

# Probabilistic Graphical Models

*Principles and Techniques*

Daphne Koller

Nir Friedman

The MIT Press  
Cambridge, Massachusetts  
London, England

©2009 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

For information about special quantity discounts, please email [special\\_sales@mitpress.mit.edu](mailto:special_sales@mitpress.mit.edu)

This book was set by the authors in  $\text{\LaTeX}$ .  
Printed and bound in the United States of America.

#### Library of Congress Cataloging-in-Publication Data

Koller, Daphne.

Probabilistic Graphical Models: Principles and Techniques / Daphne Koller and Nir Friedman.

p. cm. – (Adaptive computation and machine learning)

Includes bibliographical references and index.

ISBN 978-0-262-01319-2 (hardcover : alk. paper)

1. Graphical modeling (Statistics) 2. Bayesian statistical decision theory—Graphic methods. I.

Koller, Daphne. II. Friedman, Nir.

QA279.5.K65 2010

519.5'420285—dc22

2009008615

10 9 8 7 6 5 4 3 2 1

## Notation Index

- $|A|$  — Cardinality of the set  $A$ , 20  
 $\phi_1 \times \phi_2$  — Factor product, 107  
 $\gamma_1 \oplus \gamma_2$  — Joint factor combination, 1102  
 $p(\mathbf{Z}) \oplus g(\mathbf{Z})$  — Marginal of  $g(\mathbf{Z})$  based on  $p(\mathbf{Z})$ , 631  
 $\sum_Y \phi$  — Factor marginalization, 297  
 $X \rightleftharpoons Y$  — Bi-directional edge, 34  
 $X \rightarrow Y$  — Directed edge, 34  
 $X - Y$  — Undirected edge, 34  
 $X \leftrightarrow Y$  — Non-ancestor edge (PAGs), 1048  
 $X \circ \rightarrow Y$  — Ancestor edge (PAGs), 1048  
 $\langle x, y \rangle$  — Inner product of vectors  $x$  and  $y$ , 262  
 $\|P - Q\|_1$  —  $L_1$  distance, 1141  
 $\|P - Q\|_2$  —  $L_2$  distance, 1141  
 $\|P - Q\|_\infty$  —  $L_\infty$  distance, 1141  
 $(\mathbf{X} \perp \mathbf{Y})$  — Independence of random variables, 24  
 $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$  — Conditional independence of random variables, 24  
 $(\mathbf{X} \perp_c \mathbf{Y} \mid \mathbf{Z}, \mathbf{c})$  — Context-specific independence, 162  
 $\mathbf{I}\{\cdot\}$  — Indicator function, 32  
 $\mathcal{A}(x \rightarrow x')$  — Acceptance probability, 517  
 $\aleph$  — Template attributes, 214  
 $\alpha(A)$  — The argument signature of attribute  $A$ , 213  
 $\text{Ancestors}_X$  — Ancestors of  $X$  (in graph), 36  
 $\text{argmax}$ , 26  
 $A$  — A template attribute, 213  
 $\text{Beta}(\alpha_1, \alpha_0)$  — Beta distribution, 735  
 $\beta_i$  — Belief potential, 352  
 $\mathcal{B}_{\mathcal{I}[\sigma]}$  — Induced Bayesian network, 1091  
 $\mathcal{B}$  — Bayesian network, 62  
 $\mathcal{B}_0$  — Initial Bayesian network (DBN), 204  
 $\mathcal{B}_{\rightarrow}$  — Transition Bayesian network (DBN), 204  
 $\mathcal{B}_{\mathbf{Z}=\mathbf{z}}$  — Mutilated Bayesian network, 499  
 $\mathcal{C}(K, \mathbf{h}, g)$  — Canonical form, 609  
 $\mathcal{C}(\mathbf{X}; K, \mathbf{h}, g)$  — Canonical form, 609  
 $\mathcal{C}[v]$  — Choices, 1083  
 $\text{Ch}_X$  — Children of  $X$  (in graph), 34  
 $\mathcal{C}_i$  — Clique, 346  
 $x \sim c$  — Compatability of values , 20  
 $\text{cont}(\gamma)$  — Joint factor contraction, 1102  
 $\text{Cov}[X; Y]$  — Covariance of  $X$  and  $Y$ , 248  
 $\mathcal{D}$  — A subclique, 104  
 $\Delta$  — Discrete variables (hybrid models), 605  
 $d$  — Value of a subclique, 104  
 $\mathcal{D}^+$  — Complete data, 871  
 $\mathcal{D}$  — Empirical samples (data), 698  
 $\mathcal{D}$  — Sampled data, 489  
 $\mathcal{D}^*$  — Complete data, 912  
 $\mathcal{D}$  — Decisions, 1087  
 $\text{Descendants}_X$  — Descendants of  $X$  (in graph), 36  
 $\tilde{\delta}_{i \rightarrow j}$  — Approximate sum-product message, 435  
 $\delta_{i \rightarrow j}$  — Sum-product message, 352  
 $\text{Dim}[\mathcal{G}]$  — Dimension of a graph, 801  
 $\text{Dirichlet}(\alpha_1, \dots, \alpha_K)$  — Dirichlet distribution, 738  
 $D(P\|Q)$  — Relative entropy, 1139  
 $D_{\text{var}}(P; Q)$  — Variational distance, 1141  
 $\text{Down}^*(r)$  — Downward closure, 422  
 $\text{Down}^+(r)$  — Extended downward closure, 422  
 $\text{Down}(r)$  — Downward regions, 422  
 $do(Z := z), do(z)$  — Intervention, 1010  
 $d\text{-sep}_{\mathcal{G}}(\mathbf{X}; \mathbf{Y} \mid \mathbf{Z})$  — d-separation, 71  
 $\mathcal{E}$  — Edges in MRF, 127

- $\text{EU}[\mathcal{D}[a]]$  — Expected utility, 1059  
 $\text{EU}[\mathcal{I}[\sigma]]$  — Expected utility of  $\sigma$ , 1091  
 $\tilde{\mathbf{E}}_{\mathcal{D}}(f)$  — Empirical expectation, 490  
 $\mathbf{E}_{\mathcal{D}}[f]$  — Empirical expectation, 700  
 $\mathbf{E}_P[X]$  — Expectation (mean) of  $X$ , 31  
 $\mathbf{E}_P[X \mid \mathbf{y}]$  — Conditional expectation, 32  
 $\mathbf{E}_{X \sim P}[\cdot]$  — Expectation when  $X \sim P$ , 387
- $f(\mathbf{D})$  — A feature, 124  
 $F[\tilde{P}, \mathcal{Q}]$  — Energy functional, 385, 881  
 $\tilde{F}[\tilde{P}_{\Phi}, \mathcal{Q}]$  — Region Free Energy functional, 420  
 $\tilde{F}[\tilde{P}_{\Phi}, \mathcal{Q}]$  — Factored energy functional, 386  
 $\text{FamScore}(X_i \mid \text{Pa}_i : \mathcal{D})$  — Family score, 805  
 $\mathcal{F}$  — Feature set, 125  
 $\mathcal{F}$  — Factor graph, 123
- $\mathcal{G}$  — Directed graph, 34  
 $\mathcal{G}$  — Partial ancestral graph, 1048  
 $\Gamma$  — Continuous variables (hybrid models), 605  
 $\gamma$  — Template assignment, 215  
 $\text{Gamma}(\alpha, \beta)$  — Gamma distribution, 900  
 $\Gamma(x)$  — Gamma function, 736
- $\mathcal{H}$  — Missing data, 859  
 $\mathcal{H}$  — Undirected graph, 34  
 $\mathbf{H}_P(X)$  — Entropy, 1136  
 $\mathbf{H}_P(X \mid Y)$  — Conditional entropy, 1137  
 $\tilde{\mathbf{H}}_{\mathcal{Q}}^*(\mathcal{X})$  — Weighted approximate entropy, 415
- $\mathcal{I}$  — Influence diagram, 1088  
 $\mathcal{I}(\mathcal{G})$  — Markov independencies of  $\mathcal{G}$ , 72  
 $\mathcal{I}_\ell(\mathcal{G})$  — Local Markov independencies of  $\mathcal{G}$ , 57  
 $\mathcal{I}(P)$  — The independencies satisfied by  $P$ , 60  
 $\mathbf{I}_P(X; Y)$  — Mutual information, 1138  
 $\text{Interface}_{\mathcal{H}}(\mathbf{X}; \mathbf{Y})$  —  $\mathbf{Y}$ -interface of  $\mathbf{X}$ , 464
- $\mathcal{J}$  — Lagrangian, 1166  
 $J$  — Precision matrix, 248
- $\mathcal{K}$  — Partially directed graph, 34  
 $\mathcal{K}^+[\mathbf{X}]$  — Upward closed subgraph, 35  
 $\kappa$  — Object skeleton (template models), 214  
 $\kappa_r$  — Counting number of region  $r$ , 415  
 $\mathbf{K}_i$  — Member of a chain, 37  
 $\mathcal{K}[\mathbf{X}]$  — Induced subgraph, 35
- $\ell_{\text{PL}}(\boldsymbol{\theta} : \mathcal{D})$  — Pseudolikelihood, 970  
 $L(\boldsymbol{\theta} : \mathcal{D})$  — Likelihood function, 721  
 $\text{Local}[\mathcal{U}]$  — Local polytope, 412  
 $\ell(\hat{\boldsymbol{\theta}}_{\mathcal{G}} : \mathcal{D})$  — Maximum likelihood value, 791  
 $\ell(\boldsymbol{\theta} : \mathcal{D})$  — Log-likelihood function, 719  
 $\ell_{\mathbf{Y} \mid \mathbf{X}}(\boldsymbol{\theta} : \mathcal{D})$  — Conditional log-likelihood function, 951  
 $\text{loss}(\xi : \mathcal{M})$  — Loss function, 699
- $\mathcal{M}^*$  — Model that generated the data, 698  
 $\text{M-project-distr}_{i,j}$  — M-projection, 436  
 $M[\mathbf{x}]$  — Counts of event  $\mathbf{x}$  in data, 724  
 $\text{Marg}[\mathcal{U}]$  — Marginal polytope, 411  
 $\text{marg}_{\mathbf{W}}(\gamma)$  — Joint factor marginalization, 1102  
 $\text{MaxMarg}_f(\mathbf{x})$  — Max marginal of  $f$ , 553  
 $\mathcal{M}[\mathcal{G}]$  — Moralization of  $\mathcal{G}$ , 134  
 $\mathcal{M}$  — A model, 699  
 $\tilde{M}_{\boldsymbol{\theta}}[\mathbf{x}]$  — Expected counts, 871  
 $\tilde{\mathcal{M}}$  — Learned/estimated model, 698
- $\mathcal{N}(\mu; \sigma^2)$  — A Gaussian distribution, 28  
 $\mathcal{N}(X \mid \mu; \sigma^2)$  — Gaussian distribution over  $X$ , 616  
 $\text{Boundary}_X$  — Boundary around  $X$  (in graph), 34  
 $\text{Nb}_X$  — Neighbors of  $X$  (in graph), 34  
 $\text{NonDescendants}_X$  — Non-descendants of  $X$  (in graph), 36  
 $\mathcal{N}\mathcal{P}$ , 1149
- $\mathcal{O}$  — Outcome space, 1058  
 $O(f(\cdot))$  — “Big O” of  $f$ , 1146  
 $\mathcal{O}^\kappa[\mathcal{Q}]$  — Objects in  $\kappa$  (template models), 214
- $\mathcal{P}$ , 1149  
 $P(X \mid Y)$  — Conditional distribution, 22  
 $P(x), P(x, y)$  — Shorthand for  $P(X = x), P(X = x, Y = y)$ , 21  
 $P^*$  — Distribution that generated the data, 698  
 $P \models \dots$  —  $P$  satisfies  $\dots$ , 23  
 $\text{Pa}_X$  — Parents of  $X$  (in graph), 34  
 $\text{pa}_X$  — Value of  $\text{Pa}_X$ , 157  
 $\text{Pa}_{X_i}^{\mathcal{G}}$  — Parents of  $X_i$  in  $\mathcal{G}$ , 57  
 $\hat{P}_{\mathcal{D}}(A)$  — Empirical distribution, 703  
 $\hat{P}_{\mathcal{D}}(\mathbf{x})$  — Empirical distribution, 490  
 $\boldsymbol{\theta}$  — Parameters, 262, 720  
 $\hat{\boldsymbol{\theta}}$  — MLE parameters, 726  
 $\phi$  — A factor (Markov network), 104  
 $\phi[\mathbf{U} = \mathbf{u}]$  — Factor reduction, 110

- $\pi$  — Lottery, 1058  
 $\pi(\mathbf{X})$  — Stationary probability, 509  
 $\tilde{P}_{\Phi}(\mathcal{X})$  — Unnormalized measure defined by  $\Phi$ , 345  
 $\psi_i(\mathbf{C}_i)$  — Initial potential, 349  
 $\tilde{P}$  — Learned/estimated distribution, 698  
  
 $Q$  — Approximating distribution, 383  
 $\mathcal{Q}$  — Template classes, 214  
  
 $\mathcal{R}$  — Region graph, 419  
 $\mathbb{R}$  — Real numbers, 27  
 $\rho$  — A rule, 166  
 $\mathcal{R}$  — Rule set, 168  
  
 $\mathcal{S}$  — Event space, 15  
 $\sigma$  — Std of a Gaussian distribution, 28  
 $\sigma$  — Strategy, 1090  
 $\sigma^{(t)}(\cdot)$  — Belief state, 652  
 $Scope[\phi]$  — Scope of a factor, 104  
 $score_B(\mathcal{G} : \mathcal{D})$  — Bayesian score, 795  
 $score_{BIC}(\mathcal{G} : \mathcal{D})$  — BIC score, 802  
 $score_{CS}(\mathcal{G} : \mathcal{D})$  — Cheeseman-Stutz score, 913  
 $score_L(\mathcal{G} : \mathcal{D})$  — Likelihood score, 791  
 $score_{L_1}(\boldsymbol{\theta} : \mathcal{D})$  —  $L_1$  score, 988  
 $score_{Laplace}(\mathcal{G} : \mathcal{D})$  — Laplace score, 910  
 $score_{MAP}(\boldsymbol{\theta} : \mathcal{D})$  — MAP score, 898  
 $sep_{\mathcal{H}}(\mathbf{X}; \mathbf{Y} | \mathbf{Z})$  — Separation in  $\mathcal{H}$ , 114  
 $\text{sigmoid}(x)$  — Sigmoid function, 145  
 $\mathbf{S}_{i,j}$  — Sepset, 140, 346  
 $\text{succ}(v, c)$  — Successor (decision trees), 1083  
  
 $\mathcal{T}$  — Clique tree, 140, 347  
 $\Upsilon$  — Template clique tree, 656  
 $T$  — Decision tree, 1083  
 $\mathbf{t}(\boldsymbol{\theta})$  — Natural parameters function, 261  
 $\tau(\xi)$  — Sufficient statistics function, 261, 721  
 $\Theta$  — Parameter space, 261, 720  
 $T(\mathbf{x} \rightarrow \mathbf{x}')$  — Transition probability, 507  
  
 $\mathcal{U}$  — Cluster graph, 346  
 $\mathcal{U}$  — Response variables, 1029  
 $\mu$  — Mean of a Gaussian distribution, 28  
 $U(o)$  — Utility function, 1058  
 $\mu_{i,j}$  — Sepset beliefs, 358  
 $\text{Unif}[a, b]$  — Uniform distribution on  $[a, b]$ , 28  
 $\mathbf{Up}^*(r)$  — Upward closure, 422  
 $\mathbf{Up}(r)$  — Upward regions, 422  
  
 $\mathcal{U}$  — Utility variables, 1088  
 $U^X$  — Response variable, 1029  
  
 $Val(X)$  — Possible values of  $X$ , 20  
 $\text{Var}_P[X]$  — Variance of  $X$ , 33  
 $\text{VPI}_{\mathcal{I}}(D | X)$  — Value of perfect information, 1120  
 $\nu_r, \nu_i, \nu_{r,i}$  — Convex counting numbers, 416  
  
 $\mathbf{W}_{\langle i,j \rangle}$ , 348  
  
 $\mathcal{X}$  — The set of all variables in the domain, 21  
 $\xi$  — An assignment to  $\mathcal{X}$ , 79  
 $X, Y, Z$  — Random variables, 20  
 $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$  — Random variable sets, 20  
 $\mathbf{x}, \mathbf{y}, \mathbf{z}$  — Values of random variable sets, 20  
 $x^0, x^1$  — False/True values of  $X$ , 20  
 $\mathbf{x}(\mathbf{Y})$  — Assignment in  $\mathbf{x}$  to variables in  $\mathbf{Y}$ , 21  
 $\mathbf{x}[m]\mathbf{x}[m]$  —  $m$ 'th data instance (i.i.d. samples), 698  
 $x^i$  — The  $i$ 'th value of  $X$ , 20  
 $\mathcal{X}_R[A]$  — Ground random variables, 214  
 $\xi[m]$  —  $m$ 'th data instance (i.i.d. samples), 488  
 $\xi^{map}$  — MAP assignment, 552  
 $X^{(t)}$  —  $X$  at time  $t$ , 200  
 $X^{(t_1:t_2)}$  —  $X$  in the interval  $[t_1, t_2]$ , 200  
 $X \sim \dots$  —  $X$  is distributed according to  $\dots$ , 28  
  
 $Z$  — Partition function, 105