Royal Institute of Technology

MACHINE LEARNING 2 – DGM, CH 10

Lecture 2

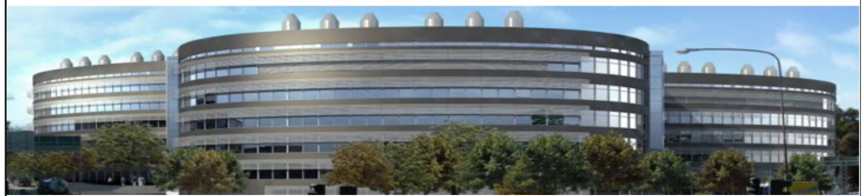
Royal Institute of Technology

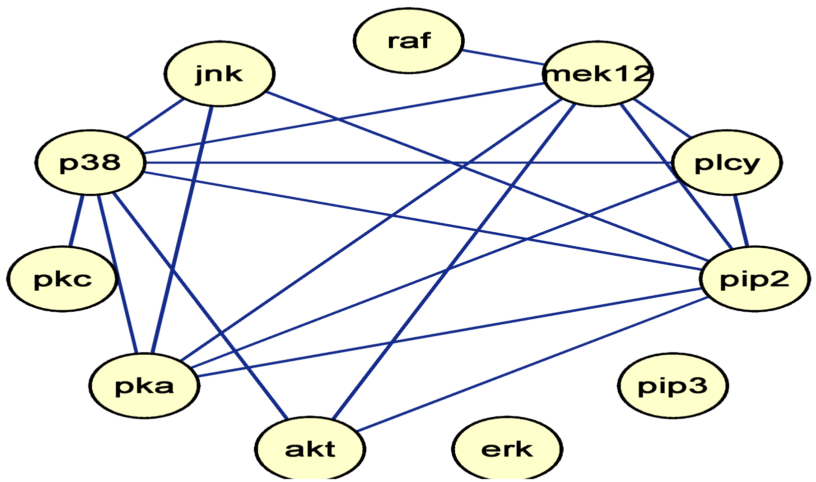
SciLifeLab

Computational Biology

Machine Learning – a main tool

Jens Lagergren





DISCOVERING GRAPH STRUCTURE

K-MEANS

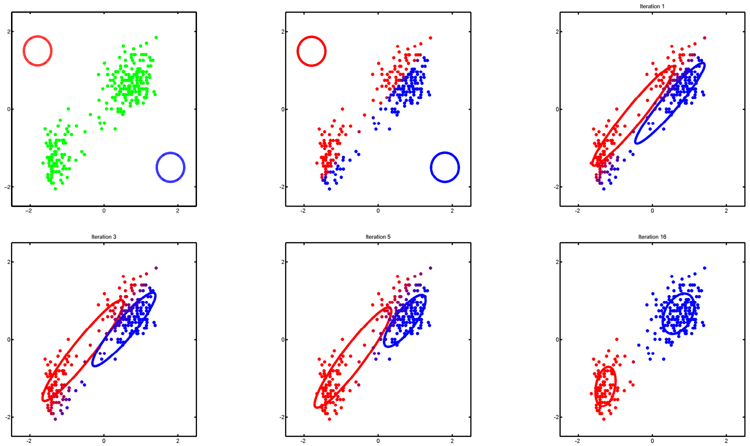
- Data vectors D={x1,…,xN}
- Randomly selected classes z1,…,zN
- Iteratively do

$$\boldsymbol{\mu}_c = \frac{1}{N_c} \sum_{n: z_n = c} \boldsymbol{x}_n, \qquad \text{where } N_c = |\{n : z_n = c\}|$$

$$z_n = \text{argmin}_c ||\boldsymbol{x}_n - \boldsymbol{\mu}_c||_2$$

- One step O(NKD), can be improved

EXAMPLE



Expected complete: notation

$$
\begin{aligned}
&\log p(\boldsymbol{x}_n|\boldsymbol{\theta}') \\
&\quad = \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}') \\
&\quad = \log \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \\
&\quad = \log E_{\boldsymbol{z}_n} \left(\frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \mid \boldsymbol{x}_n, \boldsymbol{\theta} \right) \\
&\quad \geq^{\text{Jensen}} E_{\boldsymbol{z}_n} \left(\log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \mid \boldsymbol{x}_n, \boldsymbol{\theta} \right) \\
&\quad = \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log \frac{p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}')}{p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta})} \\
&\quad = \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{x}_n, \boldsymbol{z}_n|\boldsymbol{\theta}') - \sum_{\boldsymbol{z}_n} p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \log p(\boldsymbol{z}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \\
&\quad = Q_n(\boldsymbol{\theta}'; \boldsymbol{\theta}) - R_n(\boldsymbol{\theta}; \boldsymbol{\theta})
\end{aligned}
$$

CONDITIONING

$$p(x, y) = p(y)p(x|y)$$ or $$p(x|y) = \frac{p(x, y)}{p(y)}$$

INFERENCE – THE CHAIN RULE

$$p(\underbrace{\boldsymbol{x}_{[V]}}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_V}) = p(\boldsymbol{x}_1)p(\boldsymbol{x}_2|\boldsymbol{x}_1)p(\boldsymbol{x}_3|\boldsymbol{x}_1, \boldsymbol{x}_2)\cdots p(\boldsymbol{x}_V|\boldsymbol{x}_{[V-1]})$$

- Assuming binary r.v., $p(X_V \mid X_{[V-1]})$ has 2^{V-1} parameters
- Total # parameters $\sum_{1 \leq i \leq V} 2^{i-1} = 2^V - 1$

CONDITIONAL INDEPENDENCE (MAY SAVE US)

★ X and Y are conditionally independent given Z iff

p(X, Y| Z)= P(X| Z) P(Y| Z)

★ Implies

p(X| Y, Z) = p(X, Y| Z)/p(Y| Z) = p(X| Z)

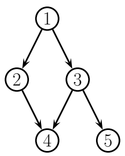
EX. WHERE IND. OBVIOUSLY FACILITATES

$$p(\underbrace{\boldsymbol{x}_{[V]}}_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_V}) = p(\boldsymbol{x}_1)p(\boldsymbol{x}_2|\boldsymbol{x}_1)p(\boldsymbol{x}_3|\boldsymbol{x}_1,\boldsymbol{x}_2)\cdots p(\boldsymbol{x}_V|\boldsymbol{x}_{[V-1]})$$

★Assume first order Markov property $\boldsymbol{x}_t \perp \boldsymbol{x}_{[t-2]}|\boldsymbol{x}_{t-1}$

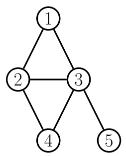
i.e., if time ordered, future independent of past given present

★Then $p(\boldsymbol{x}_{[V]}) = p(\boldsymbol{x}_1)\prod_{t=1}^{V-1} p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)$



Directed graphical model

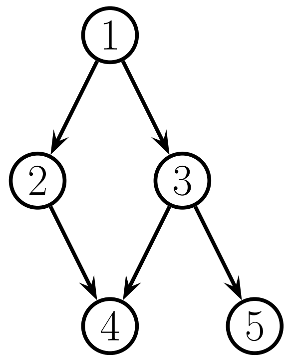
- DAG
- vertices r.v.s
- equipped with local CPDs



Undirected graphical model

- graph
- vertices r.v.s
- equipped with local “factors”

GRAPHICAL MODELS



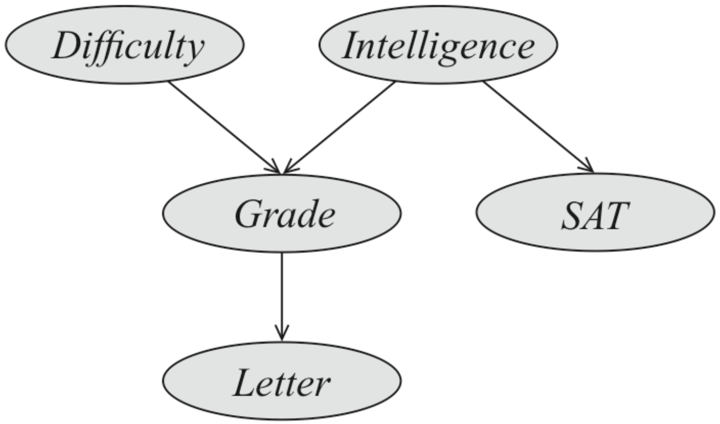
DGM

★What is the meaning of the underlying DAG? what is the semantics?

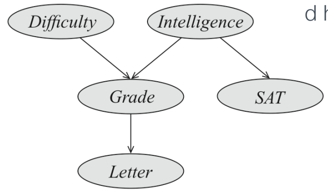
★What does a DGM mean? what is the semantics?

★Which DGMs represent a given distribution?

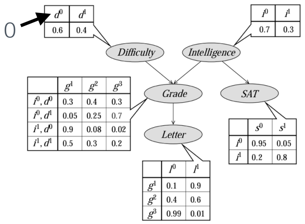
EXTENDED STUDENT EXAMPLE



RELATION DGM - DISTRIBUTIONS



d has value 0



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

BERNOULLI & BINOMIAL

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

$$\mathrm{Bin}(k|n,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

- ★ One or several (unordered) coin tosses

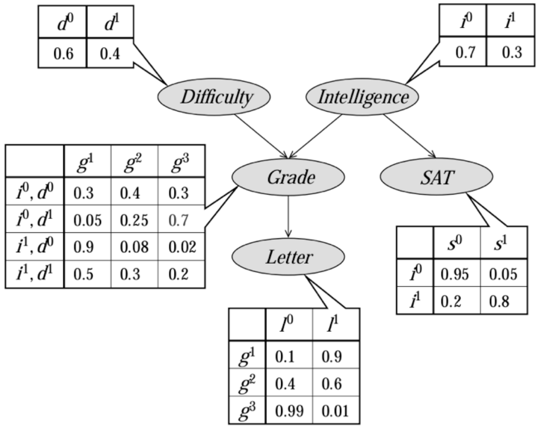
CATEGORICAL & MULTINOMIAL

$$\mathrm{Cat}(x|\boldsymbol{\theta}) = \theta_x$$

$$\mathrm{Mul}(\boldsymbol{x}|n,\boldsymbol{\theta}) = \binom{n}{x_1,\ldots,x_K}\prod_{k=1}^{K}\theta_k^{x_k}$$

- ★ One or several (unordered) coin tosses

EXTENDED STUDENT EXAMPLE



d^0	d^1
0.6	0.4

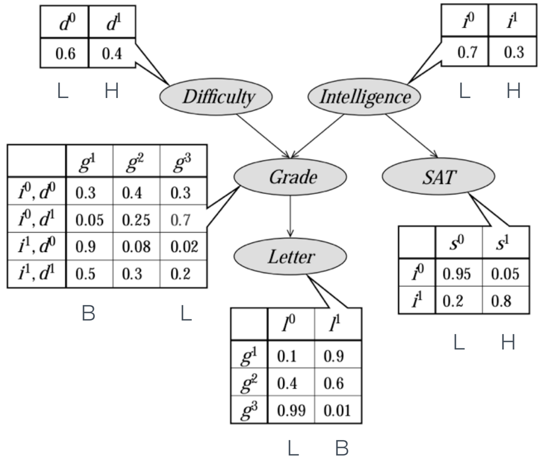
i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

EXTENDED STUDENT EXAMPLE



d^0	d^1
0.6	0.4

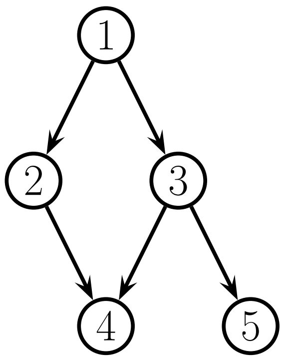
i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

B - better
H - higher
L - less



TERMINOLOGY

★ Parent

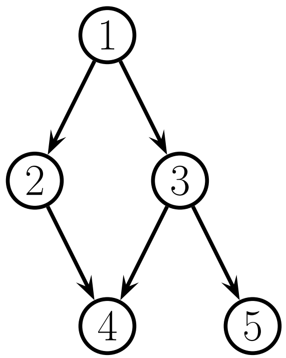
★ Child

★ Family

★ Root

★ Leaf

★Neighbors



TERMINOLOGY

★ Degree (in and out)

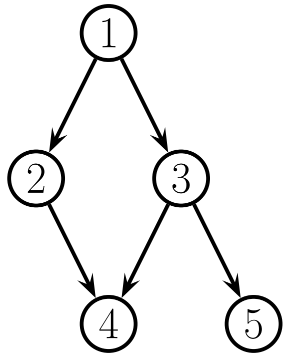
★ Cycle (directed or not)

★ Directed Acyclic Graph (DAG)

★ Topological order (parents < child)

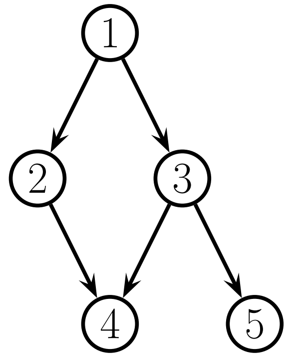
★ Path (directed or not)

★ Ancestors



TERMINOLOGY

- ★ Tree
- ★ Polytree – directed tree with multiple parents for some vertices
- ★ Forest
- ★ Subgraph
- ★ Clique
- ★ Maximal clique



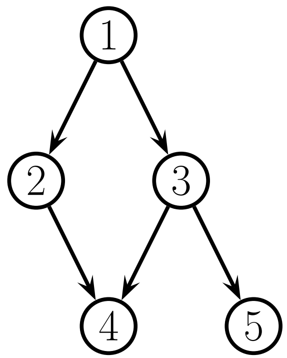
ORDERED MARKOV PROPERTY

★ The ordered Markov Property

$$\boldsymbol{x}_t \perp \boldsymbol{x}_{V \setminus \mathrm{desc}(t)} | \boldsymbol{x}_{\mathrm{pa}(t)}$$

★ In this case

$$\begin{aligned} p(\boldsymbol{x}_{[5]}) &= p(\boldsymbol{x}_1)p(\boldsymbol{x}_2|\boldsymbol{x}_1)p(\boldsymbol{x}_3|\boldsymbol{x}_1,\boldsymbol{x}_2) \\ &\quad p(\boldsymbol{x}_4|\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)p(\boldsymbol{x}_5|\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4) \\ &= p(\boldsymbol{x}_1)p(\boldsymbol{x}_2|\boldsymbol{x}_1)p(\boldsymbol{x}_3|\boldsymbol{x}_1) \\ &\quad p(\boldsymbol{x}_4|\boldsymbol{x}_2,\boldsymbol{x}_3)p(\boldsymbol{x}_5|\boldsymbol{x}_3) \end{aligned}$$



ORDERED MARKOV PROPERTY

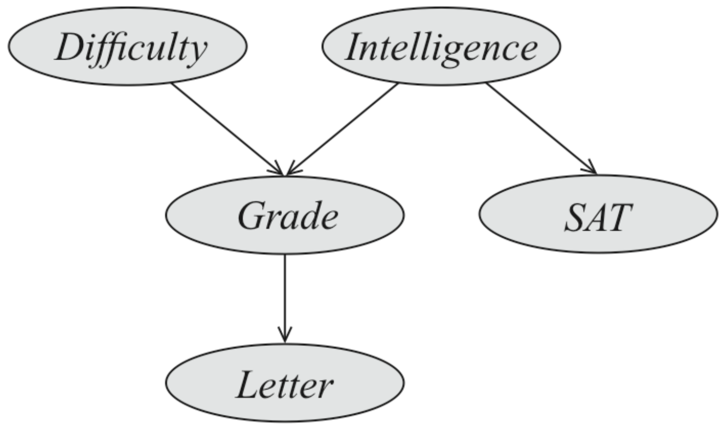
★ The ordered Markov Property

$$\boldsymbol{x}_t \perp \boldsymbol{x}_{V \setminus \mathrm{desc}(t)} | \boldsymbol{x}_{\mathrm{pa}(t)}$$

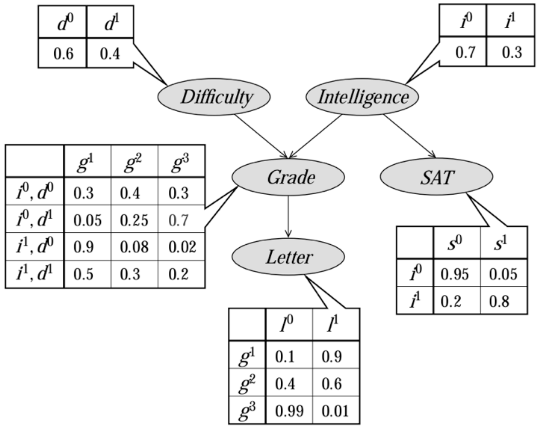
★In general, if 1,…,V topological order, the likelihood is decomposable (factorizes)

$$p(\boldsymbol{x}_{[V]}|G) = \prod_{t=1}^{V} p(\boldsymbol{x}_t|\boldsymbol{x}_{\mathrm{pa}(t)})$$

EXTENDED STUDENT EXAMPLE



EXTENDED STUDENT EXAMPLE



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

AN EXAMPLE

Difficulty (1 diff), Intelligence (1 int), Sat (1 good), Grade (1 good), Letter (1good)

D	I	S	G	L
0	1	1	1	1
1	1	1	0	0
1	1	0	0	1
1	0	0	0	0
1	1	0	0	1

$$\begin{aligned} L(\boldsymbol{\theta};\mathcal{D}) = {} & p(0,1,1,1,1|\boldsymbol{\theta})p(1,1,1,0,0|\boldsymbol{\theta}) \\ & p(1,1,0,0,1|\boldsymbol{\theta})p(1,0,0,0,0|\boldsymbol{\theta}) \\ & p(1,1,0,0,1|\boldsymbol{\theta}) \end{aligned}$$

AN EXAMPLE

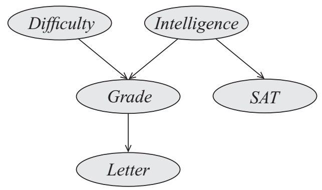
Difficulty (1 diff), Intelligence (1 int), Sat (1 good), Grade (1 good), Letter (1good)

D	I	S	G	L
0	1	1	1	1
1	1	1	0	0
1	1	0	0	1
1	0	0	0	0
1	1	0	0	1

$$\begin{aligned} L(\boldsymbol{\theta};\mathcal{D}) = {} & p(0,1,1,1,1|\boldsymbol{\theta})p(1,1,1,0,0|\boldsymbol{\theta}) \\ & p(1,1,0,0,1|\boldsymbol{\theta})p(1,0,0,0,0|\boldsymbol{\theta}) \\ & p(1,1,0,0,1|\boldsymbol{\theta}) \\ = {} & p(D=(0,1,1,1,1)|\boldsymbol{\theta}_D)\ldots \end{aligned}$$

AN EXAMPLE

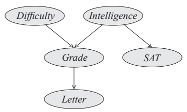
Difficulty (1 diff), Intelligence (1 int), Sat (1 good), Grade (1 good), Letter (1good)



D	I	S	G	L
0	1	1	1	1
1	1	1	0	0
1	1	0	0	1
1	0	0	0	0
1	1	0	0	1

$$\begin{aligned} L(\boldsymbol{\theta};\mathcal{D}) = {} & p(D=(0,1,1,1,1)|\boldsymbol{\theta}_D) \\ & p(I=(1,1,1,0,1)|\boldsymbol{\theta}_I) \\ & p(S=(1,1,0,0,0)|I=(1,1,1,0,1),\boldsymbol{\theta}_S) \\ & p(G=(1,0,0,0,0)|D=(0,1,1,1,1),I=(1,1,1,0,1),\boldsymbol{\theta}_G) \\ & p(L=(1,0,1,0,1)|G=(1,0,0,0,0),\boldsymbol{\theta}_L) \end{aligned}$$

AN EXAMPLE



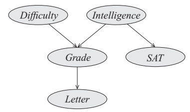
Difficulty (1 diff), Intelligence (1 int), Sat (1 good), Grade (1 good), Letter (1good)

θD	D=0	D=1
	2/5	3/5

D	I	S	G	L
0	1	1	1	1
1	1	1	0	0
1	1	0	0	1
1	0	0	0	0
1	1	0	0	1

$$L(\boldsymbol{\theta};\mathcal{D}) = p(D=(0,1,1,1,1)|\boldsymbol{\theta}_D)$$
$$p(I=(1,1,1,0,1)|\boldsymbol{\theta}_I)$$
$$p(S=(1,1,0,0,0)|I=(1,1,1,0,1),\boldsymbol{\theta}_S)$$
$$p(G=(1,0,0,0,0)|D=(0,1,1,1,1),I=(1,1,1,0,1),\boldsymbol{\theta}_G)$$
$$p(L=(1,0,1,0,1)|G=(1,0,0,0,0),\boldsymbol{\theta}_L)$$

AN EXAMPLE



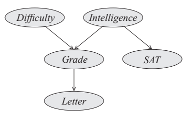
Difficulty (1 diff), Intelligence (1 int), Sat (1 good), Grade (1 good), Letter (1good)

θD	D=0	D=1
	2/5	3/5

D	I	S	G	L
0	1	1	1	1
1	1	1	0	0
1	1	0	0	1
1	0	0	0	0
1	1	0	0	1

$$L(\boldsymbol{\theta};\mathcal{D}) = \frac{2}{5}\left(\frac{3}{5}\right)^4$$
$$p(I=(1,1,1,0,1)|\boldsymbol{\theta}_I)$$
$$p(S=(1,1,0,0,0)|I=(1,1,1,0,1),\boldsymbol{\theta}_S)$$
$$p(G=(1,0,0,0,0)|D=(0,1,1,1,1),I=(1,1,1,0,1),\boldsymbol{\theta}_G)$$
$$p(L=(1,0,1,0,1)|G=(1,0,0,0,0),\boldsymbol{\theta}_L)$$

AN EXAMPLE



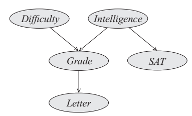
Difficulty (1 diff), Intelligence (1 int), Sat (1 good), Grade (1 good), Letter (1good)

θI	I=0	I=1
	1/4	3/4

D	I	S	G	L
0	1	1	1	1
1	1	1	0	0
1	1	0	0	1
1	0	0	0	0
1	1	0	0	1

$$L(\boldsymbol{\theta};\mathcal{D}) = \frac{2}{5}\left(\frac{3}{5}\right)^4$$
$$p(I=(1,1,1,0,1)|\boldsymbol{\theta}_I)$$
$$p(S=(1,1,0,0,0)|I=(1,1,1,0,1),\boldsymbol{\theta}_S)$$
$$p(G=(1,0,0,0,0)|D=(0,1,1,1,1),I=(1,1,1,0,1),\boldsymbol{\theta}_G)$$
$$p(L=(1,0,1,0,1)|G=(1,0,0,0,0),\boldsymbol{\theta}_L)$$

AN EXAMPLE



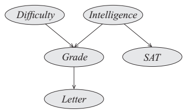
Difficulty (1 diff), Intelligence (1 int), Sat (1 good), Grade (1 good), Letter (1good)

θI	I=0	I=1
	1/4	3/4

D	I	S	G	L
0	1	1	1	1
1	1	1	0	0
1	1	0	0	1
1	0	0	0	0
1	1	0	0	1

$$L(\boldsymbol{\theta};\mathcal{D}) = \frac{2}{5}\left(\frac{3}{5}\right)^4 \frac{1}{4}\left(\frac{3}{4}\right)^4$$
$$p(S=(1,1,0,0,0)|I=(1,1,1,0,1),\boldsymbol{\theta}_S)$$
$$p(G=(1,0,0,0,0)|D=(0,1,1,1,1),I=(1,1,1,0,1),\boldsymbol{\theta}_G)$$
$$p(L=(1,0,1,0,1)|G=(1,0,0,0,0),\boldsymbol{\theta}_L)$$

AN EXAMPLE



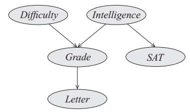
Difficulty (1 diff), Intelligence (1 int), Sat (1 good), Grade (1 good), Letter (1good)

θ_S	S=0	S=1
I=0	1	0
I=1	1/6	5/6

D	I	S	G	L
0	1	1	1	1
1	1	1	0	0
1	1	0	0	1
1	0	0	0	0
1	1	0	0	1

$$L(\boldsymbol{\theta};\mathcal{D}) = \frac{2}{5}\left(\frac{3}{5}\right)^4 \frac{1}{4}\left(\frac{3}{4}\right)^4$$
$$p(S=(1,1,0,0,0)|I=(1,1,1,0,1),\boldsymbol{\theta}_S)$$
$$p(G=(1,0,0,0,0)|D=(0,1,1,1,1),I=(1,1,1,0,1),\boldsymbol{\theta}_G)$$
$$p(L=(1,0,1,0,1)|G=(1,0,0,0,0),\boldsymbol{\theta}_L)$$

AN EXAMPLE



Difficulty (1 diff), Intelligence (1 int), Sat (1 good), Grade (1 good), Letter (1good)

θ_S	S=0	S=1
I=0	1	0
I=1	1/6	5/6

D	I	S	G	L
0	1	1	1	1
1	1	1	0	0
1	1	0	0	1
1	0	0	0	0
1	1	0	0	1

$$L(\boldsymbol{\theta};\mathcal{D}) = \frac{2}{5}\left(\frac{3}{5}\right)^4 \frac{1}{4}\left(\frac{3}{4}\right)^4 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2$$
$$p(G=(1,0,0,0,0)|D=(0,1,1,1,1),I=(1,1,1,0,1),\boldsymbol{\theta}_G)$$
$$p(L=(1,0,1,0,1)|G=(1,0,0,0,0),\boldsymbol{\theta}_L)$$

AN EXAMPLE

Difficulty (1 diff), Intelligence (1 int), Sat (1 good), Grade (1 good), Letter (1good)

	Less	Better
θ_G	G=0	G=1
D=0, I=0	1/2	1/2
D=1, I=0	3/5	2/5
D=0, I=1	1/10	9/10
D=1, I=1	2/5	3/5

D	I	S	G	L
0	1	1	1	1
1	1	1	0	0
1	1	0	0	1
1	0	0	0	0
1	1	0	0	1

$$L(\boldsymbol{\theta};\mathcal{D}) = \frac{2}{5}\left(\frac{3}{5}\right)^4 \frac{1}{4}\left(\frac{3}{4}\right)^4 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2$$
$$p(G=(1,0,0,0,0)|D=(0,1,1,1,1),I=(1,1,1,0,1),\boldsymbol{\theta}_G)$$
$$p(L=(1,0,1,0,1)|G=(1,0,0,0,0),\boldsymbol{\theta}_L)$$

AN EXAMPLE

Difficulty (1 diff), Intelligence (1 int), Sat (1 good), Grade (1 good), Letter (1good)

θ_G	G=0	G=1
D=0, I=0	1/2	1/2
D=1, I=0	3/5	2/5
D=0, I=1	1/10	9/10
D=1, I=1	2/5	3/5

D	I	S	G	L
0	1	1	1	1
1	1	1	0	0
1	1	0	0	1
1	0	0	0	0
1	1	0	0	1

$$L(\boldsymbol{\theta};\mathcal{D}) = \frac{2}{5}\left(\frac{3}{5}\right)^4 \frac{1}{4}\left(\frac{3}{4}\right)^4 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2 \frac{9}{10}\left(\frac{2}{5}\right)^3 \frac{3}{5}$$
$$p(L=(1,0,1,0,1)|G=(1,0,0,0,0),\boldsymbol{\theta}_L)$$

AN EXAMPLE

Difficulty (1 diff), Intelligence (1 int), Sat (1 good), Grade (1 good), Letter (1good)

θ_L	L=0	L=1
G=0	2/3	1/3
G=1	0	1

D	I	S	G	L
0	1	1	1	1
1	1	1	0	0
1	1	0	0	1
1	0	0	0	0
1	1	0	0	1

$$L(\boldsymbol{\theta}; \mathcal{D}) = \frac{2}{5}\left(\frac{3}{5}\right)^4 \frac{1}{4}\left(\frac{3}{4}\right)^4 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2 \frac{9}{10}\left(\frac{2}{5}\right)^3 \frac{3}{5}$$
$$p(L = (1,0,1,0,1) | G = (1,0,0,0,0), \boldsymbol{\theta}_L)$$

AN EXAMPLE

Difficulty (1 diff), Intelligence (1 int), Sat (1 good), Grade (1 good), Letter (1good)

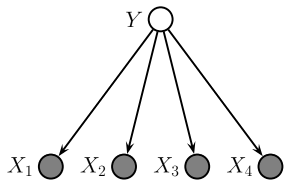
θ_L	L=0	L=1
G=0	2/3	1/3
G=1	0	1

D	I	S	G	L
0	1	1	1	1
1	1	1	0	0
1	1	0	0	1
1	0	0	0	0
1	1	0	0	1

$$L(\boldsymbol{\theta}; \mathcal{D}) = \frac{2}{5}\left(\frac{3}{5}\right)^4 \frac{1}{4}\left(\frac{3}{4}\right)^4 \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^2 \frac{9}{10}\left(\frac{2}{5}\right)^3 \frac{3}{5}\left(\frac{2}{3}\right)^2 \left(\frac{1}{3}\right)^2$$

THE LIKELIHOOD FACTORIZES

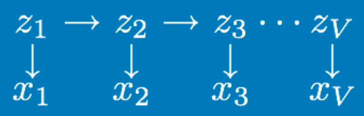
- ★ Complete data
$$\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$$
$$\boldsymbol{x}_n = \{\boldsymbol{x}_{n1}, \ldots, \boldsymbol{x}_{nV}\}$$
- ★ Likelihood
$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{\theta}) = \prod_{n=1}^{N}\prod_{v=1}^{V} p(\boldsymbol{x}_{nv}|\boldsymbol{x}_{n,\text{pa}(v)}, \boldsymbol{\theta})$$
$$= \prod_{v=1}^{V}\prod_{n=1}^{N} p(\boldsymbol{x}_{nv}|\boldsymbol{x}_{n,\text{pa}(v)}, \boldsymbol{\theta}) = \prod_{v=1}^{V} p(\mathcal{D}_v|\boldsymbol{\theta}_v)$$
where D_v is values of v together with its parents and θ_v is v's CPD
- ★ Called: decomposable likelihood (factorizes into family-factors)



SPECIAL CASE: NAIVE BAYES CLASSIFIER

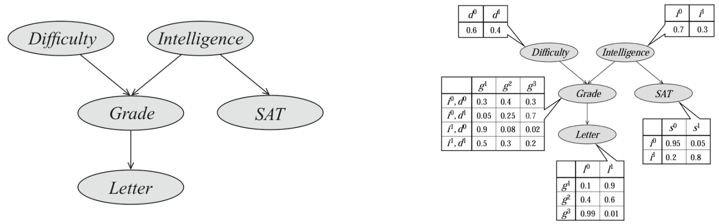
$$p(\boldsymbol{x}, y) = p(y)\prod_{t=1}^{4} p(x_t|y)$$

SPECIAL CASE: LAYERED HIDDEN MARKOV MODEL (HMM)



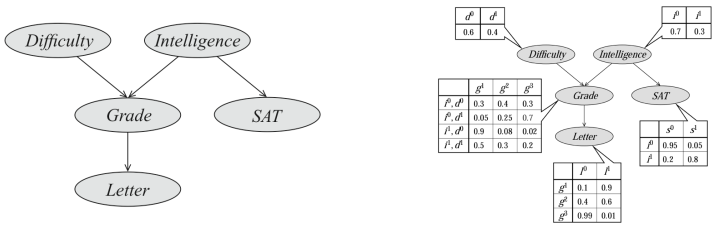
- Z_i hidden
- X_i observable
- Hidden often not observable when training, never when applying

THREE LEVELS OF COMPUTATIONAL PROBLEMS



- Inference: given G and θ, compute probabilities or marginalize
- Parameter learning: given G and D, learn θ
- Structure learning: given D, learn G and θ

THREE LEVELS

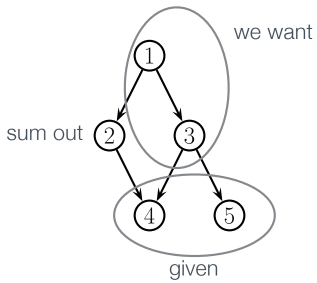


- Inference: given G and θ, compute probabilities or marginalize — Marginalize often hard
- Parameter learning: given G and D, learn θ — Easy for observable data
- Structure learning: given D, learn G and θ — Hard unless trees, doable in practice for observable

MARGINALIZE

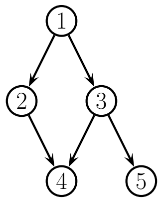
X, X'	two hidden variables
X_h	the other hidden variables
X_v	the visible variables

$$p(X=k, X'=k'|\boldsymbol{x}_v, \boldsymbol{\theta}) = \frac{p(X=k, X'=k', \boldsymbol{x}_v|\boldsymbol{\theta})}{P(\boldsymbol{x}_v|\boldsymbol{\theta})} = \frac{\sum_{x_h} p(X=k, X'=k', \boldsymbol{x}_h, \boldsymbol{x}_v|\boldsymbol{\theta})}{\sum_{x,x',x_h} p(x, x', \boldsymbol{x}_h, \boldsymbol{x}_v|\boldsymbol{\theta})}$$



MARGINALIZE

$$p(\boldsymbol{x}_h|\boldsymbol{x}_v,\boldsymbol{\theta}) = \frac{p(\boldsymbol{x}_h,\boldsymbol{x}_v|\boldsymbol{\theta})}{P(\boldsymbol{x}_v|\boldsymbol{\theta})} = \frac{p(\boldsymbol{x}_h,\boldsymbol{x}_v|\boldsymbol{\theta})}{\sum_{\boldsymbol{x}_h} p(\boldsymbol{x}_h,\boldsymbol{x}_v|\boldsymbol{\theta}}$$



- The denominator contains a marginal likelihood
- Summing out V binary hidden variables – $O(2^V)$
- K states – $O(K^V)$

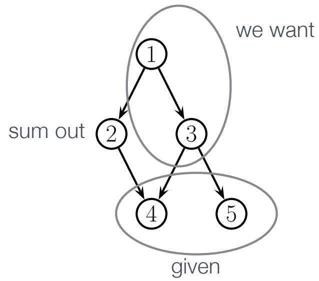
EXPECTED SUFFICIENT STATISTICS - ESS

X, X' two hidden variables

X_h the other hidden variables

X_v the visible variables

$$E[N_{k,k'}] = \sum_{\boldsymbol{x}\in\mathcal{D}} p(X=k, X'=k'|\boldsymbol{x}_v,\boldsymbol{\theta})$$



LEARNING PARAMETERS (IN GENERAL)

- "...Bayesian view, the parameters are unknown variables and should also be inferred"
- Learning from complete data $\mathcal{D} = \{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N\}$, $\boldsymbol{x}_n = \{\boldsymbol{x}_{n1},\ldots,\boldsymbol{x}_{nV}\}$
- Likelihood

$$P(\mathcal{D}|\boldsymbol{\theta}) = \prod_{n=1}^{N} P(\boldsymbol{x}_n|\boldsymbol{\theta}) = \prod_{n=1}^{N}\prod_{v=1}^{V} P(\boldsymbol{x}_{nv}|\boldsymbol{x}_{n,\mathrm{pa}(v)},\boldsymbol{\theta})$$
$$= \prod_{v=1}^{V}\prod_{n=1}^{N} P(\boldsymbol{x}_{nv}|\boldsymbol{x}_{n,\mathrm{pa}(v)},\boldsymbol{\theta}) = \prod_{v=1}^{V} P(\mathcal{D}_v|\boldsymbol{\theta}_v)$$

where D_v is values of v together with its parents and θ_v is v's CPD

CAT – NOTATION

- For a v ∊ [V],

  values $k \in [K_v]$

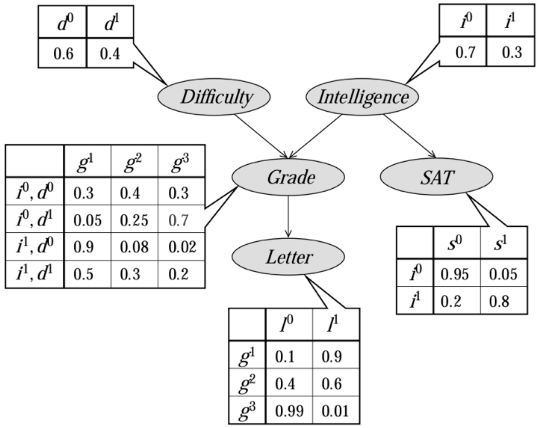
  combined values $c \in C_v = \prod_{s\in\mathrm{pa}(v)} [K_s]$ ← Cartesian product

- Cat CPDs

  where $P(x_v|x_{\mathrm{pa}(v)} = c) = \mathrm{Cat}(\boldsymbol{\theta}_{vc})$

  and $\boldsymbol{\theta}_{vck} = P(x_v = k|x_{\mathrm{pa}(v)} = c)$

EXTENDED STUDENT EXAMPLE



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

MLE FOR CAT CPDS

- Each $P(\mathcal{D}_v|\boldsymbol{\theta}_v)$, i.e., here each $\boldsymbol{\theta}_{v\mathbf{c}}$

  can be maximized independently
- So, MLE is

  $$\boldsymbol{\theta}_{v\mathbf{c}k} = N_{v\mathbf{c}k}/N_{v\mathbf{c}}$$
- where

  $$N_{v\mathbf{c}k} = \sum_{n=1}^{N} I(x_{nv} = k, x_{n,\mathrm{pa}(v)} = \mathbf{c})$$

  $$N_{v\mathbf{c}} = \sum_{n=1}^{N} I(x_{n,\mathrm{pa}(v)} = \mathbf{c})$$

c

THE END