Machine Learning

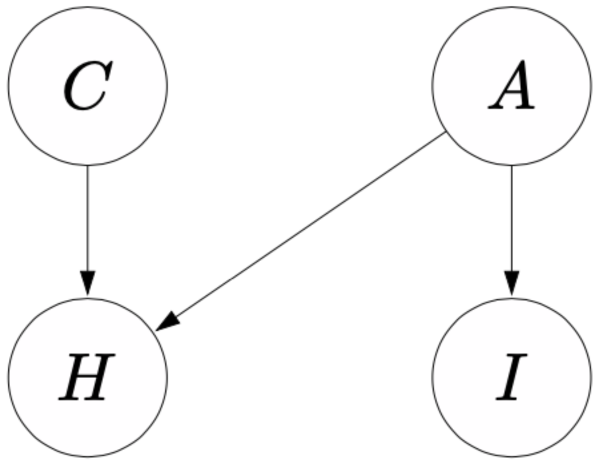
Lecture 8: Learning in BN

Sibei Yang

SIST, ShanghaiTech

Email: yangsb@shanghaitech.edu.cn

Review: Bayesian Network



$$\mathbb{P}(C = c, A = a, H = h, I = i)$$
$$= p(c)p(a)p(h \mid c, a)p(i \mid a)$$

Let $X = (X_1, \ldots, X_n)$ be random variables.

A **Bayesian network** is a directed acyclic graph (DAG) that specifies a joint distribution over X as a product of local conditional distributions, one for each node:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(x_i \mid x_{\text{Parents}(i)})$$

Review: Probabilistic Inference

Bayesian network:

$$\mathbb{P}(X = x) = \prod_{i=1}^{n} p(x_i \mid x_{\text{Parents}(i)})$$

Probabilistic inference:

$$\mathbb{P}(Q \mid E = e)$$

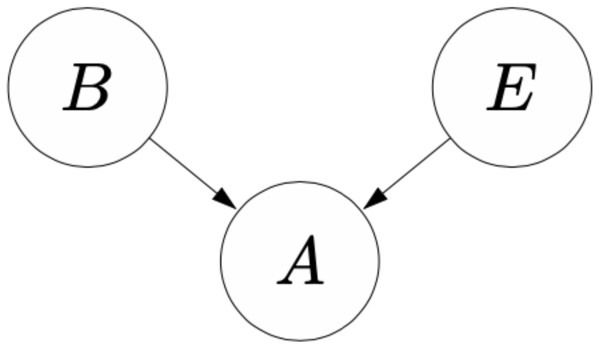
Algorithms:

- Variable elimination: general, exact
- Forward-backward: HMMs, exact
- Gibbs sampling, particle filtering: general, approximate

Outline

- Supervised Learning
- Laplace smoothing
- Unsupervised learning with EM

Learning: Where do parameters come from?



b	$p(b)$
1	?
0	?

e	$p(e)$
1	?
0	?

b	e	a	$p(a \mid b, e)$
0	0	0	?
0	0	1	?
0	1	0	?
0	1	1	?
1	0	0	?
1	0	1	?
1	1	0	?
1	1	1	?

1. Supervised Learning

Training data

$\mathcal{D}_{\text{train}}$ (an example is an assignment to X)

Parameters

θ (local conditional probabilities)

Example: one variable

Setup:

- One variable R representing the rating of a movie $\{1, 2, 3, 4, 5\}$

$$R \qquad \mathbb{P}(R = r) = p(r)$$

Parameters:

$$\theta = (p(1), p(2), p(3), p(4), p(5))$$

Training data:

$$\mathcal{D}_{\text{train}} = \{1, 3, 4, 4, 4, 4, 4, 5, 5, 5\}$$

Example: one variable

Learning:

$$\mathcal{D}_{\text{train}} \quad \Rightarrow \quad \theta$$

Intuition: $p(r) \propto$ number of occurrences of r in $\mathcal{D}_{\text{train}}$

Example:

$$\mathcal{D}_{\text{train}} = \{1, 3, 4, 4, 4, 4, 4, 5, 5, 5\}$$

$$\Downarrow$$

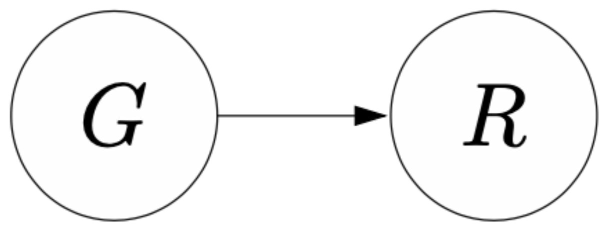
θ:

r	$p(r)$
1	0.1
2	0.0
3	0.1
4	0.5
5	0.3

Example: two variables

Variables:

- Genre $G \in \{\text{drama}, \text{comedy}\}$
- Rating $R \in \{1, 2, 3, 4, 5\}$

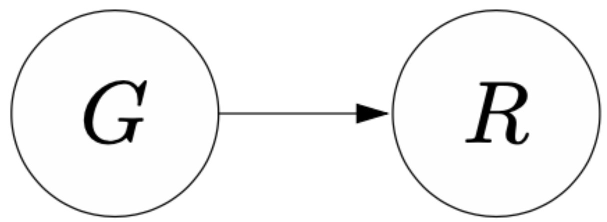


$$\mathbb{P}(G = g, R = r) = p_G(g) p_R(r \mid g)$$

$$\mathcal{D}_{\text{train}} = \{(\text{d}, 4), (\text{d}, 4), (\text{d}, 5), (\text{c}, 1), (\text{c}, 5)\}$$

Parameters: $\theta = (p_G, p_R)$

Example: two variables



$$\mathbb{P}(G = g, R = r) = p_G(g) p_R(r \mid g)$$

$$\mathcal{D}_{\text{train}} = \{(\mathsf{d}, 4), (\mathsf{d}, 4), (\mathsf{d}, 5), (\mathsf{c}, 1), (\mathsf{c}, 5)\}$$

Intuitive strategy: Estimate each local conditional distribution (p_G and p_R) separately

θ:

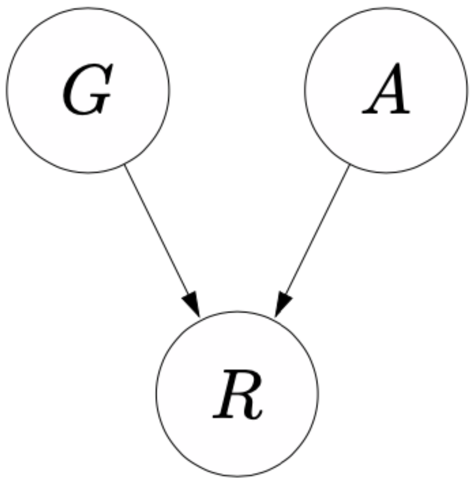
g	$p_G(g)$
d	3/5
c	2/5

g	r	$p_R(r \mid g)$
d	4	2/3
d	5	1/3
c	1	1/2
c	5	1/2

Example: v-structure

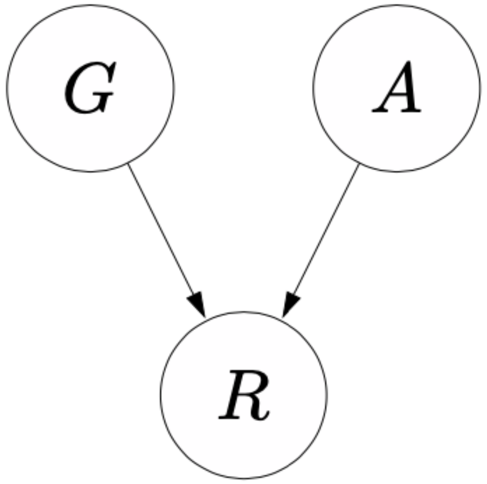
Variables:

- Genre $G \in \{\text{drama}, \text{comedy}\}$
- Won award $A \in \{0, 1\}$
- Rating $R \in \{1, 2, 3, 4, 5\}$



$$\mathbb{P}(G = g, A = a, R = r) = p_G(g) p_A(a) p_R(r \mid g, a)$$

Example: v-structure



$$\mathcal{D}_{\text{train}} = \{(\mathsf{d}, 0, 3), (\mathsf{d}, 1, 5), (\mathsf{c}, 0, 1), (\mathsf{c}, 0, 5), (\mathsf{c}, 1, 4)\}$$

Parameters: $\theta = (p_G, p_A, p_R)$

θ:

g	$p_G(g)$
d	3/5
c	2/5

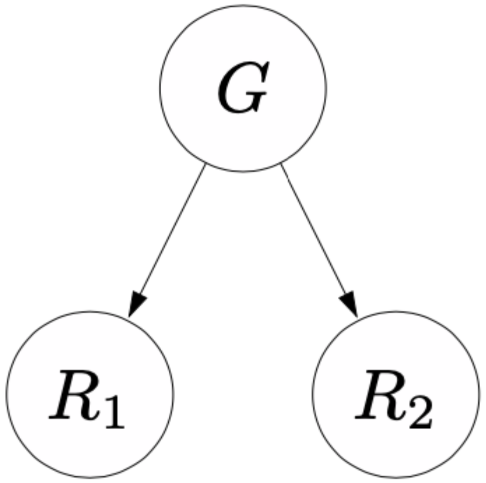
a	$p_A(a)$
0	3/5
1	2/5

g	a	r	$p_R(r \mid g, a)$
d	0	3	1
d	1	5	1
c	0	1	1/2
c	0	5	1/2
c	1	5	1

Example: inverted-v structure

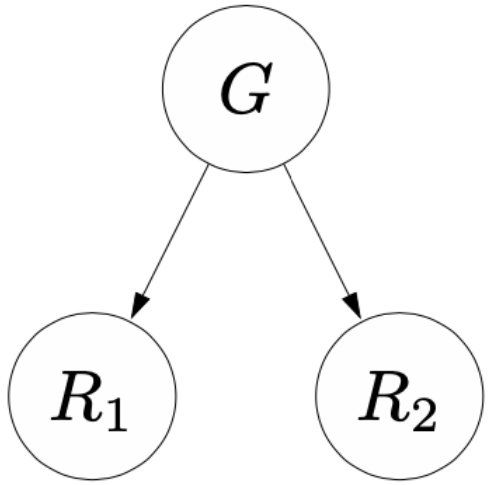
Variables:

- Genre $G \in \{\text{drama}, \text{comedy}\}$
- Jim's rating $R_1 \in \{1, 2, 3, 4, 5\}$
- Martha's rating $R_2 \in \{1, 2, 3, 4, 5\}$



$$\mathbb{P}(G = g, R_1 = r_1, R_2 = r_2) = p_G(g) p_{R_1}(r_1 \mid g) p_{R_2}(r_2 \mid g)$$

Example: inverted-v structure



$$\mathcal{D}_{\text{train}} = \{(\mathsf{d}, 4, 5), (\mathsf{d}, 4, 4), (\mathsf{d}, 5, 3), (\mathsf{c}, 1, 2), (\mathsf{c}, 5, 4)\}$$

Parameters: $\theta = (p_G, p_R)$

θ:

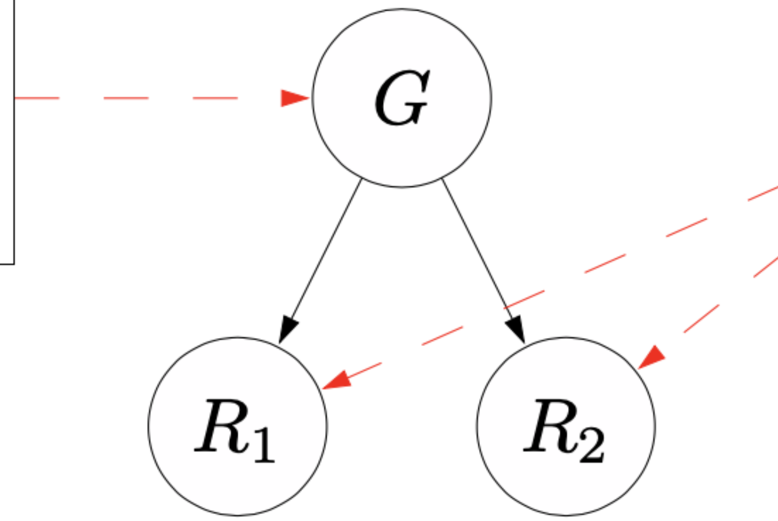
g	$p_G(g)$
d	3/5
c	2/5

g	r	$p_R(r \mid g)$
d	3	1/6
d	4	3/6
d	5	2/6
c	1	1/4
c	2	1/4
c	4	1/4
c	5	1/4

Parameter sharing

The local conditional distributions of different variables use the same parameters.

g	$p_G(g)$
c	2/5
d	3/5



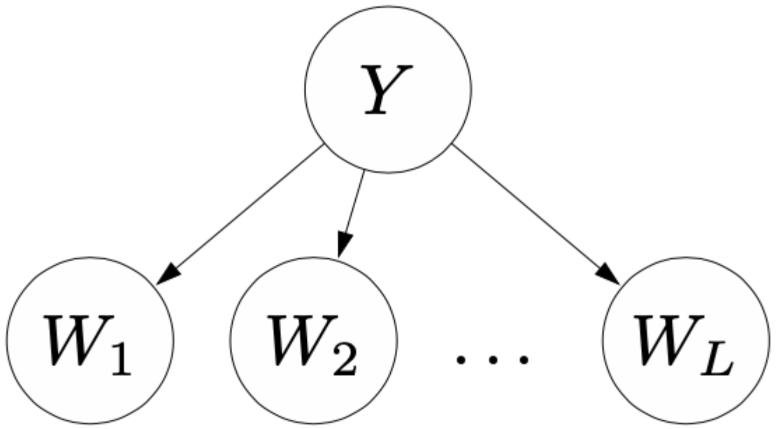
g	r	$p_R(r \mid g)$
d	3	1/6
d	4	3/6
d	5	2/6
c	1	1/4
c	2	1/4
c	4	1/4
c	5	1/4

Result: more reliable estimates, less expressive

Example: naïve bayes

Variables:

- Genre $Y \in \{\text{comedy}, \text{drama}\}$
- Movie review (sequence of words): $W_1, \dots, W_L$



$$\mathbb{P}(Y = y, W_1 = w_1, \dots, W_L = w_L) = p_{\text{genre}}(y) \prod_{j=1}^{L} p_{\text{word}}(w_j \mid y)$$

Parameters: $\theta = (p_{\text{genre}}, p_{\text{word}})$

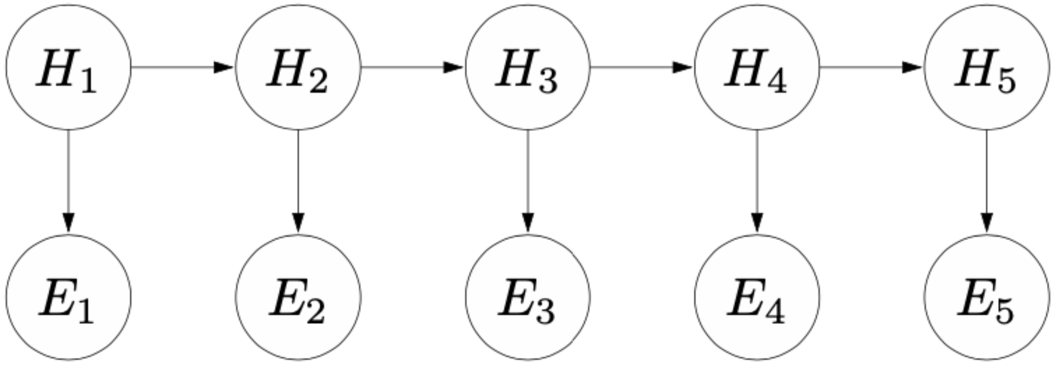
Question

If Y can take on 2 values and each W_j can take on D values, how many parameters are there?

Example: HMMs

Variables:

- $H_1, \dots, H_n$ (e.g., actual positions)
- $E_1, \dots, E_n$ (e.g., sensor readings)



$$\mathbb{P}(H = h, E = e) = p_{\text{start}}(h_1) \prod_{i=2}^{n} p_{\text{trans}}(h_i \mid h_{i-1}) \prod_{i=1}^{n} p_{\text{emit}}(e_i \mid h_i)$$

Parameters: $\theta = (p_{\text{start}}, p_{\text{trans}}, p_{\text{emit}})$

$\mathcal{D}_{\text{train}}$ is a set of full assignments to (H, E)

General case

Bayesian network: variables $X_1, \ldots, X_n$

Parameters: collection of distributions $\theta = \{p_d : d \in D\}$ (e.g., $D = \{\text{start}, \text{trans}, \text{emit}\}$)

Each variable X_i is generated from distribution p_{d_i}:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} p_{d_i}(x_i \mid x_{\text{Parents}(i)})$$

Parameter sharing: d_i could be same for multiple i

General case: learning algorithm

Input: training examples $\mathcal{D}_{\text{train}}$ of full assignments

Output: parameters $\theta = \{p_d : d \in D\}$

Count:

For each $x \in \mathcal{D}_{\text{train}}$:

For each variable x_i:

Increment $\text{count}_{d_i}(x_{\text{Parents}(i)}, x_i)$

Normalize:

For each d and local assignment $x_{\text{Parents}(i)}$:

Set $p_d(x_i \mid x_{\text{Parents}(i)}) \propto \text{count}_d(x_{\text{Parents}(i)}, x_i)$

Maximum likelihood

Maximum likelihood objective:

$$\max_\theta \prod_{x \in \mathcal{D}_{\text{train}}} \mathbb{P}(X = x; \theta)$$

Algorithm on previous slide exactly computes maximum likelihood parameters (closed form solution).

Maximum likelihood

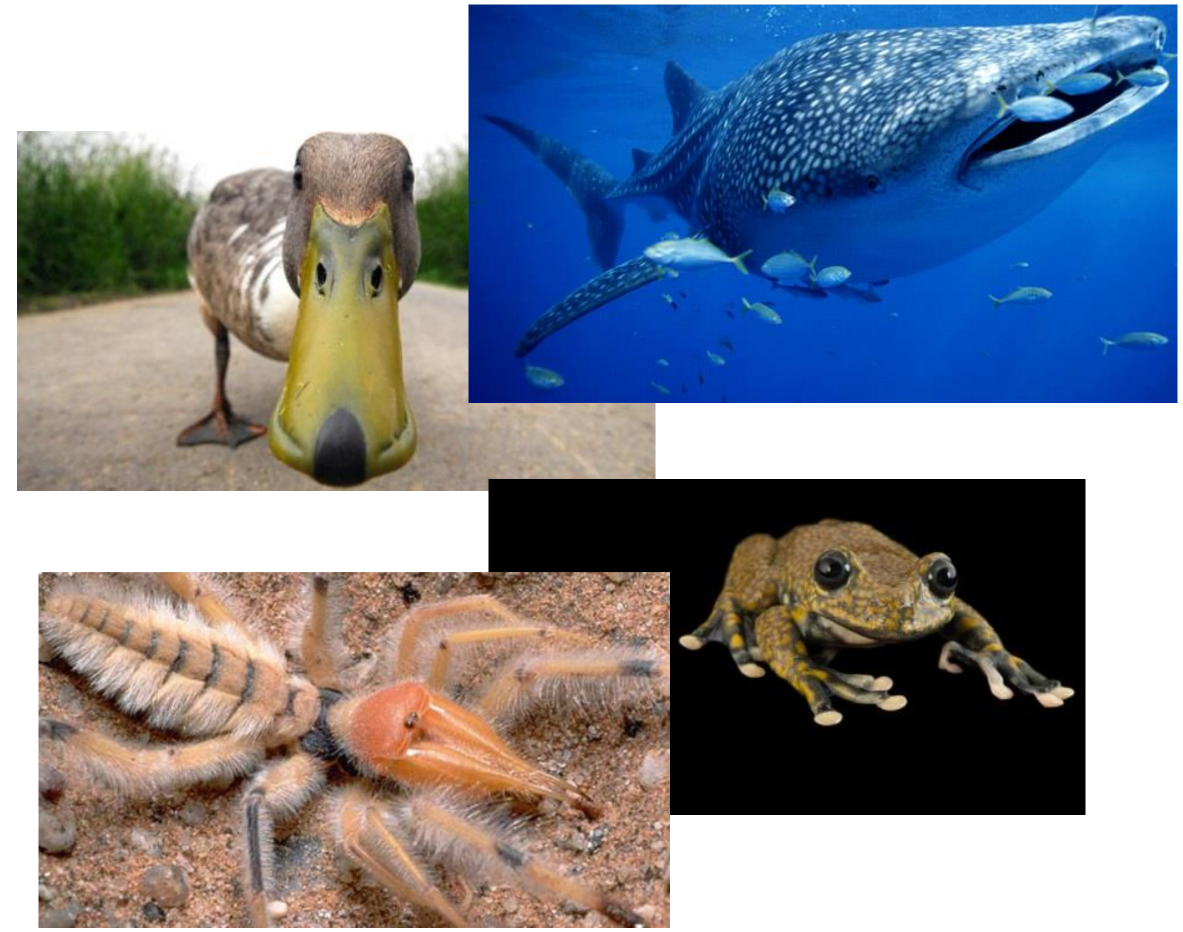
$$\mathcal{D}_{\text{train}} = \{(\mathsf{d}, 4), (\mathsf{d}, 5), (\mathsf{c}, 5)\}$$

$$\max_{p_G(\cdot)} (p_G(\mathsf{d}) p_G(\mathsf{d}) p_G(\mathsf{c})) \max_{p_R(\cdot \mid \mathsf{c})} p_R(5 \mid \mathsf{c}) \max_{p_R(\cdot \mid \mathsf{d})} (p_R(4 \mid \mathsf{d}) p_R(5 \mid \mathsf{d}))$$

- Key: decomposes into subproblems, one for each distribution d and assignment x_{Parents}
- For each subproblem, solve in closed form (Lagrange multipliers for sum-to-1 constraint)

Example: mammals vs, Non-mammals

Mammals



Non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

Example: mammals vs, Non-mammals

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Example: mammals vs, Non-mammals

Given $\boldsymbol{x} = \left(x_1, \cdots x_p\right)^T$

Goal is to predict class ω

Specifically, we want to find the value of ω that maximizes $P(\omega|\boldsymbol{x}) = P\left(\omega|x_1, \cdots x_p\right)$

$$P\left(\omega|x_1, \cdots x_p\right) \propto P\left(x_1, \cdots x_p|\omega\right)P(\omega)$$

Independence assumption among features

$$P\left(x_1, \cdots x_p|\omega\right) = P(x_1|\omega) \cdots P\left(x_p|\omega\right)$$

Example: mammals vs, Non-mammals

Tid	**Refund**	**Marital Status**	**Taxable Income**	**Evade**
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Class: $P(\omega_k) = \frac{N_{\omega_k}}{N}$

e.g., P(No) = 7/10,

P(Yes) = 3/10

For discrete attributes:

$$P(x_i|\omega_k) = \frac{|x_{ik}|}{N_{\omega_k}}$$

where $|x_{ik}|$ is number of instances having attribute x_i and belongs to class ω_k

Examples:

P(Status=Married|No) = 4/7

P(Refund=Yes|Yes)=0

Example: mammals vs, Non-mammals

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Normal distribution:

$$P(x_i \mid \omega_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} \exp\left(-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2} \right)$$

One for each (x_i, ω_j) pair

For (Income, Class=No):

If Class=No

sample mean = 110

sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} \exp\left(-\frac{(120-110)^2}{2(2975)} \right) = 0.0072$$

Example: mammals vs, Non-mammals

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:

If class=No:	sample mean=110	
	sample variance=2975	
If class=Yes:	sample mean=90	
	sample variance=25	

P(X|Class=No) = P(Refund=No|Class=No)
$\times$ P(Married| Class=No)
$\times$ P(Income=120K| Class=No)
= $4/7 \times 4/7 \times 0.0072 = 0.0024$

P(X|Class=Yes) = P(Refund=No| Class=Yes)
$\times$ P(Married| Class=Yes)
$\times$ P(Income=120K| Class=Yes)
= $1 \times 0 \times 1.2 \times 10^{-9} = 0$

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)
=> Class = No

Example: mammals vs, Non-mammals

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

2. Laplace smoothing - Scenario 1

Setup:

- You have a coin with an unknown probability of heads $p(\mathsf{H})$.
- You flip it 100 times, resulting in 23 heads, 77 tails.
- What is estimate of $p(\mathsf{H})$?

Maximum likelihood estimate:

$$p(\mathsf{H}) = 0.23 \quad p(\mathsf{T}) = 0.77$$

Scenario 2

Setup:

- You flip a coin once and get heads.
- What is estimate of $p(\mathsf{H})$?

Maximum likelihood estimate:

$$p(\mathsf{H}) = 1 \quad p(\mathsf{T}) = 0$$

Intuition: This is a bad estimate; real $p(\mathsf{H})$ should be closer to half

When have less data, maximum likelihood overfits, want a more reasonable estimate...

Regularization: Laplace smoothing

Maximum likelihood:

$$p(\mathsf{H}) = \frac{1}{1} \quad p(\mathsf{T}) = \frac{0}{1}$$

Maximum likelihood with Laplace smoothing:

$$p(\mathsf{H}) = \frac{1+1}{1+2} = \frac{2}{3} \quad p(\mathsf{T}) = \frac{0+1}{1+2} = \frac{1}{3}$$

Example: two variables

$\mathcal{D}_{\text{train}} = \{(\mathsf{d}, 4), (\mathsf{d}, 5), (\mathsf{c}, 5)\}$

Amount of smoothing: $\lambda = 1$

θ:

g	$p_G(g)$
d	3/5
c	2/5

g	r	$p_R(r \mid g)$
d	1	1/7
d	2	1/7
d	3	1/7
d	4	2/7
d	5	2/7
c	1	1/6
c	2	1/6
c	3	1/6
c	4	1/6
c	5	2/6

Regularization: Laplace smoothing

For each distribution d and partial assignment $(x_{\text{Parents}(i)}, x_i)$, add λ to $\text{count}_d(x_{\text{Parents}(i)}, x_i)$.

Then normalize to get probability estimates.

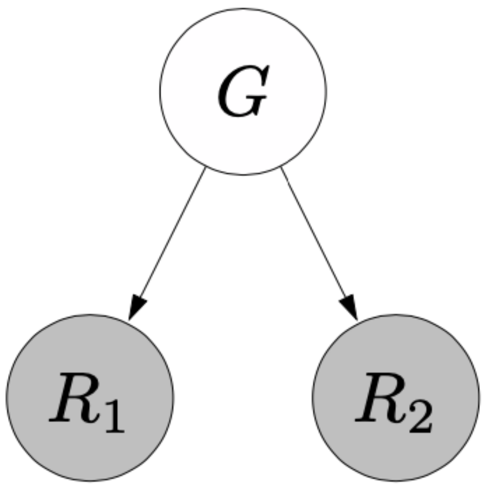
Interpretation: hallucinate λ occurrences of each local assignment

Larger $\lambda \Rightarrow$ more smoothing $\Rightarrow$ probabilities closer to uniform.

Data wins out in the end:

$$p(\text{H}) = \frac{1+1}{1+2} = \frac{2}{3} \quad p(\text{H}) = \frac{998+1}{998+2} = 0.999$$

3. Unsupervised Learning with EM: Motivation



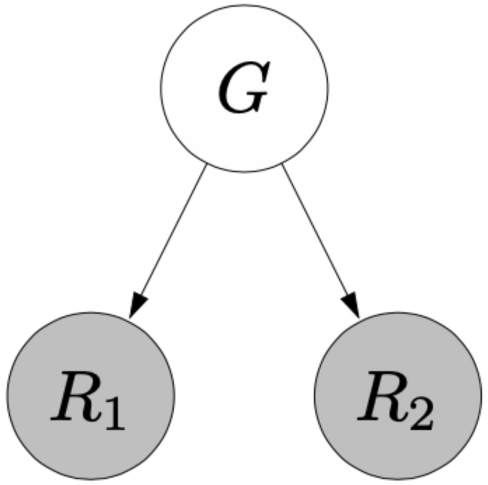
What if we **don't observe** some of the variables?

$$\mathcal{D}_{\text{train}} = \{(?, 4, 5), (?, 4, 4), (?, 5, 3), (?, 1, 2), (?, 5, 4)\}$$

Maximum marginal likelihood

Variables: H is hidden, $E = e$ is observed

Example:



$H = G \quad E = (R_1, R_2) \quad e = (4, 5)$

$\theta = (p_G, p_R)$

Maximum marginal likelihood objective:

$$\max_{\theta} \prod_{e \in \mathcal{D}_{\text{train}}} \mathbb{P}(E = e; \theta)$$

$$= \max_{\theta} \prod_{e \in \mathcal{D}_{\text{train}}} \sum_{h} \mathbb{P}(H = h, E = e; \theta)$$

Expectation Maximization

Inspiration: K-means

Variables: H is hidden, E is observed (to be e)

E-step:

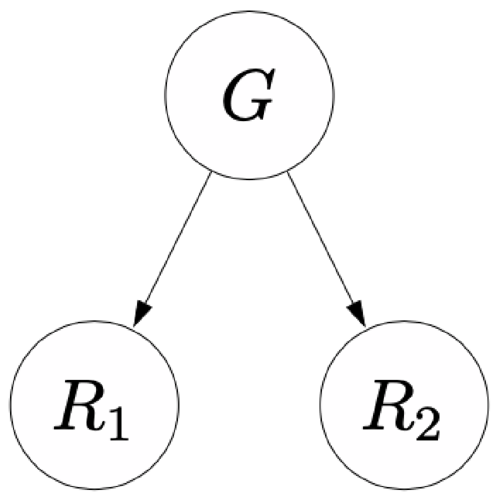
- Compute $q(h) = \mathbb{P}(H = h \mid E = e; \theta)$ for each h (use any probabilistic inference algorithm)
- Create weighted points: (h, e) with weight $q(h)$

M-step:

- Compute maximum likelihood (just count and normalize) to get θ

Repeat until convergence.

Example: one iteration of EM



$$\mathcal{D}_{\text{train}} = \{(?, 2, 2), (?, 1, 2)\}$$

θ:

g	$p_G(g)$
c	0.5
d	0.5

g	r	$p_R(r \mid g)$
c	1	0.4
c	2	0.6
d	1	0.6
d	2	0.4

E-step

(r_1, r_2)	g	$\mathbb{P}(G = g, R_1 = r_1, R_2 = r_2)$	$q(g)$
(2, 2)	c	$0.5 \cdot 0.6 \cdot 0.6$	$\frac{0.18}{0.18+0.08} = 0.69$
(2, 2)	d	$0.5 \cdot 0.4 \cdot 0.4$	$\frac{0.08}{0.18+0.08} = 0.31$
(1, 2)	c	$0.5 \cdot 0.4 \cdot 0.6$	$\frac{0.12}{0.12+0.12} = 0.5$
(1, 2)	d	$0.5 \cdot 0.6 \cdot 0.4$	$\frac{0.12}{0.12+0.12} = 0.5$

M-step

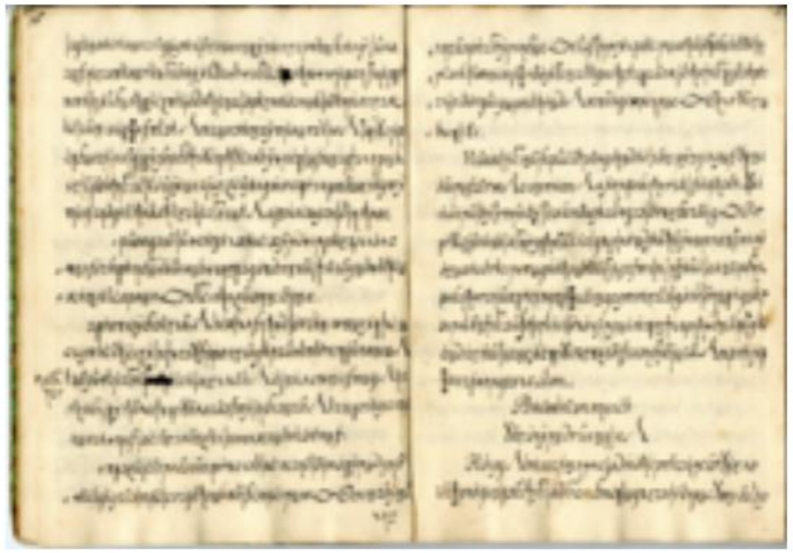
θ:

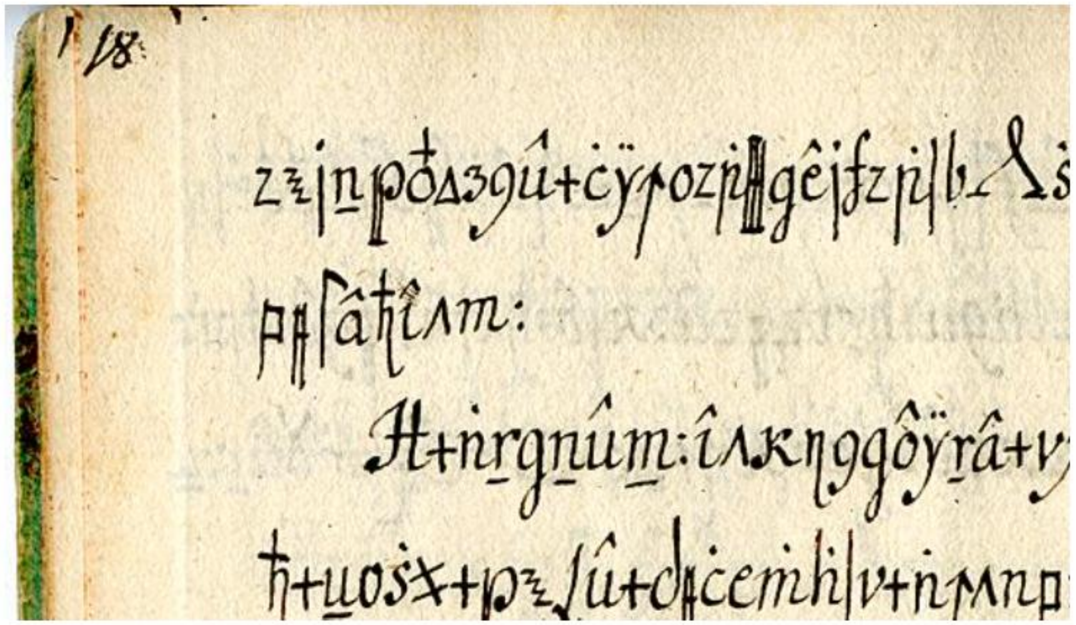
g	count	$p_G(g)$
c	0.69 + 0.5	0.59
d	0.31 + 0.5	0.41

g	r	count	$p_R(r \mid g)$
c	1	0.5	0.21
c	2	0.5 + 0.69 + 0.69	0.79
d	1	0.5	0.31
d	2	0.5 + 0.31 + 0.31	0.69

Application: decipherment

Copiale cipher (105-page encrypted volume from 1730s):





Substitution ciphers

Letter substitution table (unknown):

Plain:	`abcdefghijklmnopqrstuvwxyz`
Cipher:	`plokmijnuhbygvtfcrdxeszaqw`

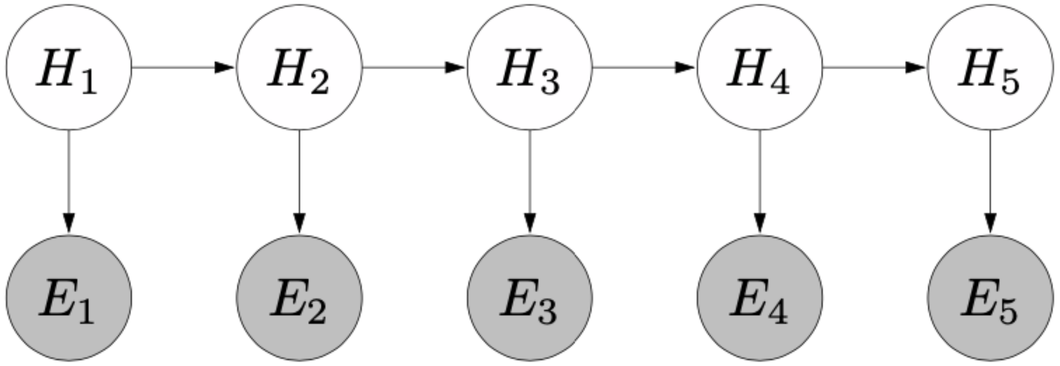
Plaintext (unknown): `hello world`

Ciphertext (known): `nmyyt ztryk`

Application: decipherment as an HMM

Variables:

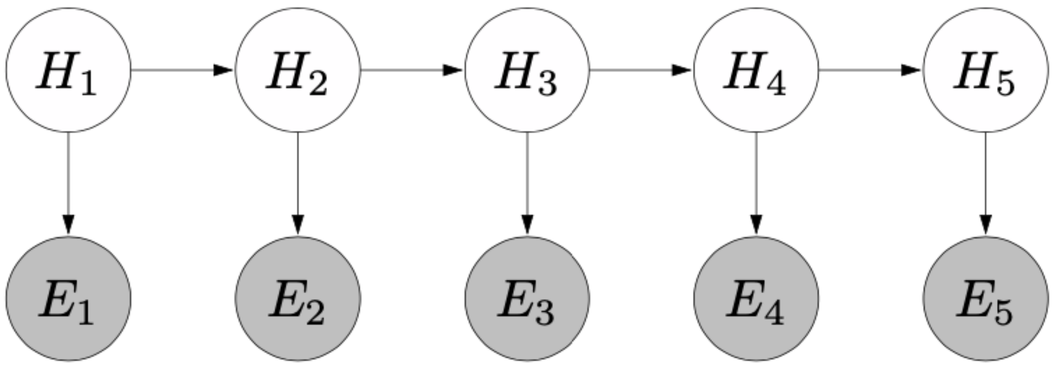
- $H_1, \ldots, H_n$ (e.g., characters of plaintext)
- $E_1, \ldots, E_n$ (e.g., characters of ciphertext)



$$\mathbb{P}(H = h, E = e) = p_{\text{start}}(h_1) \prod_{i=2}^{n} p_{\text{trans}}(h_i \mid h_{i-1}) \prod_{i=1}^{n} p_{\text{emit}}(e_i \mid h_i)$$

Parameters: $\theta = (p_{\text{start}}, p_{\text{trans}}, p_{\text{emit}})$

Application: decipherment as an HMM

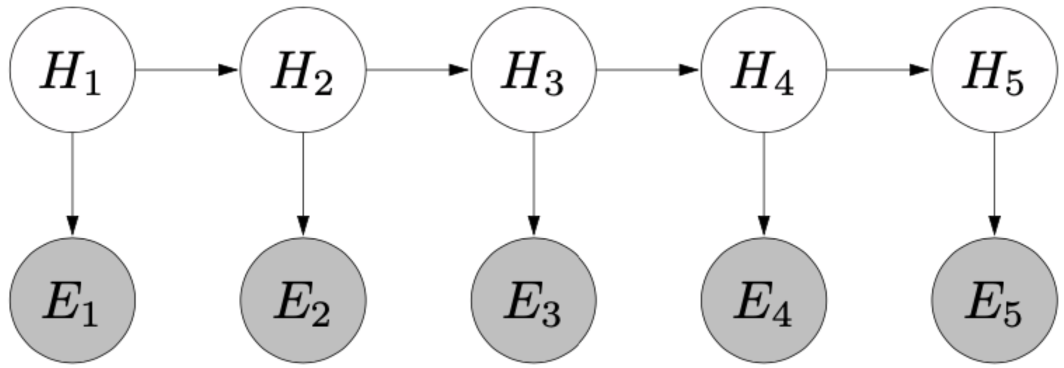


Strategy:

- p_{start}: set to uniform
- p_{trans}: estimate on tons of English text
- p_{emit}: **substitution table**, from EM

Intuition: rely on language model (p_{trans}) to favor plaintexts h that look like English

Application: decipherment as an HMM



E-step: forward-backward algorithm computes

$$q_i(h) \stackrel{\text{def}}{=} \mathbb{P}(H_i = h \mid E_1 = e_1, \ldots E_n = e_n)$$

M-step: count (fractional) and normalize

$$\text{count}_{\text{emit}}(h, e) = \textstyle\sum_{i=1}^{n} q_i(h) \cdot [e_i = e]$$

$$p_{\text{emit}}(e \mid h) \propto \text{count}_{\text{emit}}(h, e)$$

[semi-live solution]

Summary

(Bayesian network without parameters) + training examples

Learning: maximum likelihood (+Laplace smoothing, +EM)

$Q \mid E \Rightarrow$ Parameters θ (of Bayesian network) $\Rightarrow \mathbb{P}(Q \mid E; \theta)$