Multi-label Classification

Prof. Dr. Dr. Lars Shmidt Thieme
Tutor: Leandro Marinho

Information System and Machine learning lab

Definitions

- Single-label classification – set of examples associated with a single label *l* from a set of disoint labels *L, |L|>1;*
- if *|L|=2,* then it is called a *binary* classification problem,
- While if *|L|>2,* then it is called a *multi-class* classification problem
- In multi-label classification, the examples are associated with a set of labels $Y \subseteq L$
- For a input **x** the corresponding output is a vector $Y = [y_1, ..., y_L]^T$

Multi-label Classification

- Multi-class classification : direct approaches
  - Nearest Neighbor
  - Generative approach & Naïve Bayes
  - Linear classification:
    - geometry
    - Perceptron
    - K-class (polychotomous) logistic regression
    - K-class SVM
- Multi-class classification through binary classification
  - One-vs-All
  - All-vs-all
  - Others
  - Calibration

Related Tasks

- Ranking – Order the top most related labels with the new instance
- Hierarchical classification – When the labels in a data set belong to a hierarchical structure
- Hierarchical multi-label classification – When each example is labelled with more than one node of the hierarchical structure
- Multiple-label problems – only one class is the true class

Multi-Label Classification Methods

- problem transformation methods - methods that transform the multi-label classification problem either into one or more single-label classification or regression problems

- algorithm adaptation methods - methods that extend specific learning algorithms in order to handle multi-label data directly

Problem Transformation Methods

Example: Original Data

Ex.:	Sports	Religion	Science	Politics
1	X			X
2			X	X
3	X			
4		X	X	

Problem Transformation Methods

Transformed data set using PT1

Ex.:	Sports	Religion	Science	Politics
1	X			
2				X
3	X			
4		X		

PT1 - Randomly selects one of the multiple labels of each multi-label instance and discards the rest

Disadvantage: Lost of information

Problem Transformation Methods

Transformed data set using PT2

Ex.:	Sports	Religion	Science	Politics
3	X			

PT2 – Discards every multi-label instance from the multi-label data set

Disadvantage: Even more lost of information

Transformed data set using PT3

Ex.:	Sports	(Sports ^ Politics)	(Science ^ Politics)	(Science ^ Religion)
1		X		
2			X	
3	X			
4				X

PT3 – Considers each different set of labels that exist in the multi-label data set as a single label. It so learns one single-label classifier $H : X \rightarrow P(L)$ **where *P(L)* is the power set of *L***

Disadvantage: Can lead to lots of classes with small number of examples

Problem Transformation Methods

Ex.:	Sports	¬Sports
1	X	
2		X
3	X	
4		X

(a)

Ex.:	Politics	¬Politics
1	X	
2	X	
3		X
4		X

(b)

Ex.:	Religion	¬Religion
1		X
2		X
3		X
4	X	

(c)

Ex.:	Science	¬Science
1		X
2	X	
3		X
4	X	

(d)

PT4 – Learns |*L*| binary classifiers $H_l : X \rightarrow \{l, \neg l\}$ **one for each different label *l* in *L.* For the classification of a new instance x this method outputs as a set of labels the union of the labels that are output by the |*L*| classifiers:**

$$H_{PT4}(x) : \bigcup_{l \in L} \{l\} : H_l(x) = l$$

How much multi-label is a data set ?

- Not all datasets are equally multi-label. In some applications the number of labels of each example is small compared to |*L*|, while in others it is large. This could be a parameter that influences the performance of the different multi-label methods. Let *D* be a multi-label evaluation data set, consisting of |*D*| multi-label examples $(x_i, Y_i), i = 1..|D|, Y_i \subseteq L$

- Label Cardinality of D is the average number of the examples in D:

$$LC(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}|Y_i|$$

- Label denisity of D is the average number of the examples in D divided by |*L*|:

$$LD(D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i|}{|L|}$$

Evaluation Metrics

- Let *D* be a multi-label evaluation data set, consisting of |*D*| multi-label examples $(x_i, Y_i), i = 1..|D|, Y_i \subseteq L$
- Let *H* be a multi-label classifier and $Z_i = H(x_i)$ be the set of labels predicted by *H* for example x_i

$$HammingLoss(H, D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \Delta Z_i|}{|L|}$$

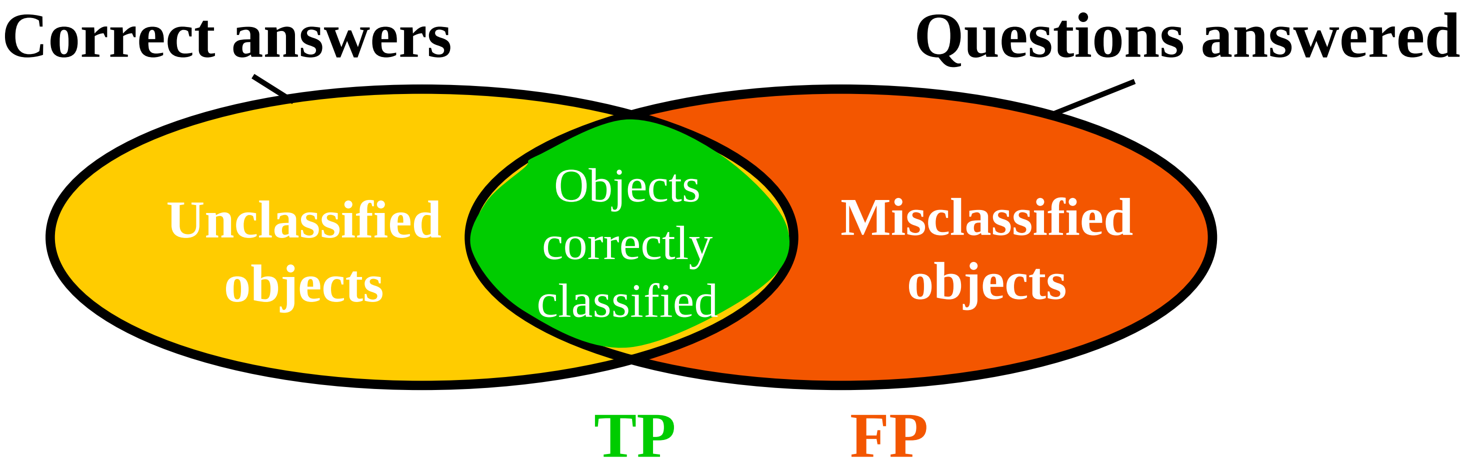
$$Recall(H, D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Y_i|}$$

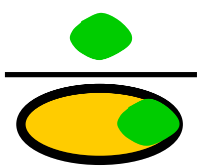
proportion of examples which were classified as label *x*, among all examples which truly have label *x*

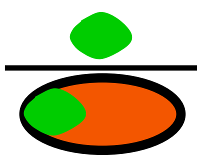
$$Precision(H, D) = \frac{1}{|D|}\sum_{i=1}^{|D|}\frac{|Y_i \cap Z_i|}{|Z_i|}$$

proportion of the examples which truly have class *x* among all those which were classified as class *x*

Precision-Recall



Recall=  = fraction of all objects correctly classified

Precision=  = fraction of all questions correctly answered

Data representation

- Labels as dummy variables
- Ex:
- X={att1,att2,att3}, L={class1, class2, class3, class4}

- 0.5,0.3,0.2,1,0,0,1
- 0.3,0.2, 0.5,0,0,1,1
- 0.5,0.1,0.2,1,1,1,0

Linear classification

- Each class has a parameter vector (w_k,b_k)
- x is assigned to class k iff $w_k^\top x + b_k \geq \max_j w_j^\top x + b_j$

- Note that we can break the symmetry and choose (w_1,b_1)=0
- For simplicity set b_k=0
(add a dimension and include it in w_k)
- So learning goal given separable data: choose w_k s.t.

$$\forall (x^i, y^i), \quad {w_{y^i}}^\top x^i \geq \max_j w_j^\top x^i$$

Three discriminative algorithms

Perceptron: $$\max_W \sum_i \left[{w_{y^i}}^\top x^i - \max_k w_k^\top x^i \right]$$

K-class logistic regression: $$\max_W \sum_i \left[{w_{y^i}}^\top x^i - \operatorname{softmax}_k w_k^\top x^i \right]$$

K-class SVM: $$\max_W \sum_i \left[{w_{y^i}}^\top x^i - \max_k (1\{k \neq y^i\} + w_k^\top x^i) \right]$$