Probabilistic Graphical Models and their Role in Databases

Amol Deshpande
University of Maryland
amol@cs.umd.edu

Sunita Sarawagi
IIT Bombay
sunita@iitb.ac.in

Probabilistic graphical models provide a framework for compact representation and efficient reasoning about the joint probability distribution of several interdependent variables. This is a classical topic with roots in statistical physics. In recent years, spurred by several applications in unstructured data integration, sensor networks, image processing, bio-informatics, and code design, the topic has received renewed interest in the machine learning, data mining, and database communities. Techniques from graphical models have also been applied to many topics directly of interest to the database community including information extraction, sensor data analysis, imprecise data representation and querying, selectivity estimation for query optimization, and data privacy. As database research continues to expand beyond the confines of traditional enterprise domains, we expect both the need and applicability of probabilistic graphical models to increase dramatically over the next few years. With this tutorial, we are aiming to provide a foundational overview of probabilistic graphical models to the database community, accompanied by a brief overview of some of the recent research literature on the role of graphical models in databases.

Part I: Foundations of graphical models

The first part of the tutorial will be an approachable introduction to the foundations of probabilistic graphical models [15]. We will cover basic representation issues of directed and undirected graphical models, and algorithms for answering various kinds of queries on such models. Finally, we will cover techniques for learning the parameters and structure of a graphical model given example data.

Representation

Graphical models provide a compact representation of the full joint distribution of a set of variables: $X_1 \ldots X_n$ as a graph whose nodes are the variables and whose edges connect variables that interact directly. Variables that are not directly connected are conditionally independent under some combination of the other variables. A key concept in the graphical model formalism is the use of these independencies to represent the full joint distribution as a product of *factors* involving smaller subsets of variables.

We will discuss three types of models in detail:

- *Directed graphical models*, popularly known as *Bayesian networks*, are typically used to represent causal or assymetric interactions amongst variables. Some popular examples of Bayesian networks are Hidden Markov Models, Kalman Filters, and QMR networks.
- *Probabilistic Relational Models* [8] are a class of Bayesian networks particularly suited for representing relational databases. A PRM contains a relational component that describes the relational schema of the domain, and a probabilistic component that captures the probabilistic dependencies that hold in the domain. The result is a rich and powerful formalism that allows reasoning about data that is inherently relational in nature.
- *Undirected graphical models (MRFs, CRFs)*, or Markov Random Fields, are useful for representing distributions over variables where there is no natural directionality of the influence of one variable over another and where the interactions are more symmetric. Examples are the interactions between atoms in a molecular structure, the dependency between the labels of pixels of an image, or the interactions between environmental properties sensed by geographically co-located sensors [6].

Inference Queries

Once the prior knowledge about the environment or the application domain has been encoded in the form of a probabilistic graphical model (with the probability distribution parameters possibly learned from historical data), three kinds of tasks may be performed using the model:

- *Adding evidence*: additional knowledge about the variables, possibly gained by observing them, needs to be incorporated into the model.
- *Marginal probability queries over a small subset of variables:* At any point, we may wish to find out the marginal probability distribution over a subset of the variables.
- *Computing most likely labels of a subset of variables:* These queries are similar to above, but we are only interested in the most likely values of the subset of variables. In many cases, these queries can be executed more efficiently than marginal probability queries.

We will discuss algorithms for both exact and approximate evaluation of these tasks.

Part II: Database Applications

In the second part, we will survey several representative database applications where graphical models have proven to be useful.

VLDB '07, September 23-28, 2007, Vienna, Austria.

- **Probabilistic databases:** An increasing number of real-world applications are demanding support for managing, storing, and querying probabilistic data in relational database systems. This has led to a recent resurgence of research in this area, spanning a wide range of issues from theoretical development of data models and data languages to practical aspects such as indexing [9, 4] and creating imprecise databases to represent uncertain data sources [11]. Probabilistic graphical models present perhaps the most attractive option for both capturing and representing the uncertainty in the data and also for efficient evaluation of queries over such data [25, 4].

- **Information extraction and integration:** Modern techniques for information extraction rely on a number of inter-related clues to automatically extract structured entities from unstructured text. Some of the most successful information extraction systems are therefore based on graphical models for combining evidence. We will review early probabilistic extraction systems based on Hidden Markov Models [7, 1] and Maximum Entropy Models [18] and then cover the state-of-the-art methods based on Conditional Random Fields [16] and Max-margin markov networks [26]. We will present various graphical models for extraction, starting with traditional chain models for plain text to segmentation models [24, 3, 23] for exploiting matches with existing entities, and general graph models for extracting from visual 2D layouts as in web pages.

- **Sensor data management:** Many common sensor processing tasks can be seen as applications of specific instances of graphical models, specifically dynamic Bayesian networks [20, 19, 14], to streaming sensor data. Examples of such tasks include (a) eliminating measurement noise (sometimes called "white noise") from the observed data using filtering techniques such as *Kalman filters* [27], (b) predicting missing readings using historical data, (c) inferring *hidden* variables using *hidden Markov models* [22]), and (d) automatically detecting novel or anomalous behavior [17]. Probabilistic graphical models have also been shown to be useful in approximate querying [6] and data collection in sensor networks [13, 2].

- **Selectivity estimation for query optimization:** Probabilistic graphical models can be used to capture the correlations present in the data, to aid in better selectivity estimation. Getoor et al. [10] used probabilistic relational models (PRMs) for this purpose. Deshpande et al. [5] proposed *dependency-based histograms*, based on a class of undirected models, called *decomposable models*. Pavlov et al. [21] explore similar techniques for approximate querying over large sparse binary transaction data. Ilyas et al. [12] identify and exploit dependency structures similar to graphical models for discovering soft functional dependencies.

Presenters

Amol Deshpande is an Assistant Professor of Computer Science at the University of Maryland at College Park. He received his PhD from UC Berkeley in 2004, and his bachelors degree from IIT Bombay. His research interests include adaptive query processing, data streams, sensor networks and statistical modeling of data. He is a recipient of the NSF CAREER Award.

Sunita Sarawagi researches in the fields of databases, data mining, machine learning and statistics. She is associate professor at IIT Bombay. Prior to that she was a research staff member at IBM Almaden Research Center. She got her PhD in databases from the University of California at Berkeley and a bachelors degree from IIT Kharagpur. She has several publications in databases and data mining including a best paper award at the 1998 ACM SIGMOD conference and several patents.

References

[1] V. R. Borkar, K. Deshmukh, and S. Sarawagi. Automatic text segmentation for extracting structured records. In *SIGMOD*, 2001.

[2] D. Chu, A. Deshpande, J. Hellerstein, and W. Hong. Approximate data collection in sensor networks using probabilistic models. In *ICDE*, 2006.

[3] W. W. Cohen and S. Sarawagi. Exploiting dictionaries in named entity extraction: Combining semi-markov extraction processes and data integration methods. In *SIGKDD*, 2004.

[4] N. Dalvi and D. Suciu. Management of probabilistic data: Foundations and challenges. In *PODS*, 2007.

[5] A. Deshpande, M. Garofalakis, and R. Rastogi. Independence is good: Dependency-based histogram synopses for high-dimensional data. In *SIGMOD*, 2001.

[6] A. Deshpande, C. Guestrin, S. Madden, J. M. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. In *VLDB*, 2004.

[7] D. Freitag, A. McCallum. Information extraction with HMM structures learned by stochastic optimization. In *AAAI*, 2000.

[8] N. Friedman, L. Getoor, D. Koller, and A. Pfeffer. Learning probabilistic relational models. In *IJCAI*, 1999.

[9] M. Garofalakis and D. Suciu, editors. *IEEE Data Eng. Bull. Special Issue on Probabilistic Data Management*. March 2006.

[10] L. Getoor, B. Taskar, and D. Koller. Selectivity estimation using probabilistic models. In *SIGMOD*, 2001.

[11] R. Gupta and S. Sarawagi. Creating probabilistic databases from information extraction models. In *VLDB*, 2006.

[12] I. F. Ilyas et al. Cords: automatic discovery of correlations and soft functional dependencies. In *SIGMOD*, 2004.

[13] A. Jain, E. Chang, and Y.-F. Wang. Adaptive stream resource management using Kalman Filters. In *SIGMOD*, 2004.

[14] B. Kanagal and A. Deshpande. Online filtering, smoothing and probabilistic modeling of streaming data. Technical Report CS-TR-4867, University of Maryland, 2007.

[15] D. Koller and N. Friedman. *Structured Probabilistic Models*. In preparation, 2006.

[16] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

[17] M. Markou and S. Singh. Novelty detection: a review - part 1: statistical approaches. *Signal Processing*, 83(12), 2003.

[18] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy markov models for information extraction and segmentation. In *ICML*, 2000.

[19] V. Mihajlovic and M. Petkovic. Dynamic bayesian networks: A state of the art. Univ. of Twente Document Repository, 2001.

[20] K. Murphy. A brief introduction to graphical models and bayesian networks, 1998.

[21] D. Pavlov, H. Mannila, and P. Smyth. Beyond independence: Probabilistic models for query approximation on binary transaction data. *IEEE TKDE*, 2003.

[22] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. 77:257–286, 1989.

[23] S. Sarawagi. Efficient inference on sequence segmentation models. In *ICML*, 2006.

[24] S. Sarawagi and W. W. Cohen. Semi-markov conditional random fields for information extraction. In *NIPS*, 2004.

[25] P. Sen and A. Deshpande. Representing and querying correlated tuples in probabilistic databases. In *ICDE*, 2007.

[26] B. Taskar. *Learning Structured Prediction Models: A Large Margin Approach*. PhD thesis, Stanford University, 2004.

[27] G. Welch and G. Bishop. An introduction to Kalman filter. `http://www.cs.unc.edu/~welch/kalman/kalmanIntro.html`, 2002.