Heading and Administrative Matters

- Course ID: 599
- Course Title: Advanced Topics in Machine Learning: Statistical Relational Learning (SRL)
- Prerequisites: CSCI567 (Machine Learning) or permission by instructor
- Semester and day/time: Spring 2009, Tuesdays 5:00pm – 7:50pm
- Professors contact information
    - Name: Sofus A. Macskassy
    - Office: SAL 216
    - Office hours: by appointment
    - Email: macskass@usc.edu
    - Homepage: http://www.cs.rutgers.edu/~sofmac
    - TA: TBD

Introduction and Purposes

Statistical relational learning (SRL) is revolutionizing the field of automated learning and discovery by moving beyond the conventional analysis of entities in isolation to analyze networks of interconnected entities. In relational domains such as bioinformatics, citation analysis, epidemiology, fraud detection, intelligence analysis, and web analytics, there is often limited information about any one entity in isolation; instead it is the connections among entities that are of crucial importance to pattern discovery. Conventional machine learning techniques have two primary assumptions that limit their application in relational domains. First, algorithms for propositional data assume that data instances are recorded in homogeneous structures (i.e., a fixed number of attributes for each entity) but relational data instances are usually more varied and complex (e.g., molecules have different numbers of atoms and bonds). Second, the algorithms assume that data instances are independent but relational data often violate this assumption---dependencies may occur either as a result of direct relations or through chaining multiple relations together. For example, scientific papers have dependencies through both citations (direct) and authors (indirect).

This seminar will provide an introduction to recent research in statistical relational learning. The course will survey recent approaches that combine probabilistic and logical representations to model relational and network datasets, focusing on fundamental challenges in representation, learning, and inference. We will review conventional graphical models and inductive logic programming approaches as needed for background.

Classes will consist of instructor presentations, student presentations, and group discussions. Students will be required to (1) read, discuss, and present research papers, and (2) complete a semester-long class project. Potential projects include: investigating the performance of SRL algorithms, analyzing data with SRL models, design and implementation of SRL model/algorithm extensions.

Course Requirements and Grades

- Material
    - Required:
        - Course readings (to be provided)

- Optional:
  - Introduction to Statistical Relational Learning, L. Getoor and B. Taskar, editors, MIT Press, 2007.

- Grading breakdown
  - Assignments:
    - Response papers to weekly readings
      - Papers should be at least half a page and include, at minimum, a summary and two points of critique, question or praise of the work.
    - Paper presentations
      - Students are expected to present at least 3 papers throughout the semester (10-15 minutes per presentation; one paper at one lecture).
    - Leading class discussion
      - Students will be expected to start discussions in two lectures. This will include summarize paper responses (5-10) and have two or more questions prepared to get discussion started.
    - Class participation
      - Students are expected to attend lectures and participate in discussion.
    - Class project
      - A class project involving SRL in some manner.

  - Grade breakdown
    - Response papers to weekly readings: 20%
      Response papers will be graded on a scale of 1 to 5. Lowest two grades will be dropped.
    - Paper presentations: 20%
    - Leading class discussion: 20%
    - Class participation: 10%
    - Class project: 30%

Course Readings/Class Sessions

- Tentative schedule and readings. Final schedule will be presented first day of class.
  - 1/13: Introduction and example applications
    - Background: Getoor, L. and C. Diehl (2005). Link Mining: A Survey. SIGKDD Explorations, December, 2005, Volume 7 Issue 2.
    - McGovern, A., L. Friedland, M. Hay, B. Gallagher, A. Fast, J. Neville, and D. Jensen (2003). Exploiting Relational Structure to Understand Publication Patterns in High-Energy Physics. SIGKDD Explorations, December 2003, Volume 5, Issue 2, pages 165-172.
    - Neville, J., O. Simsek, D. Jensen, J. Komoroske, K. Palmer and H. Goldberg (2005). Using Relational Knowledge Discovery to Prevent Securities Fraud. In Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
  - 1/20: Uncertainty
    - Tversky, A., and D. Kahneman (1974). Judgment under uncertainty: heuristics and biases. Science 185:1124-1131.

- Dawes, R., D. Faust, and P. Meehl (1989). Clinical versus actuarial judgment. Science 243:1668-1674.

- 1/27: Markov Relational Fields and Conditional Relational Models
  - S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. IEEE Trans. PAMI, 6:721 - 741, 1984.
  - On the Statistical Analysis of Dirty Pictures, Julian Besag, Journal of the Royal Statistical Society B, vol. 48, 1986, pp. 259-302.
  - Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. John Lafferty, Andrew McCallum and Fernando Pereira. ICML-2001.
  - C. Anderson, P. Domingos and D. Weld, Relational Markov Models and their Application to Adaptive Web Navigation. Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining (pp. 143-152), 2002. Edmonton, Canada: ACM Press.
- 2/3: Conditional Relational Models (cont'd)
  - Flach, P. and N. Lachiche (2004). Naive Bayesian classification of structured data. Machine Learning, 57(3): 233-269.
  - Neville, J., D. Jensen, L. Friedland and M. Hay (2003). Learning Relational Probability Trees. In Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
  - Popescul, A., L. Ungar, S. Lawrence, D. Pennock (2003). Statistical Relational Learning for Document Mining. In Proceedings of IEEE International Conference on Data Mining (ICDM 2003).
  - Perlich, C., and F. Provost (2006). ACORA: Distribution-based Aggregation for Relational Learning from Identifier Attributes. Machine Learning, 62 (1/2) 65-105.
- 2/10: Graphical Models
  - Project proposals due
  - E. Charniak (1991). Bayesian Networks without Tears. AI magazine.
  - Koller, D., N. Friedman, L. Getoor, and B. Taskar (2007). Graphical Models in a Nutshell. *Introduction to Statistical Relational Learning*, L. Getoor and B. Taskar, editors, to appear.
- 2/17: Probabilistic Relational Models
  - Getoor, L., N. Friedman, D. Koller, and A. Pfeffer (2001). Learning Probabilistic Relational Models. Relational Data Mining, S. Dzeroski and N. Lavrac, Eds., Springer-Verlag.
  - Taskar, B., P. Abbeel, and D. Koller (2002). Relational Markov Networks. Introduction to Statistical Relational Learning, L. Getoor and B. Taskar, editors, to appear.
  - Neville, J., and D. Jensen (2007). Relational Dependency Networks. Introduction to Statistical Relational Learning, L. Getoor and B. Taskar.
- 2/24: Inductive Logic Programming and Probabilistic logic models
  - Dzeroski, S. (2007). Inductive Logic Programming in a Nutshell. Introduction to Statistical Relational Learning, L. Getoor and B. Taskar, editors, to appear.

        - Kersting, K., and L. De Raedt (2007). Bayesian Logic Programming: Theory and Tools. Introduction to Statistical Relational Learning, L. Getoor and B. Taskar, editors, to appear.
        - Richardson, M., and P. Domingos (2006). Markov Logic Networks. Machine Learning, 62, 107-136.
    - 3/3: Representational Issues and Learning Issues
        - Heckerman, D., C. Meek, and D. Koller (2004). Probabilistic Models for Relational Data. Microsoft Research Technical Report, MSR-TR-2004-30.
        - Milch, B., B. Marthi, S. Russell, D. Sontag, D. Ong, and A. Kolobov (2007). BLOG: Probabilistic Models with Unknown Objects. Introduction to Statistical Relational Learning, L. Getoor and B. Taskar, editors, to appear.
        - Jensen, D. and J. Neville (2002). Linkage and Autocorrelation Cause Feature Selection Bias in Relational Learning. In Proceedings of the 19th International Conference on Machine Learning.
        - Jensen, D., J. Neville and M. Hay (2003). Avoiding Bias When Aggregating Relational Data with Degree Disparity. In Proceedings of the 20th International Conference on Machine Learning.
    - 3/10: Inference Issues
        - Project progress report due
        - Macskassy, S. and F. Provost (2006). Classification in Networked Data: A toolkit and a univariate case study. Journal of Machine Learning Research, 8(May):935-983, 2007.
        - Taskar, B. V. Chatalbashev and D. Koller (2004). Learning Associative Markov Networks. In Proceedings of the 21st International Conference on Machine Learning (ICML04).
        - Milch, B., B. Marthi, D. Sontag, S. Russell, D. Ong, and A. Kolobov (2005). Approximate Inference for Infinite Contingent Bayesian Networks. In Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS).
        - Milch, B., and S. Russell (2006). General-Purpose MCMC Inference over Relational Structures. In Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI).
    - 3/17 – Spring recess. No class.
    - 3/24: Dynamic Relational Models
        - C. Anderson, P. Domingos, and D. Weld (2002). Relational Markov models and their application to adaptive Web navigation. In Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining.
        - Sanghai, S., P. Domingos, and D. Weld (2003). Dynamic Probabilistic Relational Models. In Proceedings of the 18th International Joint Conference on Artificial Intelligence.
    - 3/31: Group Discovery
        - Kubica, J., A. Moore, J. Schneider, and Y. Yang (2002). Stochastic link and group detection. In Proceedings of the 18th National Conference on Artificial Intelligence, pages 798--804.

        - Neville, J. and D. Jensen (2005). Leveraging Relational Autocorrelation with Latent Group Models. In Proceedings of the Fifth IEEE International Conference on Data Mining.
    - 4/7: Link Prediction
        - Taskar, B., Abbeel, P., Wong, M., and Koller, D. (2003). Label and Link Prediction in Relational Data. In Advances in Neural Information Processing Systems (NIPS) 16.
        - Liben-Nowell, D. and Kleinberg, J. (2003). The Link Prediction Problem for Social Networks. In Proceedings of the 12th International Conference on Information and Knowledge Management (CIKM).
    - 4/14: Entity resolution
        - Pasula, H., B. Marthi, B. Milch, S. Russell and I. Shpitser (2003). Identity uncertainty and citation matching. In NIPS 15.
        - Culotta, A. and A. McCallum (2005). Joint deduplication of multiple record types in relational data. In Proceedings of the 14th Conference on Information and Knowledge Management.
    - 4/21: Graph Mining
        - Cook, D. and L. Holder (2000). Graph-Based Data Mining. IEEE Intelligent Systems, 15(2), pages 32-41.
        - Dehaspe, L., H. Toivonen, R. King (1998). Finding frequent substructures in chemical compounds. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining.
    - 4/28: Project presentations
        - Students will present their research projects.

- Policies related to late or make-up work, if relevant.

Statement for Students with Disabilities
Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me (or to TA) as early in the semester as possible. DSP is located in STU 301 and is open 8:30 a.m.–5:00 p.m., Monday through Friday. The phone number for DSP is (213) 740-0776.

Statement on Academic Integrity
USC seeks to maintain an optimal learning environment. General principles of academic honesty include the concept of respect for the intellectual property of others, the expectation that individual work will be submitted unless otherwise allowed by an instructor, and the obligations both to protect one's own academic work from misuse by others as well as to avoid using another's work as one's own. All students are expected to understand and abide by these principles. *Scampus*, the Student Guidebook, contains the Student Conduct Code in Section 11.00, while the recommended sanctions are located in Appendix A: http://www.usc.edu/dept/publications/SCAMPUS/gov/. Students will be referred to the Office of Student Judicial Affairs and Community Standards for further review, should there be any suspicion of academic dishonesty. The Review process can be found at: http://www.usc.edu/student-affairs/SJACS/.