Probabilistic Graphical Models (PGMs)

Part 1

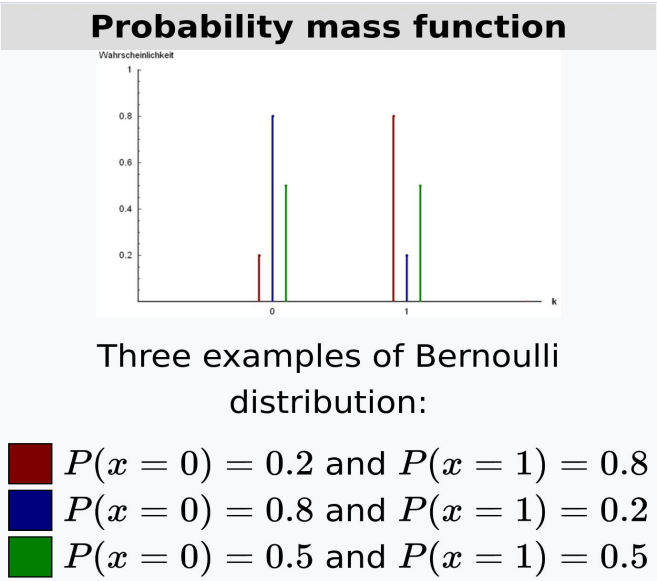
Anna Saranti

2026.05.03

Probabilistic Graphical Models (PGMs) (1/5)

- Representations of joint probability distributions over RVs
- The graph represents the probabilistic relationships
- Discrete: categorical values
  Continuous: real values
- Domain of RV: The set of all possible values
  Remember the previous lecture $\mathcal{A}_{\mathbf{X}}$

Probabilistic Graphical Models (PGMs) (2/5)



- Example of a Bernoulli distribution:
  `https://en.wikipedia.org/wiki/Bernoulli_distribution`
  coss a toin - 2 possible outcomes

Probabilistic Graphical Models (PGMs) (3/5)

- Visible RVs:
  - have outcomes that can be directly observed
  - their values are contained in the dataset
- Hidden/Latent RVs:
  - defined by human experts using the domain knowledge of the problem
  - their outcomes are not directly accessible
  - represent latent causes of visible random variables
  - improve the accuracy and interpretability of model

Probabilistic Graphical Models (PGMs) (4/5)

- How to specify the dependencies?
  Need direction, type and intensity
- Graphs: Nodes are RVs, Edges are dependencies
- Undirected models (called Markov networks) represent symmetric probabilistic interactions - no dependency with direction, only factors that represent the degree of the strength of the connection
- Directed acyclic graph (DAG), otherwise circular reasoning would be possible

Probabilistic Graphical Models (PGMs) (5/5)

- The joint distribution represented by the model is mathematically expressed by the chain rule: $P(X_1, X_2, \cdots X_N) = \prod_{n=1}^{N} P(X_n | Parents_G(X_n))$
- Expresses the factorization of the distribution
- $P(X_n | Parents_G(X_n))$ conditional distributions represent local probability models that have their local likelihood and the estimation of their parameters
- The value that a random variable will take is in general dependent on the values of its parent(s) random variable(s)

Independence of two RVs (1/2)

- Independence denotes a situation where knowing about the value of one RV does not add any new information about the value of another
- Conditional independence: knowledge about a particular RV C turns previously dependent RVs A and B to independent:

  $$P(A, B|C) = P(A|C)P(B|C)$$

- The graph structure of the model can be used to find which variables are conditionally independent

Independence of two RVs (2/2)

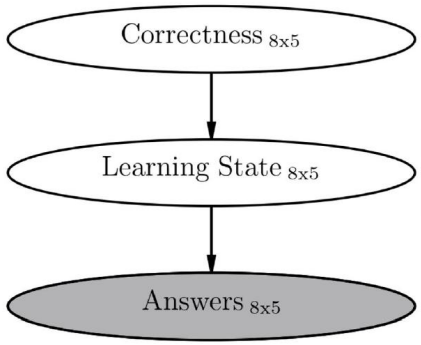
- The independence of two random variables A and B is not conditioned on the knowledge of just one variable, but a set of variables $\mathcal{C}$ -
  directional or d-separation of A and B with regard to $\mathcal{C}$
- Markov blanket: the smallest set
  can be computed from specific rules

- **The model reflects our belief [the human's belief] about the modelled problem**

Conditional Probability Distributions (1/2)

- Discrete RVs $\rightarrow$ Conditional Probability Table (CPT):
  - Number of rows of each CPT equals the number of combinations of all possible values of all parent variable(s)
  - Number of columns equals the number of values of the child variable
  - Each row of a CPT has values that sum up to 1.0

Conditional Probability Distributions (2/2)



wrong	correct
0.5	0.5

	operand	···	unclassified	correct
wrong	1/7	···	1/7	0.0
correct	0.0	···	0.0	1.0

	···	36	37	38	39	40	41	42	43	···
operand	···	0.083	0.0	0.0	0.0	0.0	0.0	0.083	0.0	···
intrusion	···	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	···
consistency	···	0.0	0.0	0.0	0.0	0.0	0.058	0.058	0.058	···
off-by	···	0.0	0.0	0.25	0.25	0.0	0.25	0.25	0.0	···
add/sub/div	···	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	···
pattern	···	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	···
unclassified	···	0.0	0.018	0.0	0.0	0.0	0.0	0.0	0.0	···
correct	···	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	···

- Visible: gray, hidden: white
- Conditional Independence Property:
  $\mathbf{Answes_q} \perp \mathbf{Correctness_q} \mid \mathbf{Learning\ State_q}$

Categorical Distribution (1/2)

- Bernoulli distribution - domain with 2 values:

$$P(x|\theta) = \begin{cases} \theta, & \text{if } x = 0 \\ 1-\theta, & \text{if } x = 1 \end{cases}$$

- Domain with many values:

$$p(x = k) = \theta_k, \ \sum_{k=1}^{K} \theta_k = 1$$

Dependencies and Inferrence

- The $\mathbf{Correctness_q}$ RV influences the $\mathbf{Learning\ State_q}$ and in turn $\mathbf{Learning\ State_q}$ is a cause to the particular $\mathbf{Answers_q}$ RV.
- The $\mathbf{Learning\ State_q}$ and the $\mathbf{Correctness_q}$ are inferred by the answers of the students in question q.

- Factorization of the joint probability distribution:
$P(\mathbf{Correctness_q}, \mathbf{Learning\ State_q}, \mathbf{Answers_q}) =$
$P(\mathbf{Correctness_q})\ P(\mathbf{Learning\ State_q}|\mathbf{Correctness_q})$
$P(\mathbf{Answers_q}|\mathbf{Learning\ State_q})$

Bayes Theorem (1/5)

- Dataset $\mathcal{D} = \{d[1], \cdots, d[N]\}$,
  $d[n]$ is one data sample
- Assume that the data samples $\mathcal{D}$ are independent and identically distributed (i.i.d.)
- The joint probability distribution $P_{\mathcal{M}}$ defined by the model $\mathcal{M}$ with parameters Θ
- The parameter learning's goal is to increase the likelihood of the data given the model: $P(\mathcal{D}|\mathcal{M})$ or equivalently the log-likelihood: $log\ P(\mathcal{D}|\mathcal{M})$ w.r.t. the set of the parameters Θ of the model

Bayes Theorem (2/5)

- Generative model: The likelihood expresses the probability of the data given a particular model
- If it were this model to have generated the data, which parameters would it have?
- A model that assigns a higher likelihood to the data $\mathcal{D}$ approximates the true distribution (the one that has generated the data) better.

Bayes Theorem (3/5)

- To initialize the parameters Θ one needs first to define their prior distribution, **which expresses our beliefs about them before seeing any data samples**
- **Our beliefs about the outcome of the toss of a coin can be that we will encounter "heads" with probability around 50% and "tails" also with probability 50%**
  But if we encounter a coin where 9 of 10 tosses the outcome is "heads", we revise our belief of a uniform distribution of the outcomes - we no longer believe that the coin is fair

Bayes Theorem (4/5)

- The revised belief is called posterior distribution
- The observation is called evidence
- The first posterior is computed when there was at least one observation in one or more variables of the model
- New evidence $\rightarrow$ the old posterior becomes the new prior

Bayes Theorem (5/5)

- $$\overbrace{P(\Theta|\mathcal{D})}^{posterior} = \frac{\overbrace{P(\mathcal{D}|\Theta)}^{likelihood}\overbrace{P(\Theta)}^{prior}}{\underbrace{P(\mathcal{D})}_{marginal\ likelihood}}$$

Priors and posteriors of categorical distributions (1/5)

- Phenomenon with 2 possible outcomes:
  Bernoulli distribution with Beta prior:

  $$P(\theta|a,b) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}, a, b > 0, 0 \leq \theta \leq 1$$

  a and b are the hyperparameters that represent the number of **pseudocounts** - the number of times we've encountered each of the 2 outcomes in previous experiments (or our belief about those outcomes in general)

- The prior is defined by the human
  The pseudocounts "are in your head"

Priors and posteriors of categorical distributions (2/5)

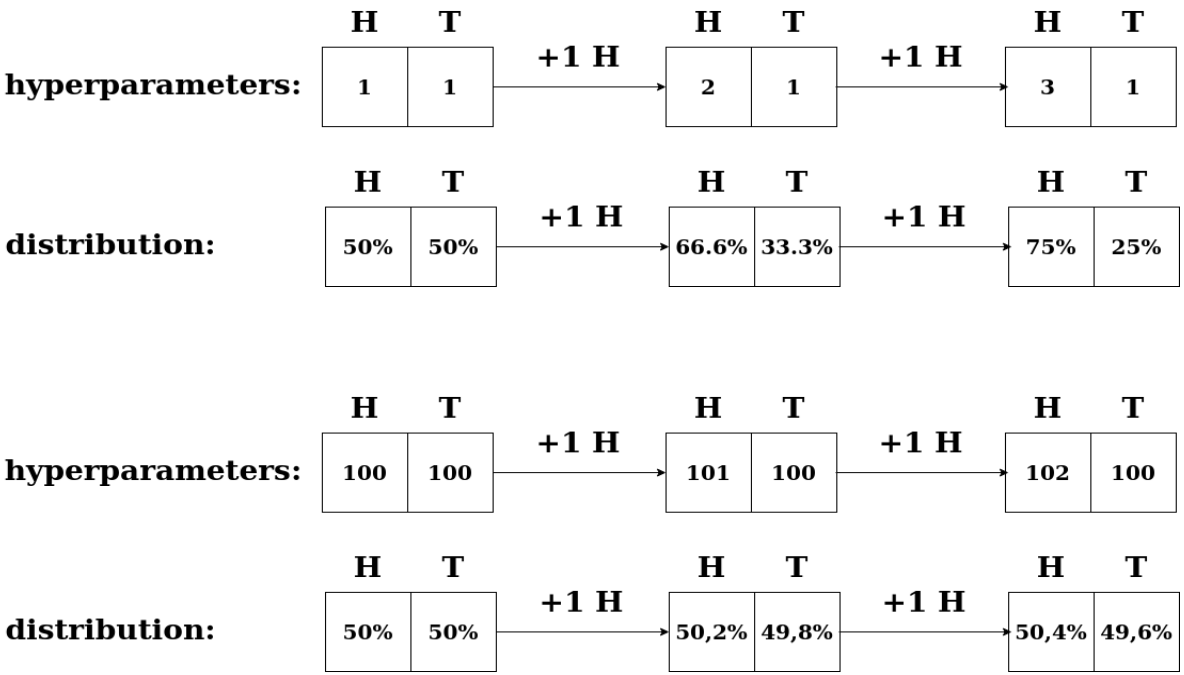
- Phenomenon with k possible outcomes:
  Categorical distribution with Dirichlet prior:

  $$Dirichlet(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Pi_{k=1}^{K}\Gamma(\alpha_k)}\Pi_{k=1}^{K}\theta_k^{\alpha_k-1}$$

  where $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$

- As in the Beta prior, the α_k are the hyperparameters that represent the number of pseudocounts of each possible outcome

Priors and posteriors of categorical distributions (3/5)



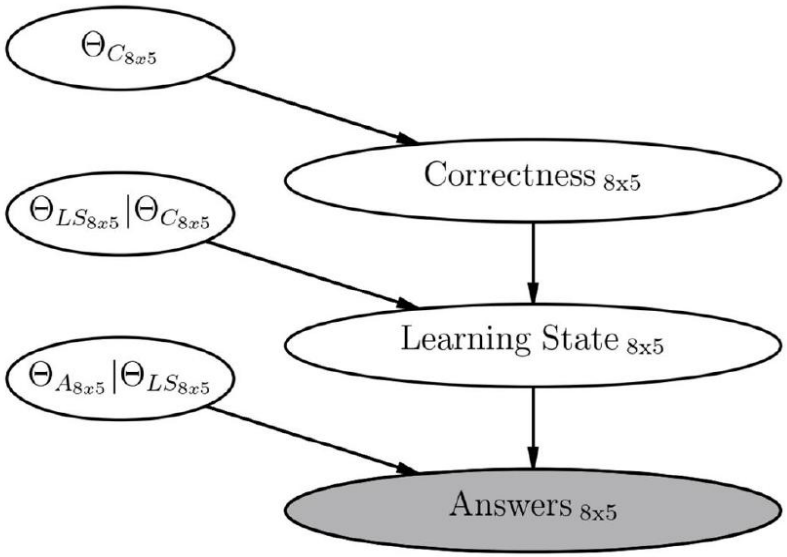
Bayes Theorem (4/5)

- Six-sided dice: $\mathit{Dirichlet}(1,1,1,1,1,1)$
  vs. $\mathit{Dirichlet}(100,100,100,100,100,100)$
- Parameters have the uniform distribution: $(\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6})$
- But in the second case more data will be necessary to make the parameters shift from the uniform distribution
- $P(\theta) = \mathit{Dirichlet}(a_1,\cdots,a_k)$
  after seeing $M[k]$ outcomes $\forall$ possible outcome k
  $P(\theta|\mathcal{D}) = \mathit{Dirichlet}(a_1+M[1],\cdots,a_k+M[k])$

Bayes Theorem (5/5)

- "weak" vs. "strong" prior
- If the prior "agrees" with the posterior - OK
- With a "strong" prior we will need more data to counteract our beliefs
- There exist other priors that show preference for one of the outcomes...
  How much data will you need to be "persuaded" that your beliefs are wrong?
- The posterior is also a belief -
  it is the prior of your next experiment...

Example with prior and posterior



wrong	correct
w_1	w_2

	operand	···	unclassified	correct
wrong	$\pi_{1\|wrong}$	···	$\pi_{k\|wrong}$	0.0
correct	0.0	···	0.0	1.0

	···	x	···
operand	···	···	···
···	···	$\theta_{x\|k}$	···
correct	···	···	···

- Bayesian parameter learning is applicable when all variables are visible -
  if not use Expectation-Maximization (EM)

Sufficient Statistics

- All possible outcomes of an RV are present at least one sample of the dataset
- For example: estimating the bias of a coin
  You can start with any possible prior
  If you do 10 experiments - 10 throwing the coin
  and you only see heads (**H**)
  you do not have sufficient statistics!

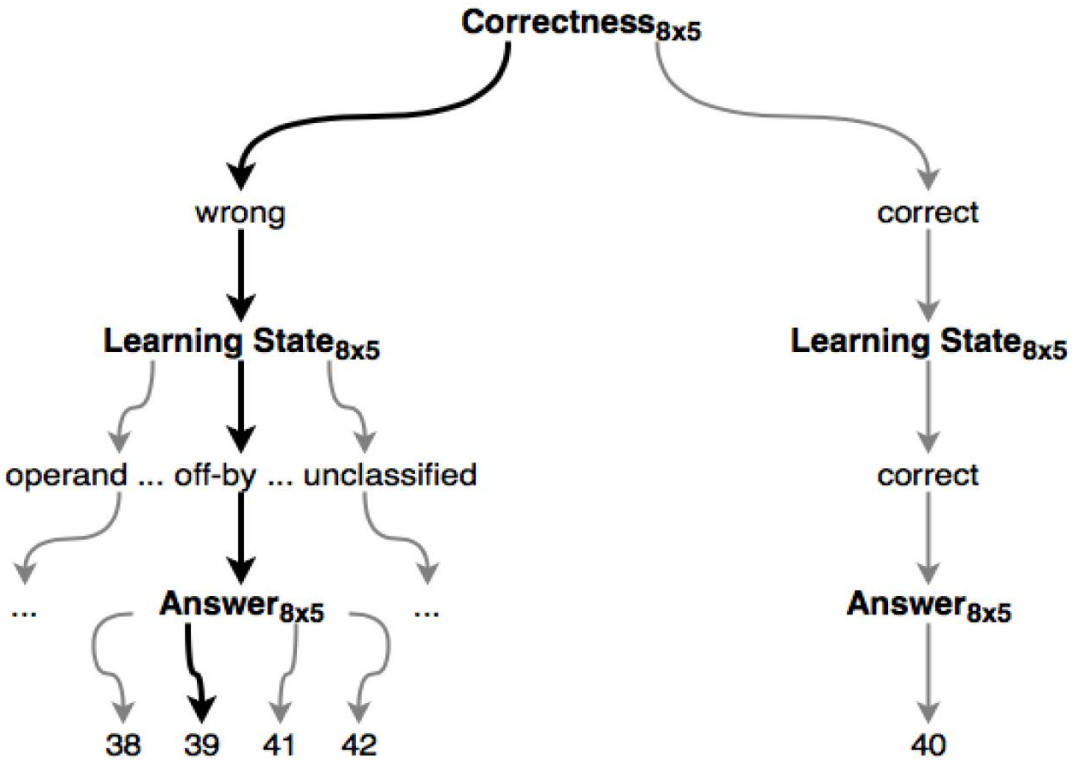
  You need to see at least one tails (**T**)
  to have sufficient statistics to say something
  about the bias of the coin

Sampling (1/2)

- Sample!
  in the example seen before, predict the answer of a particular question
- Forward sampling handles the RVs of the model in a topological order
- Firstly, it samples a value from $\mathbf{Correctness_q}$ and then uses the sampled answer c to choose the next probability distribution $P(\mathbf{Learning\ State_q}|\mathbf{Correctness_q} = c)$ (which is a row in the Conditional Probability Table of $\mathbf{Learning\ State_q}$) to sample from, and so on.

Sampling (2/2)

- Forward sampling path



Probabilistic Query / Inference (1/5)

- Causal reasoning: start with the knowledge of the causes as evidence and provide information about the effects
- Evidential reasoning (also called explanation): it has the opposite direction;
  it involves situations where effects lead to the specification of causes
- Considering the direction of time **sometimes** helps - evidential reasoning infers the past probability distribution from the current set of data whereas causal reasoning makes a prediction for the future given the data

Probabilistic Query / Inference (2/5)

- $P(\boldsymbol{Y}|\boldsymbol{E} = e) = \frac{P(\boldsymbol{Y},e)}{P(e)}$
  Computes the posterior of the subset of random variables represented by $\boldsymbol{Y}$ (target of query) given observations e of the subset of evidence variables denoted by $\boldsymbol{E}$
- $\mathrm{MAP}(\boldsymbol{Y}|\boldsymbol{E} = e) = \arg\max_{y} P(y, e)$
  The MAP query, which is also called most probable explanation (MPE), is a query that maximizes the posterior of the joint distribution of a subset of random variables $\boldsymbol{Y}$

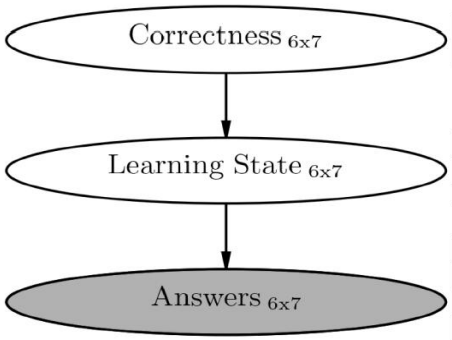
Probabilistic Query / Inference (3/5)

- Assume in the example that we need to compute

$$P(\mathbf{C_{6\times 7}}, \mathbf{LS_{6\times 7}}, \mathbf{A_{6\times 7}} = 40)$$

- The faulty answer 40 for the question 6×7 eliminates all cases for which the answer is not equal to 40;
  it can belong only to two potential error types: "operand" and "off-by"

Probabilistic Query / Inference (4/5)



wrong	correct
0.158	0.842

	operand	intrusion	consistency	off-by	add/sub/div	pattern	unclassified	correct
wrong	0.3367	0.0791	0.1638	0.1031	0.0014	0.0726	0.2433	0.0
correct	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

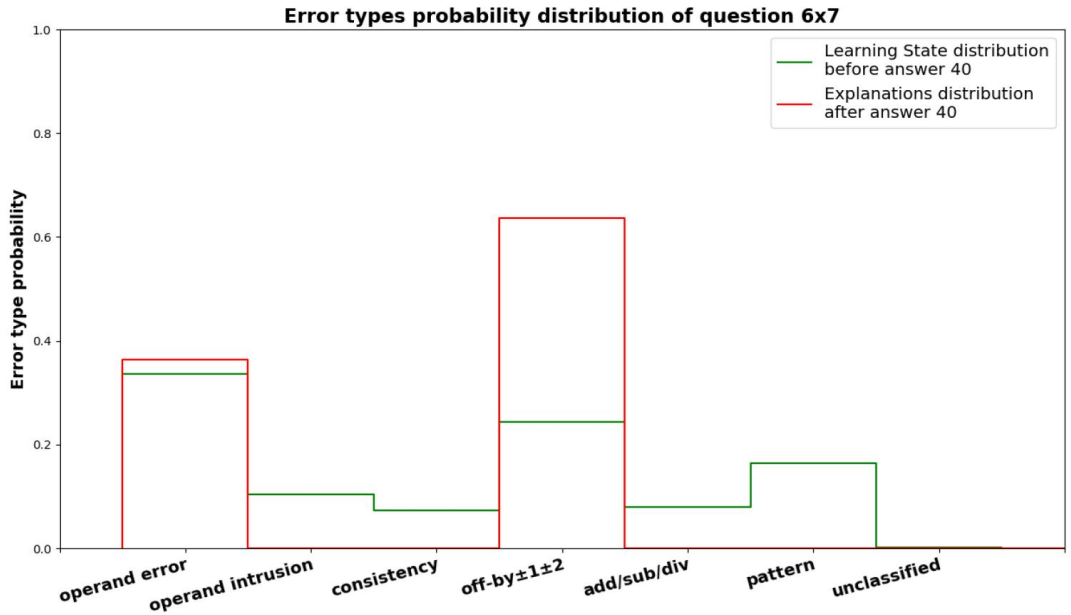
	...	40	...
...	...		...
operand	...	0.035	...
off-by	...	0.202	...
...	...		...

$C_{6\times7}$	$LS_{6\times7}$	$A_{6\times7}$	unnormalized proportions
wrong	**operand**	40	$0.158 \cdot 0.336 \cdot 0.035 = 1.85 \cdot 10^{-3}$
wrong	**off-by**	40	$0.158 \cdot 0.103 \cdot 0.202 = 3.28 \cdot 10^{-3}$

- Sum of the unnormalized proportions,
  $1.85 \cdot 10^{-3} + 3.28 \cdot 10^{-3} = 5.14 \cdot 10^{-3}$

Probabilistic Query / Inference (5/5)

$C_{6\times7}$	$LS_{6\times7}$	$A_{6\times7}$	normalized probabilities
wrong	**operand**	40	$1.85 \cdot 10^{-3} / 5.14 \cdot 10^{-3} = 0.36$
wrong	**off-by**	40	$3.28 \cdot 10^{-3} / 5.14 \cdot 10^{-3} = 0.64$



pgmpy

- Parameter Learning and Structure Learning:

  `https://pgmpy.org/api/parameter_estimation.html`
  `https://pgmpy.org/guides/causal_discovery.html`

Literature (1/2)

- Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.
- Barber, David. Bayesian reasoning and machine learning. Cambridge University Press, 2012.
  $\rightarrow$ contains excercises
- Pfeffer, Avi. Practical probabilistic programming. Simon and Schuster, 2016.
  $\rightarrow$ contains implementation in Java

Literature (2/2)

- Daphne Kollers PGM online class: `https://www.youtube.com/watch?v=GqMzbbaN6T4&list=PLzERW_Obpmv-_TkPEmCyzaJUGHtl7S01i`
- Glymour, Madelyn, Judea Pearl, and Nicholas P. Jewell. Causal inference in statistics: A primer. John Wiley & Sons, 2016.
  → contains the do-calculus