PROBABILISTIC MODELS FOR STRUCTURED DATA

Course Project

Instructor: Yizhou Sun
yzsun@cs.ucla.edu

January 14, 2020
Overview

• Goal: design a probabilistic graphical model to solve real-world problems, and write a report that is potentially submitted to some venue for publication

• Teamwork
  • 3-4 people per group

• Milestones
  • Team formation due date: Week 2 (1pt as participation)
  • Proposal due date: Week 5 (5pt)
  • Presentation due date: 3/12/2020 in class (20pt)
  • Final report due date: 3/13/2020 (15pt)
    • What to submit: project report and code
Report Guideline

• Format: no more than 8-page, ACM SIG template: https://www.acm.org/publications/proceedings-proceedings-template-16dec2016:

  1. Title with group information (group # and name, group member names)
  2. Abstract
  3. Introduction of the overall goal and background
  4. Problem definition and formalization
  5. Methods description (detailed steps)
  6. Experiments design and Evaluation

• 7. Related work
  8. Conclusion

• 9. References
### Breakdown Points

Problem 1: Paper Classification in Directed Citation Network

- Cora Dataset:
  - Cora.zip

- Label: Each paper is associated with a research topic
  - There is a hierarchy structure in the dataset, please use the top hierarchy as labels

- Feature: Each paper has words extracted from title
• Task:
  • Design a probabilistic graphical model to leverage the citation links to classify papers into research topics

• Questions to address:
  • How to take the asymmetry in citation relation into the potential function design?
    • Design asymmetry potential function and implement it correctly
  • Will the consideration of asymmetry improve the classification accuracy?
    • Compare with the solution that simply ignores the asymmetry
• Evaluation:
  • Hide p% labels as test, use the remaining as training
    • Vary p to see its impact to the classification accuracy
  • Evaluation metric for multi-label classification
Problem 2: Node Classification in Heterogeneous Bibliographic Network

- Dataset
  - four_area.zip

- Label: authors and venues are associated with one of the four research areas, i.e., DB, DM, ML, IR
  - Label information can be found on DBLP_four_area.zip

- Feature: Only Papers are associated with text information
• Task:
  • Design a probabilistic graphical model to classify all the objects into four category in the network

• Questions to address:
  • How to leverage different types of links in the network?
    • Design different types of potential functions for different types of links by assuming different parameters
  • Will the consideration of type information for links improve the performance?
    • Compare the solution that treats all the links equally
• Evaluation:
  • Hide p% labels as test, use the remaining as training
    • Vary p to see its impact to the classification accuracy
  • Evaluation metric for multi-label classification
  • Evaluation when multiple types of nodes exist
Project 3: Polarity Detection for Twitter Users

• Dataset: Crawl Twitter Users following Political figures, their following, retweet, and reply behaviors, as well as their tweets

• Task: Design a probabilistic graphical model to classify all the users into two polarities
Project 4: Knowledge Completion for Knowledge Graphs via Higher-Order Dependency Modeling

- Datasets: Knowledge Graphs, such as YAGO, FreeBase, and NELL
- Task: Design a probabilistic graphical model to that can leverage higher-order dependency to solve knowledge graph completion tasks
  - i.e., <h,r,?>
Project 5: Construct CS Taxonomy from Wiki

- Task: construct taxonomy for terms related to computer science
  - E.g., root node: “computer science”

![Diagram of computer science taxonomy](https://www.researchgate.net/figure/Computer-Science-Taxonomy_fig1_260318181)
Project 6: NER for Wiki Pages in CS

- Task: Conduct NER task for text of wiki pages
  - Categories: concept (e.g., machine learning, deep learning); algorithm (e.g., CNN); application (e.g., self driving car); dataset (e.g., ImageNet), etc.