Discrete Geometry for Risk and AI

Dr. Paul Larsen

April 14, 2022

Why discrete geometry?

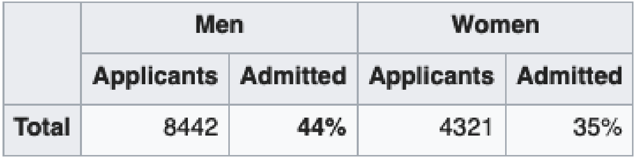
- Recent history: Dissatisfaction with deep learning, only “curve fitting”, alternatives via *causal graphical models* [Pea19]
- Less recent history: graphical models among first non-rules based AI approaches [Dar09]
- Geometrical formulations of statistical objects, e.g. graphical models and probability polytopes

For today . . .

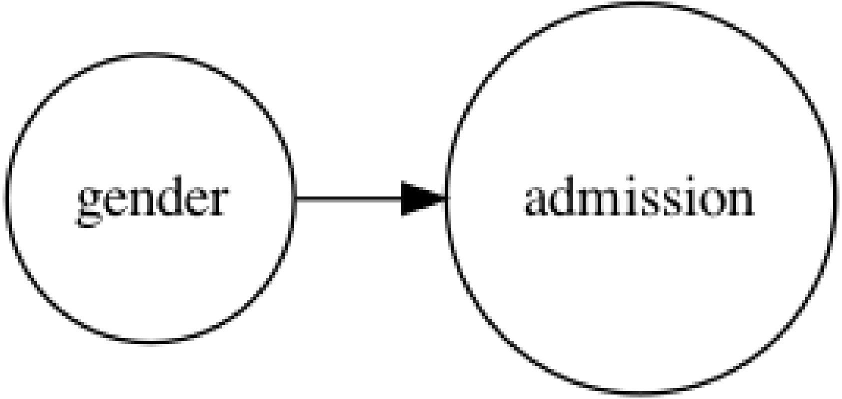
- Graphical models
- Probability polytopes
- Geometry of Simpson’s Paradox

Directed graphical model: university admission gender bias

Simpson paradox preview

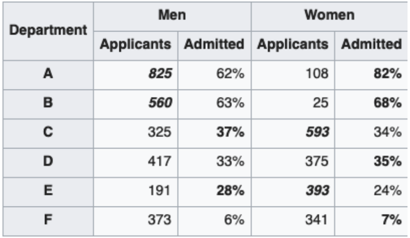


	Men		Women	
	Applicants	Admitted	Applicants	Admitted
Total	8442	44%	4321	35%

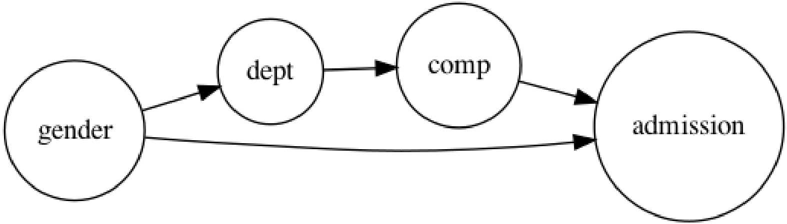


Directed graphical model: university admission gender bias

Simpson paradox preview



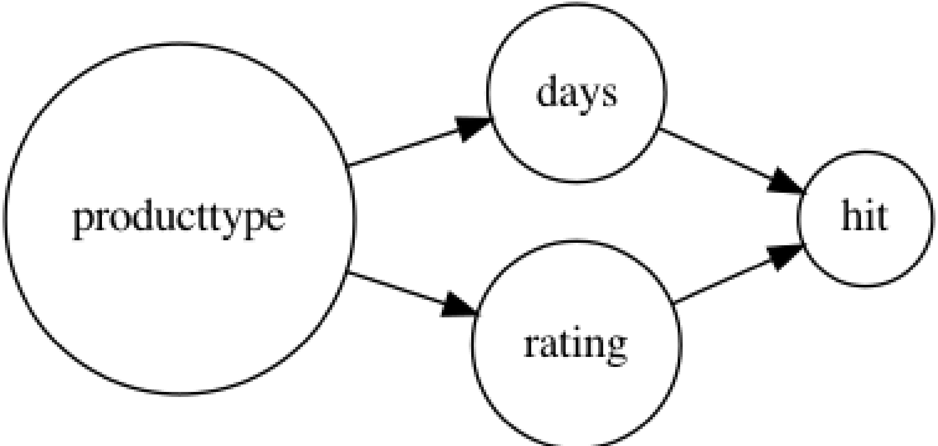
Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	***825***	62%	108	**82%**
B	***560***	63%	25	**68%**
C	325	**37%**	***593***	34%
D	417	33%	375	**35%**
E	191	**28%**	***393***	24%
F	373	6%	341	**7%**



Sources: [Wik] [BHO75]

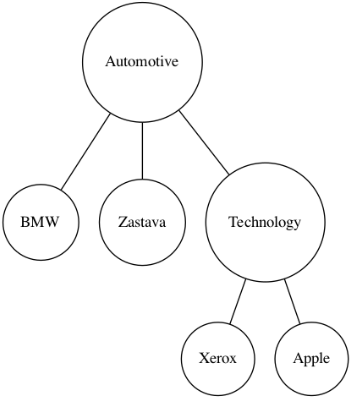
Directed graphical model: hit rate for insurance quotes

- product type: financial, liability, property
- days: number of days to generate quote
- rating: measure of premium paid expected claims
- hit: 0 if quote refused, 1 if accepted



Undirected graphical model: credit default risk [FGMS12]

- Nodes take values 0 (healthy) or 1 (default)
- Industry nodes connect to other industry nodes
- Individual firm nodes connect only to corresponding industry node



Graph definitions

Definition

A *graph* is a pair of sets (V, E), where V is called the set of *vertices* (or *nodes*) and E is called the set of *edges*, such that the set of edges corresponds injectively to pairs of vertices.

Notes

- Typically ‘pairs of vertices‘ does not include self-pairs, but this can be relaxed, leading to graphs with with loops.
- The injectivity requirement can also be relaxed, leading to *multigraphs*.

Graphical models

Definition

(Informal) A graphical model is a graph whose nodes represent variables and whose edges represent direct statistical dependencies between the variables.

Why graphical models?

- For probability distributions admitting a graphical model representation, then graph properties (*d-separation*) imply conditional independence relations.
- Conditional independence relations reduce the number of parameters required to specify a probability distribution.
- Graphical models come in two flavors depending on their edges: directed (aka *Bayesian Networks*) and undirected (aka *random Markov fields*).

Directed acyclic graphs

Definition

A graph $G = (V, E)$ is a *directed acyclic graph* (denoted also *DAG*) if all edges have an associated direction, and no edge path consistent with the directions forms a cycle.

If there is a directed path from X_i to X_j, then X_i is called a *parent* of X_j, and $Pa(X_j) \subseteq V$ is the set of all parents of X_j.

Definition

If $X = (X_1, \ldots, X_m)$ admits a DAG G, then X_G is a *DAG model* if the distribution of X decomposes according to G, i.e.

$$P(X) = \prod_{i \in \{1,\ldots,m\}} P(X_i | Pa(X_i))$$

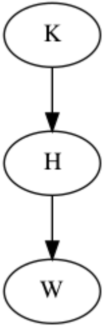
Example: Karma and weight-lifting

Take K to be your Karma, H to be the hours you spend in the gym lifting weight each day, and then W be the weight you can bench press on a given day. For simplicity, all random variables are binary.

karma	hours	weight
1	0	1
1	1	1
0	1	0
1	0	1
1	0	1

Decomposition example: Karma and weight-lifting

Suppose $X = (K, H, W)$ admits the graph



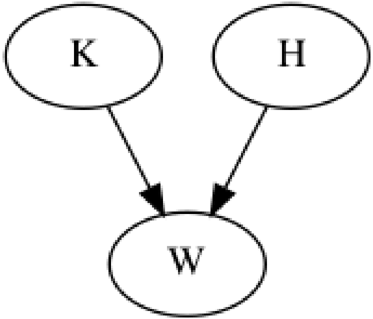
Then $P(K, H, W) = P(K)\,P(H|K)P(W|H)$.

Definition

A DAG of the form above is called a *chain*.

Decomposition example: Karma and weight-lifting

Suppose $X = (K, H, W)$ admits the graph



Then $P(K, H, W) = P(K)\, P(H)\, P(W|K, H)$.

Definition

A DAG of the form above is called a *collider* at W.

Conditional independence

Recall that two random variables X, Y are *independent* if, for all x, y, $P(X = x, Y = y) = P(X = x)P(Y = y)$.

Definition

Let $X = (X_1, \dots, X_m)$ be a probability distribution, and let A, B, C be pair-wise disjoint subsets of $1, \dots, m$, and define $X_A = (X_i)_{i \in A}$. Then X_A, X_B are *conditionally depenedent given* X_C if and only if

$$\begin{aligned} P(X_A = x_A, X_B = x_B | X_C = x_c) \\ = P(X_A = x_a | X_C = x_c) P(X_B = x_B | X_C = x_c) \end{aligned}$$

for all x_A, x_B, x_C.

For X_A, X_B conditionally independent given X_C, we write $(X_A \perp\!\!\!\perp X_B | X_C)$. See e.g. [DSS08] for a precise formulation.

Conditional independence and d-separation teaser

First example of discrete geometry helping statistics: conditional independence in a DAG model (X, G) can be detected in properties of G[1]. More precisely,

Theorem

If (X, G) is a DAG model, then d-separation implies conditional independence.

See e.g. [PGJ16], chapter 2.

---

[1] The required graph properties are combinatorial, but can also be understood geometrically, see e.g. [DSS08].

More definitions before d-separation

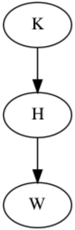


Figure: Chain

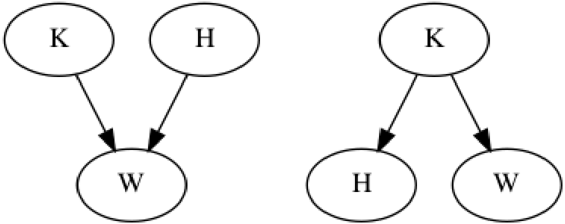


Figure: Collider at W, Fork at K

d-separation in DAGs

Definition

An undirected path p in a DAG G is *blocked* by a set of nodes S if and only if

1. p contains a chain of nodes $X \to Y \to Z$, or a fork $X \leftarrow Y \to Z$ such that $Y \in S$, or
2. p contains a collider $X \to Y \leftarrow Z$ such that $Y \notin S$ and no descendant of Y is in S.

Definition

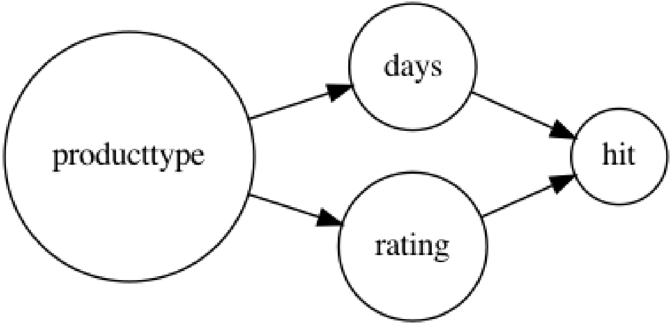
If a set of nodes S blocks every path between two nodes X and Y, then X and Y are called *d-separated conditional on* S, and we write

$$(X \perp\!\!\!\perp Y|S)_G$$

.

By the d-separation teaser theorem, $(X \perp\!\!\!\perp Y|S)_G$ implies conditional independence.

d-separation example: hit rate for insurance



All paths from $\mathrm{product\_type}$ to hit are blocked by $\{\mathrm{days}, \mathrm{rating}\}$, hence $(\mathrm{product\_type} \perp\!\!\!\perp \mathrm{hit} | \mathrm{days}, \mathrm{rating})_G$.

Probability polytopes

Goal: Use geometric interpretation of multivariate discrete random variables to generate interesting fake data with few(er) parameters.

Example: The family of all $X \sim$ *Bernoulli* can be represented as

$$\Delta_1 = \{(p_0, p_1) : p_i \geq 0, \sum p_i = 1\} \subseteq R^2$$

Example: Consider the collider graph for Karma-influenced weight-lifting (K, H, W). Then all possible conditional probability tables for $(W|K, H)$ can be parametrized as

$$\{(p_{w|k,h}) : p_{w|k,h} \geq 0, \sum_{w} p_{w|k,h} = 1 \text{ for } (k, h) \in \{0, 1\}^2\} \subseteq R^8$$

In general, the space of multivariate discrete random variable distributions is a *polytope*, see e.g. [DSS08], Ch. 1.

H- and V-representations of polytopes

Definition

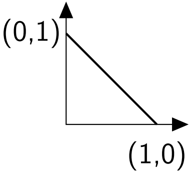
An *H-polyhedron* is an intersection of closed halfspaces, i.e. a set $P \subseteq R^d$ presented in the form

$$P = P(A, z) = \{x \in R^d : Ax \leq z\} \text{ for some } A \in R^{md}, z \in R^m.$$

If P is bounded (i.e. compact), then it is called a *polytope*.

Definition

(Informal) A *V-polytope* is the convex hull of a finite set of vertices $conv(V) \in R^d$. See [Zie12] for a precise definition.



Example: The V-representation for all *Bernoulli* distributions is

The main theorem of polytopes

Theorem

A subset $P \subset R^d$ is the convex hull of a finite point set (a V-polytope)

$$P = conv(V) \textit{ for some } V \in R^{dn}$$

if and only if it is a bounded intersection of halfspaces (an H-polytope)

$$P = P(A, z) \textit{ for some } A \in R^{md}, z \in R^m$$

See [Zie12] for a proof.

Applying the main theorem to conditional probability tables

For the Karma weight-lifting example, all conditional probability tables for $(W|K, H)$ that satisfy $E(W|K = 0) = 0$ (bad Karma, no weight) and $E(W|H = 0) = 0.2$ can be written as an $H - \mathit{polytope}$ as above with additional constraints

$$\sum_{w,h} w \, p_{w|0,h} = 0$$

$$\sum_{w,k} w \, p_{w|k,0} = 0.2$$

By converting this H-representation to a V-representation, we can generate random conditional probability tables subject to expectation constrains.

For an example, see the implementation of ProbabilityPolytope of https://munichpavel.github.io/fake-data-for-learning/.

Geometry of Simpson's Paradox

Motivating example and notation

From Primer on Causality by Pearl, Glymour, Jewell, table 1.1. is

Subpopulation	No Treatment	Treatment
Female	55 of 80 recover (69%)	192 of 263 recover (73%)
Male	234 of 270 recover (87%)	81 of 87 recover (93%)
Total	289 of 350 (83%))	273 of 350 (78%)

we consider the counts above as being derived from a space of counts along dimensions $(\mathrm{RECOVERED}, \mathrm{GENDER}, \mathrm{TREATED})$ of $\mathbb{N}^2 \times \mathbb{N}^2 \times \mathbb{N}^2$:

$$U = (u_{ijk})$$

where

$$u_{ijk} = \text{counts of } \mathrm{RECOVERED} = i, \mathrm{GENDER} = j, \mathrm{TREATED} = k$$

Geometry of Simpson's Parardox

Notation for counts, II

$$\begin{aligned}
u_{000} &= \text{Count of non-recovered females who received no treatment}\\
u_{100} &= \text{Count of recovered females who received no treatment}\\
u_{010} &= \text{Count of non-recovered males who received no treatment}\\
u_{110} &= \text{Count of recovered males who received no treatment}\\
u_{001} &= \text{Count of non-recovered females who received treatment}\\
u_{101} &= \text{Count of recovered females who received treatment}\\
u_{011} &= \text{Count of non-recovered males who received treatment}\\
u_{111} &= \text{Count of recovered males who received treatment}
\end{aligned}$$

Subpopulation	No Treatment	Treatment
Female	u_{100} of u_{+00} recover (ratio $\frac{u_{100}}{u_{+00}}$)	u_{101} of u_{+01} recover (ratio $\frac{n_{101}}{n_{+01}}$)
Male	u_{110} of u_{+10} recover (ratio $\frac{u_{110}}{u_{+10}}$)	u_{111} of u_{+11} recover (ratio $\frac{u_{111}}{u_{+11}}$)
Total	u_{1+0} of u_{++0} recover $\left(\frac{u_{1+1}}{u_{++0}}\right)$	n_{1+1} of u_{++1} recover (ratio $\frac{u_{1+1}}{u_{++1}}$)

Geometry of Simpson's Paradox

Converting counts u_{ijk} to probabilities p_{ijk} (exercise), have quadratic inequalities for this Simpson's paradox example:

$$\begin{aligned} p_{101}p_{+00} - p_{100}p_{+01} &> 0 \\ p_{111}p_{+10} - p_{110}p_{+11} &> 0 \\ p_{1+1}p_{++0} - p_{1+0}p_{++1} &< 0 \end{aligned}$$

References I

[BHO75] P. J. Bickel, E. A. Hammel, and J. W. O'Connell, *Sex Bias in Graduate Admissions: Data from Berkeley*, Science **187** (1975), no. 4175, 398–404.

[Dar09] Adnan Darwiche, *Modeling and reasoning with bayesian networks*, Cambridge university press, 2009.

[DSS08] Mathias Drton, Bernd Sturmfels, and Seth Sullivant, *Lectures on algebraic statistics*, vol. 39, Springer Science & Business Media, 2008.

[FGMS12] Ismail Onur Filiz, Xin Guo, Jason Morton, and Bernd Sturmfels, *Graphical models for correlated defaults*, Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics **22** (2012), no. 4, 621–644.

[FPP98] D. Freedman, R. Pisani, and R. Purves, *Statistics*, W.W. Norton, 1998.

[Pea19] Judea Pearl, *The limitations of opaque learning machines*, Possible Minds: Twenty-Five Ways of Looking at ai (2019), 13–19.

References II

[PGJ16] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell, *Causal inference in statistics: A primer*, John Wiley & Sons, 2016.

[Wik] Wikipedia, *Simpson's paradox*, https://en.wikipedia.org/wiki/Simpson's_paradox.

[Zie12] Günter M Ziegler, *Lectures on polytopes*, vol. 152, Springer Science & Business Media, 2012.