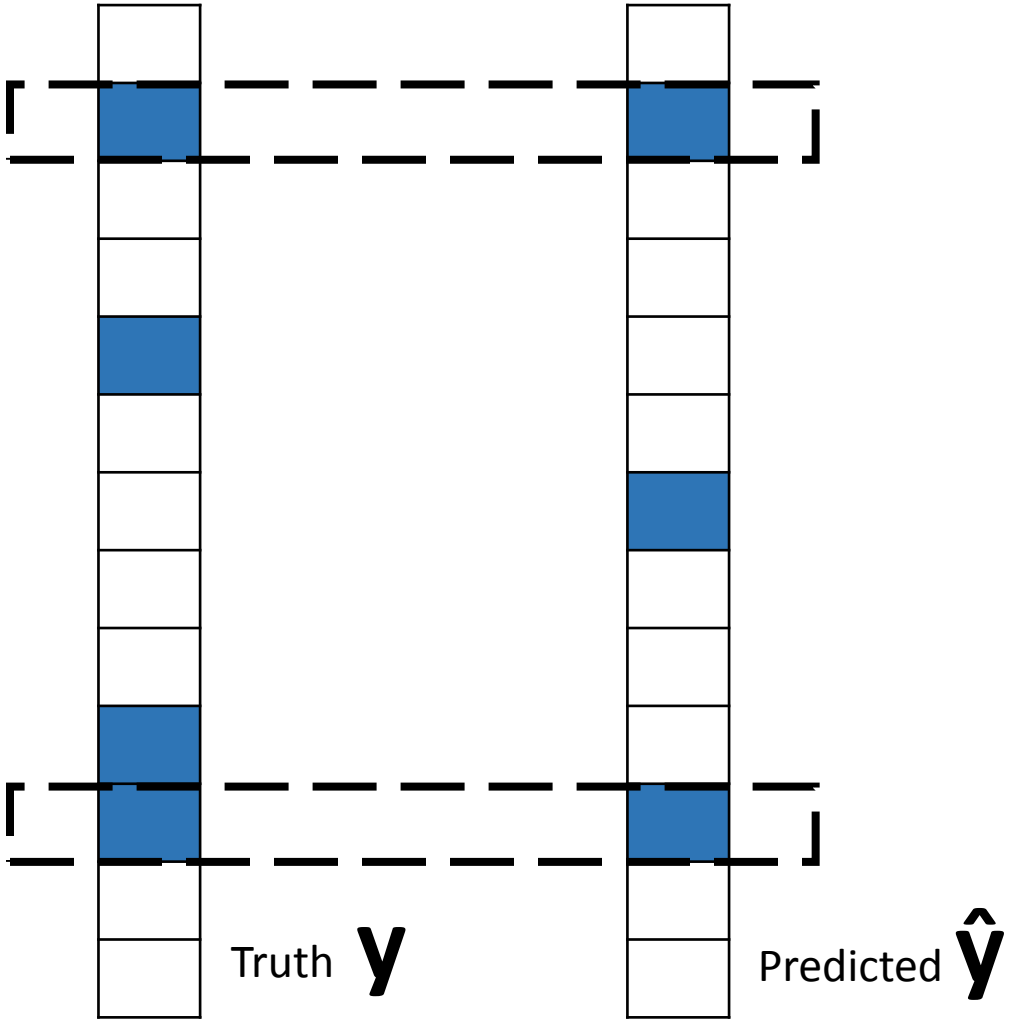


Hamming Loss



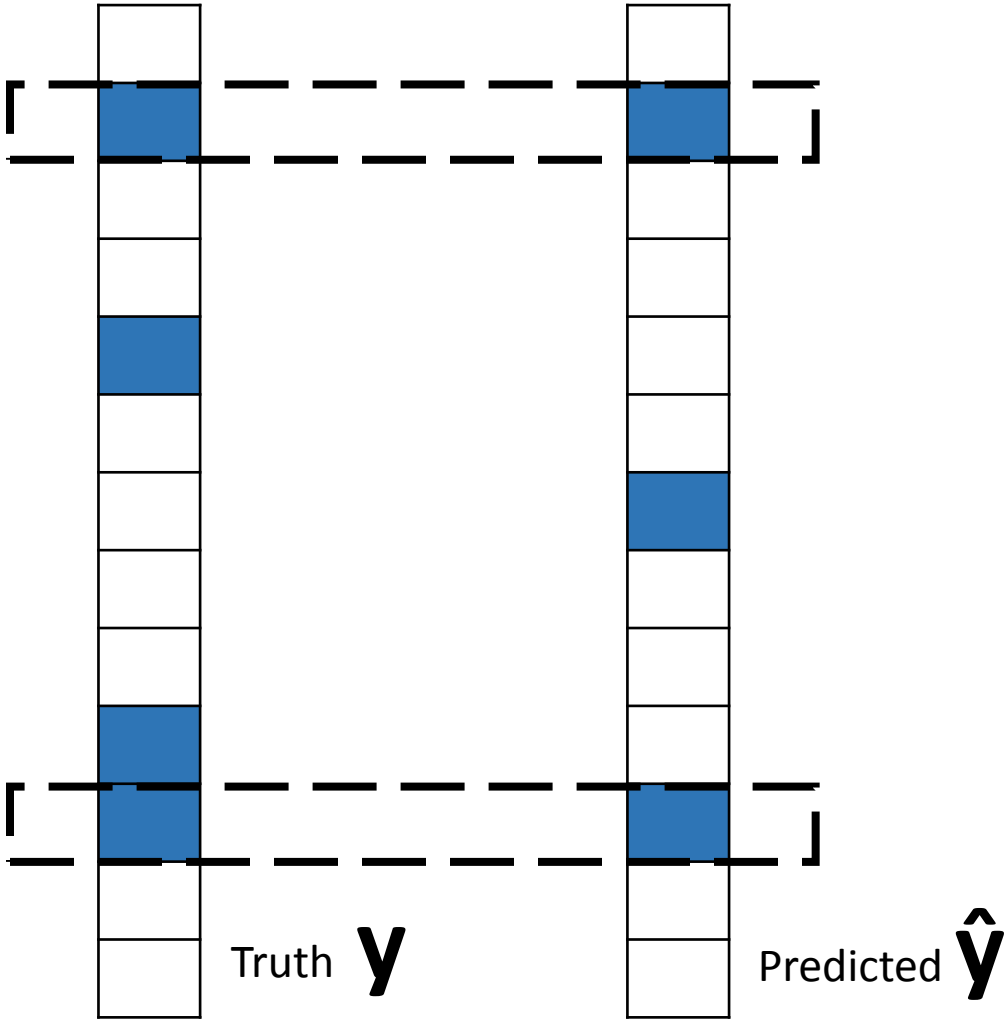
- $(|\mathbf{y}|+|\hat{\mathbf{y}}|-2|\mathbf{y}\cap\hat{\mathbf{y}}|)/L$
  $=(|\mathbf{y}\,\Delta\,\hat{\mathbf{y}}|/L)$
  $= 3/13 = 0.23$
- *Symmetric difference*
- *What if* $|\mathbf{y}| >> |\hat{\mathbf{y}}|$?

Precision



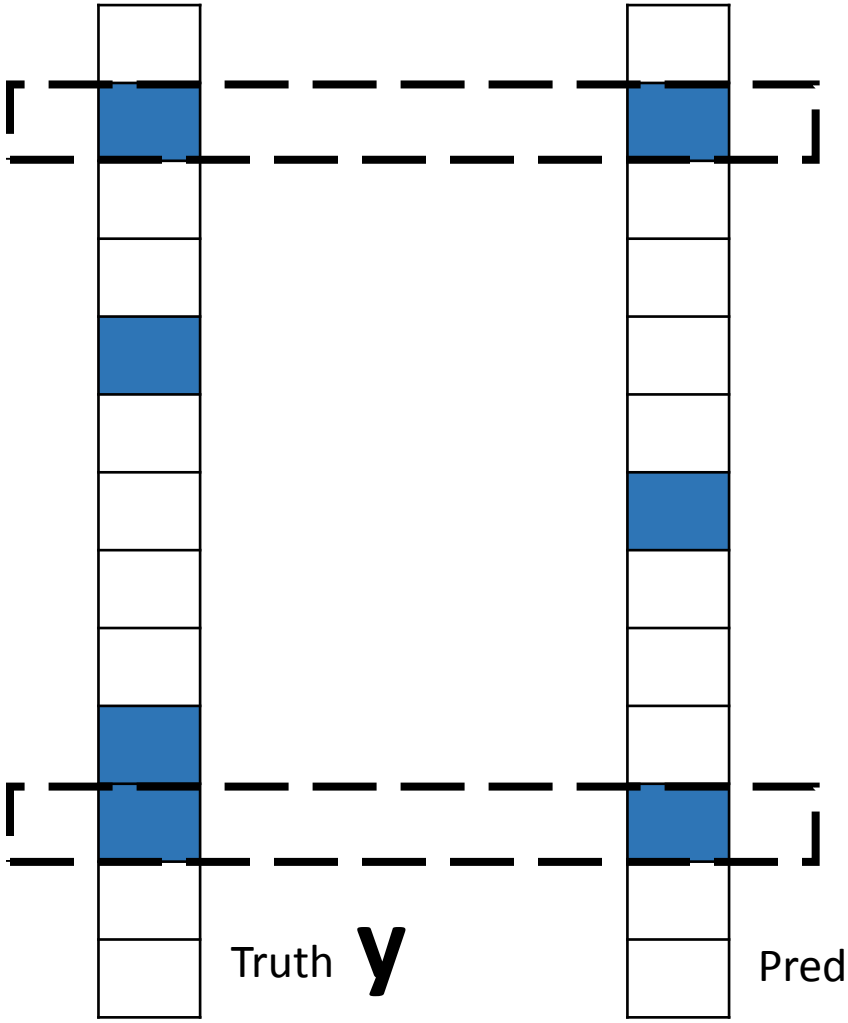
- $|\mathbf{y} \cap \hat{\mathbf{y}}| / |\hat{\mathbf{y}}|$ = 2/3 = 0.66

Recall



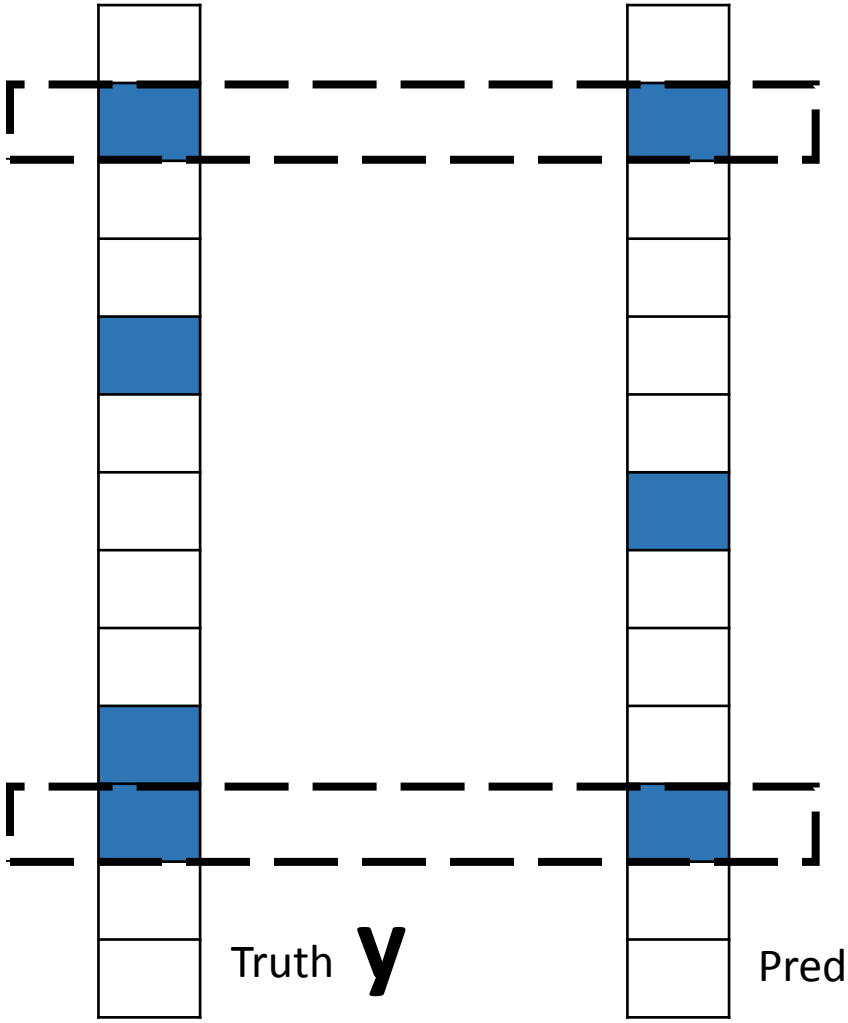
- $|\mathbf{y} \cap \hat{\mathbf{y}}| / |\mathbf{y}|$ = 2/4 = 0.5
- *What if* $|\mathbf{y}| >> |\hat{\mathbf{y}}|$?

F-measure



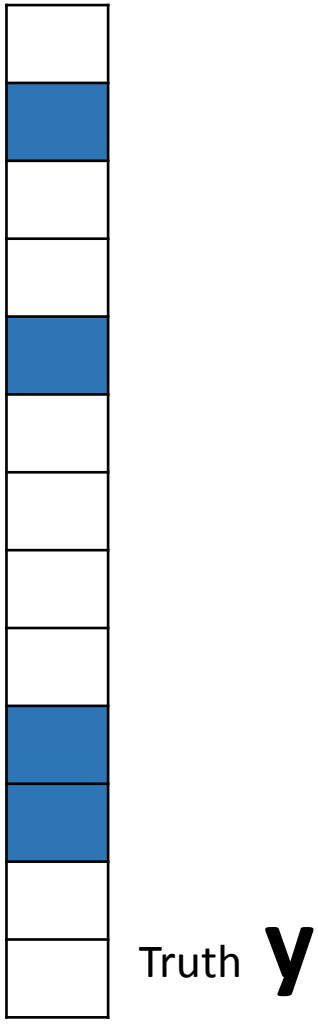
- Harmonic mean of precision and recall
- $2|\mathbf{y} \cap \hat{\mathbf{y}}| / (|\mathbf{y}| + |\hat{\mathbf{y}}|) = 0.57$
- What if $|\mathbf{y}| >> |\hat{\mathbf{y}}|$?

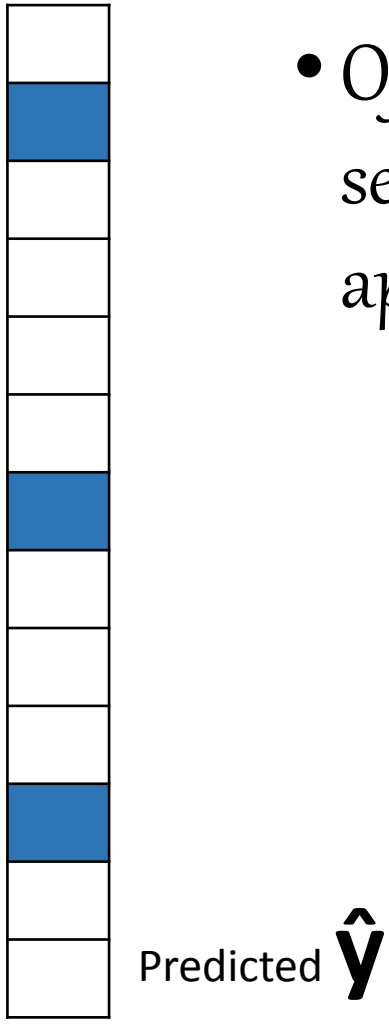
Jaccard Distance



- $|\mathbf{y} \cap \hat{\mathbf{y}}| / |\mathbf{y} \cup \hat{\mathbf{y}}|$ = 2/5 = 0.4
- *What if* $|\mathbf{y}| >> |\hat{\mathbf{y}}|$?

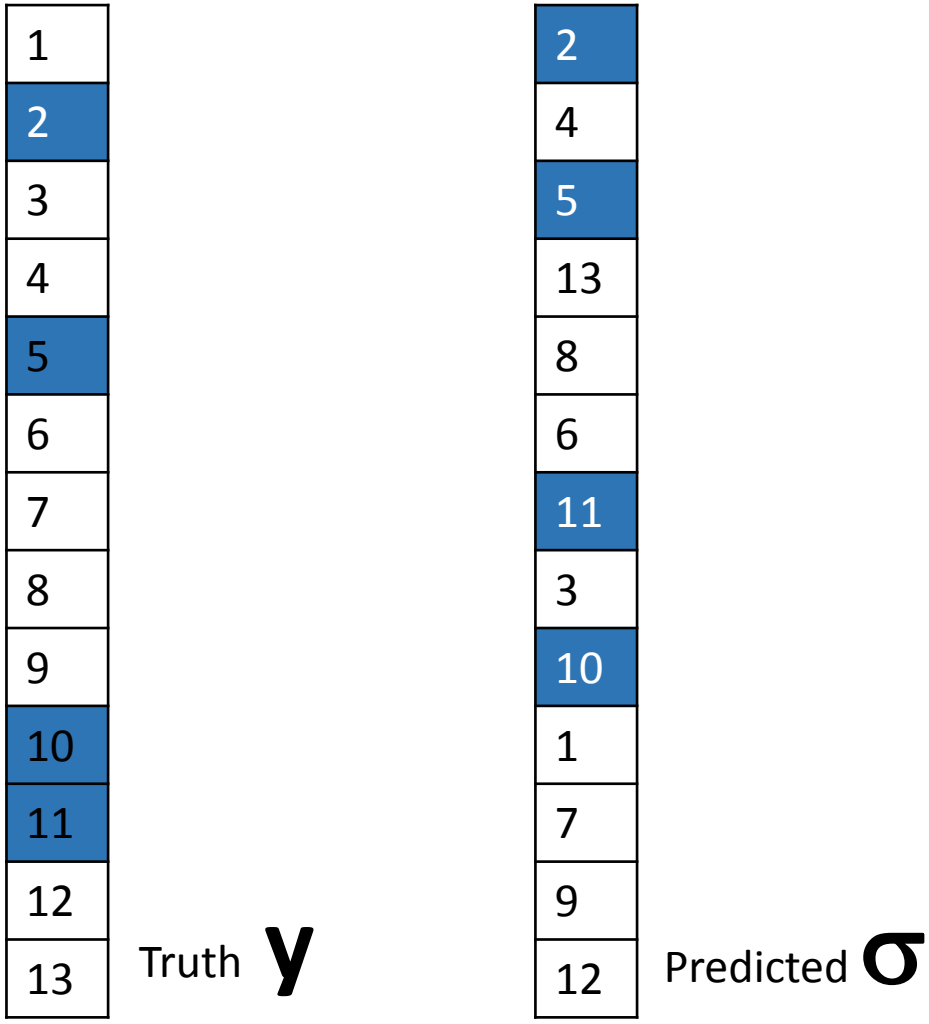
Classification Metrics



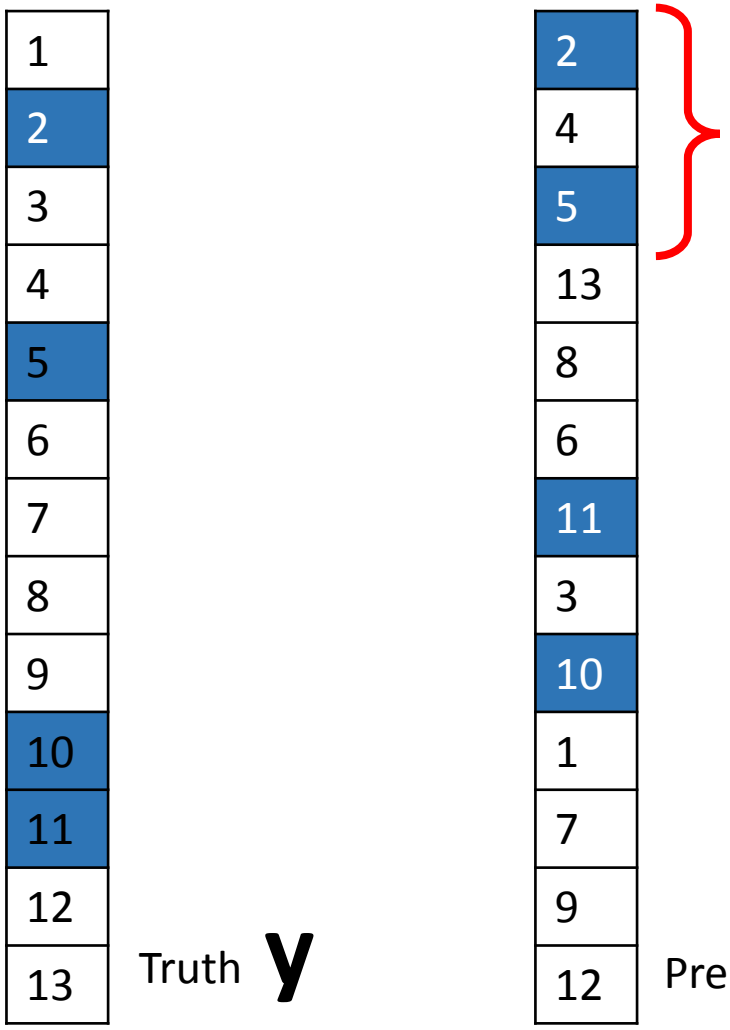


- Of these, only precision seems to be (mildly) appropriate for cases with
  - eXtremely large number of labels
  - Smaller prediction budgets
  - Missing labels in truth

Ranking Metrics



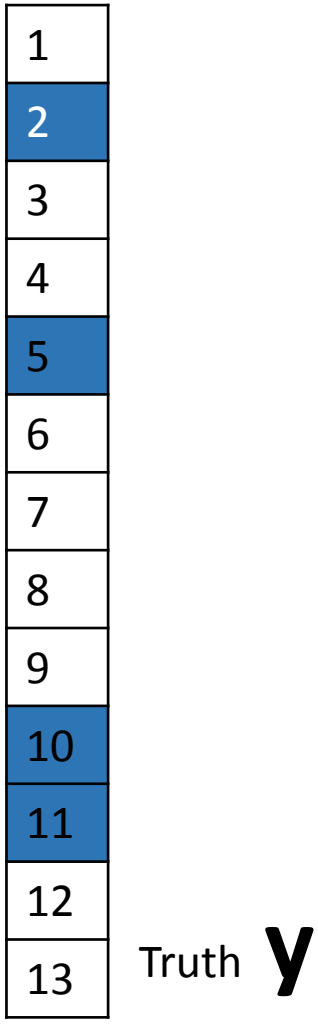
Precision@k



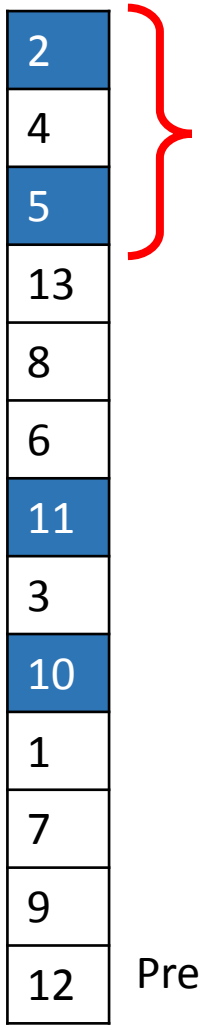
- Precision@1 = 100%
- Precision@2 = 50%
- Precision@3 = 66%

- Very appropriate for budget constrained prediction settings

Mean Average Precision



Truth **y**

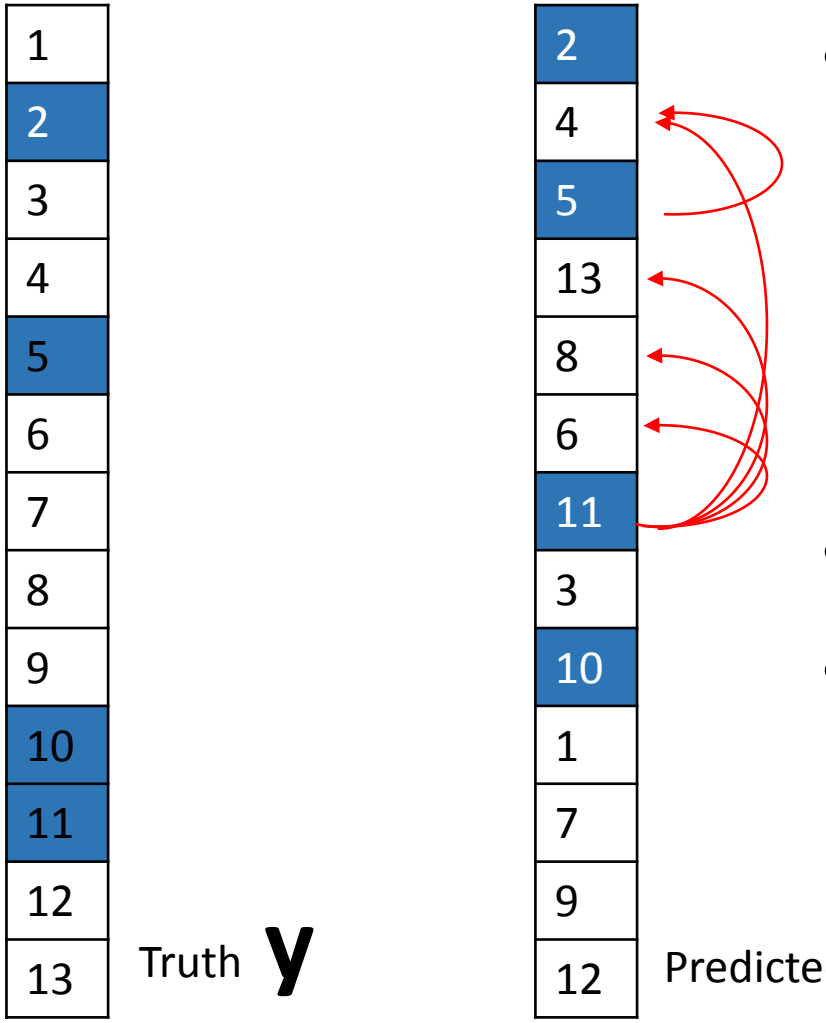


Predicted **σ**

- Precision@1 = 100%
- Precision@2 = 50%
- …
- Precision@13 = 13.7%

- MAP = 46.56%
- Usefulness for large L??

Area under the ROC curve



- Count mis-orderings
  - For 2: none
  - For 5: 1
  - For 11: 4
  - For 10: 5
- Total violations: 10
- AUC = 1 – 10/(4*9)

  = 0.72

Mean Reciprocal Rank

Truth y
1
2
3
4
5
6
7
8
9
10
11
12
13

Predicted σ
2
4
5
13
8
6
11
3
10
1
7
9
12

- Penalize rankings that rank "on" labels low
- Rank of 2 = 1
- Rank of 5 = 3
- Rank of 11 = 7
- Rank of 10 = 9
- MRR = ¼* $(1/1+1/3+1/7+1/9)$

  = 0.39 = 1/(2.52)

Solution Strategies

a.k.a. how to compress a decade worth of literature into an hour long talk

Notation and Formulation

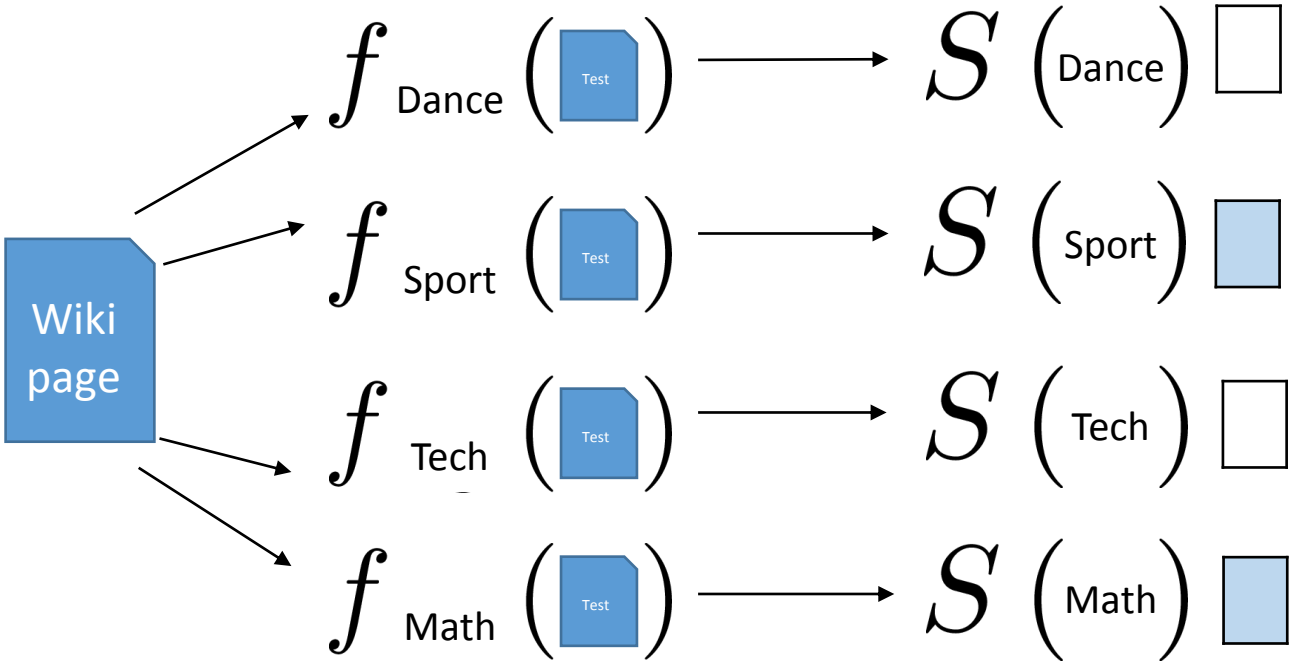
- Abstract problem: We have "documents" that are to be assigned a *subset* of L labels
- Representation
  - Documents: vectors in D dimensions
  - Labels: vectors in L dimensions (Boolean hypercube)
- Training set
  - $(\mathbf{x}_1,\mathbf{y}_1)$, $(\mathbf{x}_2,\mathbf{y}_2)$, $(\mathbf{x}_3,\mathbf{y}_3)$, …, $(\mathbf{x}_n,\mathbf{y}_n)$
  - $\mathbf{x}_i \in R^D$, $\mathbf{y}_i \in \{0,1\}^L$

The Three Pillars of Multi-label Learning

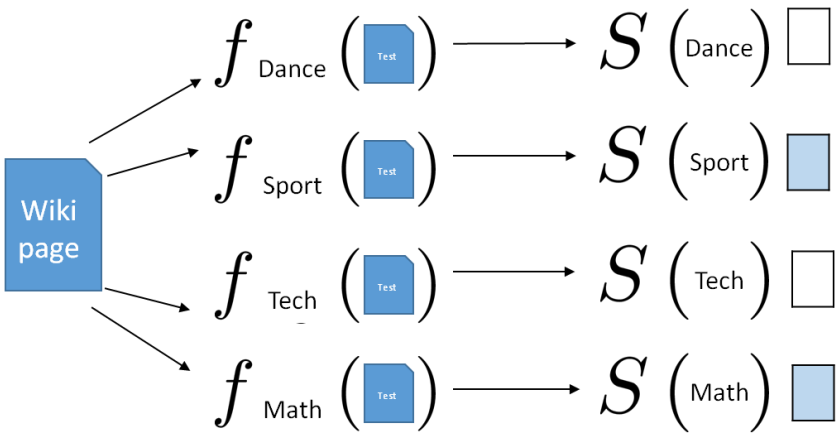
- 1-vs-All or Binary Relevance Methods
- Embedding or Dimensionality Reduction Methods
- Tree or Ensemble Methods

1-vs-All Methods

- Predict scores for each label separately
  - Threshold or rank scores to make predictions



1-vs-All Methods



Questions

- Are the L classifiers trained separately/jointly?
- If jointly then what "joins" the classifiers?

Benefits

- Extremely flexible model
- In-depth theoretical analysis possible

Considerations

- Training time ☹
- Test time ☹
- Model size ☹

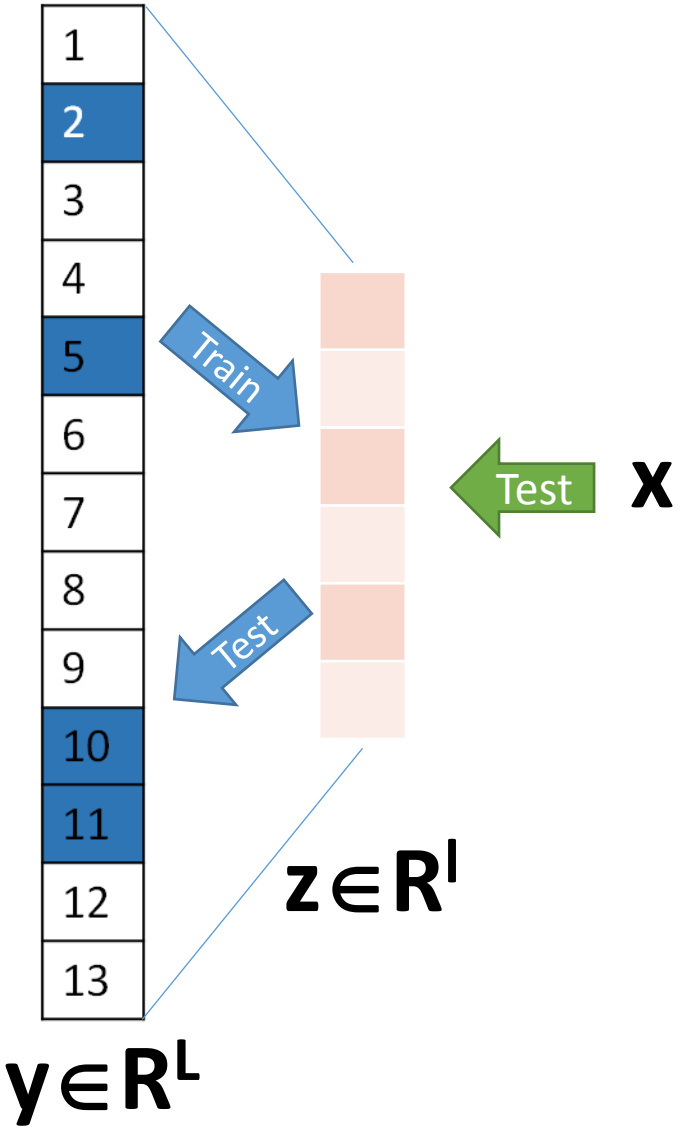
1-vs-All Methods

- Binary Relevance methods
  - Treat each label as a separate classification problem
  - Formulation (on board)
  - Also includes so-called *plug-in* methods, *submodular* methods
- Margin methods
  - Ensure scores of "on" labels are larger ~~than~~ much larger those of "off" labels
  - Formulation (on board)
- *Structured Loss* minimization methods
  - Formulation (sketch on board)

Embedding Methods

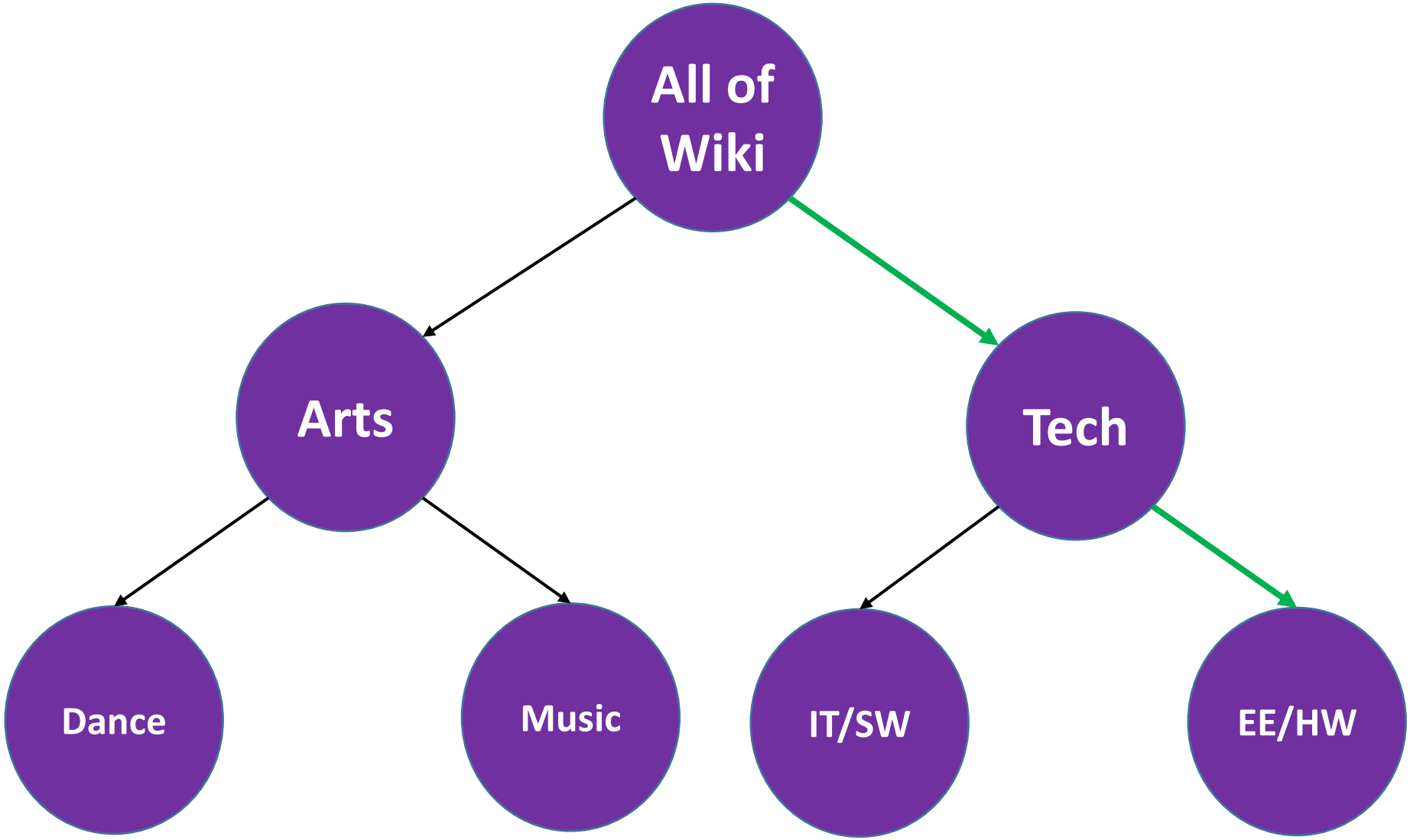
- Since L >>>1 and also has redundancies, reduce L
- Dimensionality reduction!!
- Nice theory, results, but expensive in prediction, training
- Questions
  - How to embed labels (linear/non-linear)
  - How to predict in the embedding space
  - How to "pull back" to the label space
  - Single/multiple embeddings
- CS, BCS, PLST, CPLST, LEML, SLEEC

Embedding Methods



- How to embed labels
  - RP(CS), CCA, PCA, Low local distortion proj., Learnt projections
- How to pull back
  - Sparse recovery, Nearest neighbor, Learnt projections
- Considerations
  - Training time ☺
  - Test time ☹
  - Model size ☺

Tree Methods



Tree Methods

- Partition the space of documents into several bins
  - To ease life, perform hierarchical partitioning as a tree
- At each leaf perform some classification task to predict
  - To increase efficiency, use several trees (forest)
- Questions
  - Partitioning criterion (clustering, ranking, classification)
  - Leaf action (constant labeling, use of another multi-labeler)
  - Ensemble size and aggregation method (single, multiple)
  - LPSR, MLRF, FAST-XML
- Consideration: good accuracy, fast prediction, huge models

The Three Pillars of Multi-label Learning

Name	"Accuracy"	Scalability	Prediction Cost	Model Size	Well Understood?
1-vs-All	Meh!	Yikes!	Are you kidding me!	Did I not make myself clear?	Now we are talking! Excellent
Embedding	Good/ Best	Good/ Best	Good	Good	Good
Tree	Good/ Best	Good/ Best	Best	Large	Meh!