Multi-Label Classification: Label Dependence, Loss Minimization, and Reduction Algorithms

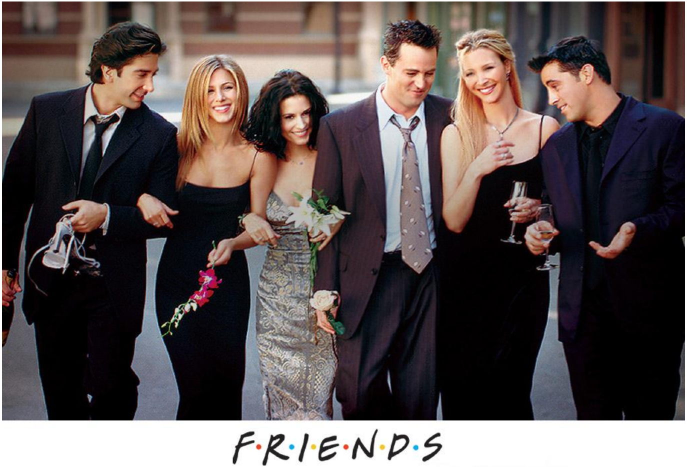
Krzysztof Dembczyński

Intelligent Decision Support Systems Laboratory (IDSS), Poznań University of Technology, Poland

April 2013, Poznań, Poland

Multi-label classification (MLC) is a prediction problem in which several class labels are assigned to single instances simultaneously.

Object detection on images



Characters:

- Ross, Rachel, Monica, Chandler, Phoebe, Joey

Multi-Label Classification

- Training data: $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$, $\boldsymbol{y}_i \in \{0,1\}^m$.
- **Predict** the vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$.

	X_1	X_2	Y_1	Y_2	$\ldots$	Y_m
$\boldsymbol{x}_1$	5.0	4.5	1	1		0
$\boldsymbol{x}_2$	2.0	2.5	0	1		0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\boldsymbol{x}_n$	3.0	3.5	0	1		1
$\boldsymbol{x}$	4.0	2.5	?	?		?

Multi-Label Classification

- Training data: $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$, $\boldsymbol{y}_i \in \{0,1\}^m$.
- **Predict** the vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$.

	X_1	X_2	Y_1	Y_2	$\ldots$	Y_m
$\boldsymbol{x}_1$	5.0	4.5	1	1		0
$\boldsymbol{x}_2$	2.0	2.5	0	1		0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\boldsymbol{x}_n$	3.0	3.5	0	1		1
$\boldsymbol{x}$	4.0	2.5	1	1		0

Straight-forward Approaches

- **Binary Relevance**: Decompose the problem to m binary classification problems.
- **Label Powerset**: Treat each label combination as a new meta-class and use any multi-class classification method.

	X_1	X_2	Y_1	Y_2	$\ldots$	Y_m
$\boldsymbol{x}_1$	5.0	4.5	1	1		0
$\boldsymbol{x}_2$	2.0	2.5	0	1		0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\boldsymbol{x}_n$	3.0	3.5	0	1		1
$\boldsymbol{x}$	4.0	2.5	?	?		?

Two Main Issues in Multi-Label Classification

- **Exploiting interdependence between labels** – the different class labels have to be predicted simultaneously.
- **A multitude of different loss functions** – different performance measures can be defined for multi-label predictions.

Theoretical Framework for Multi-Label Classification

Let us start with Conventional Classification

- Let Y be the random response variable and y its realization that take values from set $G = \{1, \ldots, K\}$.
- Similarly, let $\boldsymbol{X}$ be a random vector of features describing examples, and $\boldsymbol{x}$ be a realization of the random vector.
- The task is to find a function $h(\boldsymbol{x})$ that for a given object $\boldsymbol{x}$ predicts accurately the actual value of y.

Classification Problem

- We assume that data are coming from distribution

$$P(Y, \boldsymbol{X}) \,.$$

- Since we predict the value of Y for a given object $\boldsymbol{x}$, we are interested in conditional distribution:

$$P(Y = k | \boldsymbol{X} = \boldsymbol{x}) = \frac{P(Y = k, \boldsymbol{X} = \boldsymbol{x})}{P(\boldsymbol{X} = \boldsymbol{x})}$$

- It is reasonable to choose response k for which $P(Y = k | \boldsymbol{X} = \boldsymbol{x})$ is the largest.

Prediction

- This corresponds to minimization of the so-called 0/1 loss function:

$$\ell_{0/1}(y, f(\boldsymbol{x})) = \begin{cases} 0, & \text{if } y = h(\boldsymbol{x})\,, \\ 1, & \text{otherwise}\,. \end{cases}$$

- The solution of the following **risk** minimization problem:

$$\begin{aligned} y^* &= \arg\min_{h(\boldsymbol{x})} \mathbb{E}_{Y|\boldsymbol{x}} \ell_{0/1}(Y, h(\boldsymbol{x})) \\ &= \arg\min_{h(\boldsymbol{x})} \sum_{k \in G} P(Y = k | \boldsymbol{X} = \boldsymbol{x}) \ell_{0/1}(k, h(\boldsymbol{x})) \end{aligned}$$

is, in fact, k for which the conditional probability is the largest:

$$y^* = \arg\max_{k} P(Y = k | \boldsymbol{X} = \boldsymbol{x})$$

Getting Back to Multi-Label Classification

- The difference to binary classification is that instead of random variable Y we have a random vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_m)$.
- Vector $\boldsymbol{y} = (y_1, y_2, \ldots, y_m) \in \{0, 1\}^m$ is a realization of random vector $\boldsymbol{Y}$.
- So, data are coming from distribution

$$P(\boldsymbol{Y}, \boldsymbol{X}) \,.$$

- And the task is to find a function $\boldsymbol{h}(\boldsymbol{x})$ that for a given object $\boldsymbol{x}$ predicts accurately the actual value of $\boldsymbol{y}$.

Multi-Label Classification Problem

- Since we predict the value of $\boldsymbol{Y}$ for a given object $\boldsymbol{x}$, we are interested in conditional distribution:

$$P(\boldsymbol{Y}=\boldsymbol{y}|\boldsymbol{X}=\boldsymbol{x})=\frac{P(\boldsymbol{Y}=\boldsymbol{y},X=\boldsymbol{x})}{P(\boldsymbol{X}=\boldsymbol{x})}$$

Multi-Label Classification Problem

- Since we predict the value of $\boldsymbol{Y}$ for a given object $\boldsymbol{x}$, we are interested in conditional distribution:

$$P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}) = \frac{P(\boldsymbol{Y} = \boldsymbol{y}, X = \boldsymbol{x})}{P(\boldsymbol{X} = \boldsymbol{x})}$$

- It is reasonable to choose response $\boldsymbol{y}$ for which

Multi-Label Classification Problem

- Since we predict the value of $\boldsymbol{Y}$ for a given object $\boldsymbol{x}$, we are interested in conditional distribution:

$$P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}) = \frac{P(\boldsymbol{Y} = \boldsymbol{y}, X = \boldsymbol{x})}{P(\boldsymbol{X} = \boldsymbol{x})}$$

- It is reasonable to choose response $\boldsymbol{y}$ for which ... **?**

Multi-Label Classification Problem

- Since we predict the value of $\boldsymbol{Y}$ for a given object $\boldsymbol{x}$, we are interested in conditional distribution:

$$P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}) = \frac{P(\boldsymbol{Y} = \boldsymbol{y}, X = \boldsymbol{x})}{P(\boldsymbol{X} = \boldsymbol{x})}$$

- It is reasonable to choose response $\boldsymbol{y}$ for which ... **?**
  - $P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})$ is the largest?

Multi-Label Classification Problem

- Since we predict the value of $\boldsymbol{Y}$ for a given object $\boldsymbol{x}$, we are interested in conditional distribution:

$$P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}) = \frac{P(\boldsymbol{Y} = \boldsymbol{y}, X = \boldsymbol{x})}{P(\boldsymbol{X} = \boldsymbol{x})}$$

- It is reasonable to choose response $\boldsymbol{y}$ for which ... **?**
  - $P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})$ is the largest?
  - $P(Y_i = y_i | \boldsymbol{X} = \boldsymbol{x})$ are the largest?

Multi-Label Classification Problem

- Since we predict the value of $\boldsymbol{Y}$ for a given object $\boldsymbol{x}$, we are interested in conditional distribution:

$$P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x}) = \frac{P(\boldsymbol{Y} = \boldsymbol{y}, X = \boldsymbol{x})}{P(\boldsymbol{X} = \boldsymbol{x})}$$

- It is reasonable to choose response $\boldsymbol{y}$ for which ... **?**
  - $P(\boldsymbol{Y} = \boldsymbol{y} | \boldsymbol{X} = \boldsymbol{x})$ is the largest?
  - $P(Y_i = y_i | \boldsymbol{X} = \boldsymbol{x})$ are the largest?
  - ... ?
  - ... ?
  - ... ?

Multi-Label Loss Functions

- Subset 0/1 loss:

$$\ell_{0/1}(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) = \mathbb{1}[\boldsymbol{y} \neq \boldsymbol{h}(\boldsymbol{x})]$$

- Hamming loss:

$$\ell_H(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x})) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}[y_i \neq h_i(\boldsymbol{x})]$$

- F-measure-based loss:

$$\ell_F = 1 - F(\boldsymbol{y}, \boldsymbol{h}) = 1 - \frac{2 \sum_{i=1}^{m} y_i h_i}{\sum_{i=1}^{m} y_i + \sum_{i=1}^{m} h_i} \in [0, 1]$$

- Rank loss:

$$\ell_{\mathsf{rnk}}(\boldsymbol{y}, \boldsymbol{h}) = w(\boldsymbol{y}) \sum_{(i,j)\,:\, y_i > y_j} \left(\mathbb{1}[h_i(\boldsymbol{x}) < h_j(\boldsymbol{x})] + \frac{1}{2} \mathbb{1}[h_i(\boldsymbol{x}) = h_j(\boldsymbol{x})] \right)$$

Reduction Algorithms

- Reduction: reusing solutions to simple, core problems in order to solve more complex problems. (ICML 2009 Tutorial)
- Properties of reduction algorithms:
  - Assumptions behind a given reduction algorithm,
  - Statistical consistency and regret bounds,
  - Generalization bounds,
  - Learning and inference complexity.

Binary Relevance

- BR trains for each label independent classifier:
  - Does BR assume label independence?
  - Is it consistent for any loss function?
  - What is its complexity?

Binary Relevance

- The **risk minimizer**

$$\boldsymbol{h}^*(\boldsymbol{x}) = \arg\min_{\boldsymbol{h}} \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{x}} \ell(\boldsymbol{Y}, \boldsymbol{h}),$$

  for Hamming loss are the **marginal modes**:

$$h_i^*(\boldsymbol{x}) = \arg\max_{y_i \in \{0,1\}} P(Y_i = y_i \,|\, \boldsymbol{x}), \quad i = 1, \ldots, m$$

- It can be proved that BR is **consistent** for Hamming loss **without** any additional assumption on **label independence**.
- Learning and inference is linear in m (however, faster algorithms exist).

Label Powerset

- LP treats each label combination as a new meta-class and use any multi-class classification method
  - What are the assumptions behind LP?
  - Is it consistent for any loss function?
  - What is its complexity?

Label Powerset

- The risk minimizer for subset $0/1$ loss is the **joint mode**:

$$\boldsymbol{h}^*(\boldsymbol{x}) = \arg \max_{\boldsymbol{y} \in \{0,1\}^m} P(\boldsymbol{y} \,|\, \boldsymbol{x})$$

- Since LP treats the multi-label problem as a multi-class problem, it can be proved that LP is **consistent** for subset $0/1$ loss.
- Moreover, if used with probabilistic multi-class classifier, it estimates the joint conditional distribution for given $\boldsymbol{x}$.
- Unfortunately, learning and inference are basically exponential in m (however, this complexity is somehow constrained by the number of training examples).

Hamming Loss vs. Subset 0/1 Loss

- The risk minimizers of Hamming and subset $0/1$ loss have a different structure: marginal modes vs. joint mode.

$\boldsymbol{y}$	$P(\boldsymbol{y})$
0 0 0 0	0.30
0 1 1 1	0.17
1 0 1 1	0.18
1 1 0 1	0.17
1 1 1 0	0.18

Hamming loss minimizer:	1 1 1 1
subset $0/1$ loss minimizer:	0 0 0 0

- Under specific conditions, these two loss minimizers are provably equivalent: joint mode ≥ 0.5, conditional independence.
- However, minimization of the subset $0/1$ loss may result in a large error for the Hamming loss and vice versa.

Synthetic Data

Table: Results on two synthetic data sets.

Conditional independence

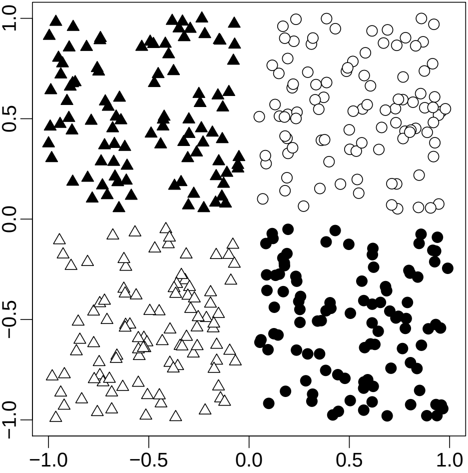
classifier	Hamming loss	subset 0/1 loss
BR	0.4208($\pm$.0014)	0.8088($\pm$.0020)
LP	0.4212($\pm$.0011)	0.8101($\pm$.0025)
Bayes Optimal	0.4162	0.8016

Conditional dependence

classifier	Hamming loss	subset 0/1 loss
BR	0.3900($\pm$.0015)	0.7374($\pm$.0021)
LP	0.4227($\pm$.0019)	0.6102($\pm$.0033)
Bayes Optimal	0.3897	0.6029

Synthetic Data

Figure: Data set composed of two labels: the first label is obtained by a linear model, while the second label represents the XOR problem.



Synthetic Data

Table: Results of three classifiers on this data set.

classifier	Hamming loss	subset 0/1 loss
BR Linear SVM	0.2399($\pm$.0097)	0.4751($\pm$.0196)
LP Linear SVM	0.0143($\pm$.0020)	0.0195($\pm$.0011)
Bayes Optimal	0	0

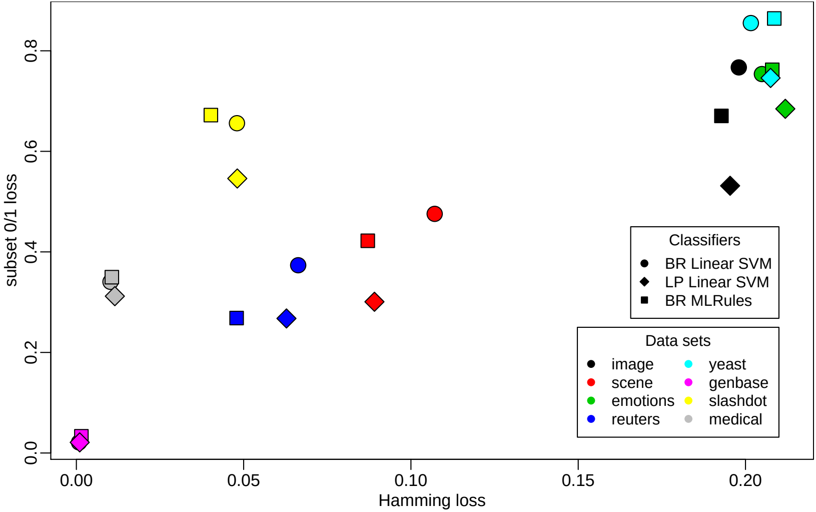
Synthetic Data

Table: Results of three classifiers on this data set.

classifier	Hamming loss	subset 0/1 loss
BR Linear SVM	0.2399($\pm$.0097)	0.4751($\pm$.0196)
LP Linear SVM	0.0143($\pm$.0020)	0.0195($\pm$.0011)
BR MLRules	**0.0011($\pm$.0002)**	**0.0020($\pm$.0003)**
Bayes Optimal	0	0

Benchmark Data

Figure: Results of three classifiers on 8 benchmark data sets.



Summary

- BR performs well for Hamming loss, but fails for subset $0/1$ loss.
- LP takes the label dependence into account, but the conditional one: it is well-tailored for the subset $0/1$ loss, but fails for the Hamming loss.
- LP may gain from the expansion of the feature or hypothesis space.
- One can easily tailor LP for solving the Hamming loss minimization problem, by marginalization of the joint probability distribution that is a by-product of this classifier.

Conclusions

- Modeling of label dependence,
- A multitude of loss functions,
- Reduction algorithms,
- Results presented for subset 0/1 loss and Hamming loss,
- Similar results for F-measure and rank loss.

This project is partially supported by the Foundation of Polish Science under the Homing Plus programme, co-financed by the European Regional Development Fund.

INNOVATIVE ECONOMY
NATIONAL COHESION STRATEGY

FNP
Foundation for Polish Science

EUROPEAN UNION
EUROPEAN REGIONAL
DEVELOPMENT FUND