Aalto University
School of Science
and Technology

Structured Output Learning with A Random Sample of Spanning Trees

Hongyu Su

Helsinki Institute for Information Technilogy HIIT
Department of Information and Computer Science
Aalto University

November 18, 2014

Multilabel Classification

- Multilabel classification is an important research field in machine learning.
  - For example, a document can be classified as "science", "genomics", and "drug discovery".
  - Each input variable $\mathbf{x} \in \boldsymbol{\mathcal{X}}$ is simultaneously associated with multiple output variables $\mathbf{y} \in \boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Y}} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_k$.
  - The goal is to find a mapping function that predicts the best values of an output given an input $f \in \mathcal{H} : \boldsymbol{\mathcal{X}} \rightarrow \boldsymbol{\mathcal{Y}}$.
- The central problems of multilabel classification:
  - The exponential sized output space $\boldsymbol{\mathcal{Y}}$ in the number of microlabels.
  - The dependency of microlabels to be exploited to improve the prediction performance.

Category of Algorithms

- Flat multilabel classification:
  - Multiple output variables are treated as a "flat" vector.
  - For example, ML-KNN, ADABOOST.MH, ...
- Structured output prediction:
  - There is an *output graph* connecting multiple labels.
    - A set of nodes corresponds to the multiple labels.
    - A set of edges represents the correlation between labels.
  - Hierarchical classification:
    - The output graph is a rooted tree or a graph with parent-child relationships defining the different levels of granularities.
    - For example, SSVM, ...
  - Graph labeling:
    - The output graph often takes a general form (e.g., a tree, a chain).
    - For example, M^3N, CRF, MMCRF, ...