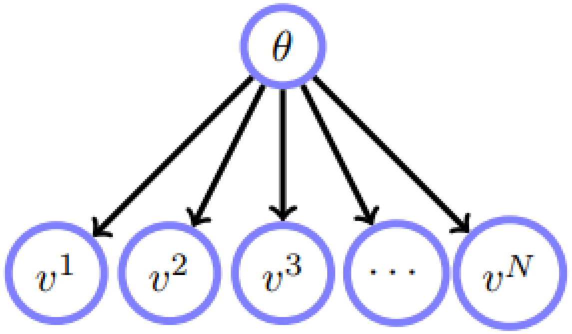
Chapter 9
Learning as Inference

9.1 Learning as Inference

- 9.1.1 Learning the bias of a coin
  - Coin toss of N times
  - Goal: to estimate the probability that the coin will be a head
  - We write $v^n = 1$ if on toss n the coin comes up heads,
    and $v^n = 0$ if it is tails
  - Let $\theta = p(v^n = 1)$, which is called the bias of the coin

$$p(v^1, \dots, v^N, \theta) = p(\theta) \prod_{n=1}^{N} p(v^n | \theta)$$



belief network

- Learning refers to using the observations $v^1, \dots, v^N$ to infer θ
- For simplicity, let $\mathcal{V} = (v^1, \dots, v^N)$
- Posterior

$$p(\theta|\mathcal{V}) = \frac{p(\mathcal{V}, \theta)}{p(\mathcal{V})} = \frac{p(\mathcal{V}|\theta)p(\theta)}{p(\mathcal{V})}$$

$$\begin{aligned} p(\theta|\mathcal{V}) &\propto p(\mathcal{V}|\theta)p(\theta) \\ &= p(v^1|\theta) \cdots p(v^N|\theta)p(\theta) \\ &= p(\theta)\theta^{N_H}(1-\theta)^{N_T} \end{aligned}$$

where $N_H = \#\text{head}$, $N_T = \#\text{tail}$

$$\boxed{N_H = \sum_{n=1}^{N} \mathbb{I}[v^n = 1]}$$

- MAP

$$\operatorname*{argmax}_{\theta} p(\theta|\mathcal{V})$$

- For simplicity we assume that $\theta \in \{0.1, 0.5, 0.8\}$ and

$$p(\theta = 0.1) = 0.15,\ \ p(\theta = 0.5) = 0.8,\ \ p(\theta = 0.8) = 0.05$$

  - This prior expresses that we have
    - 80% belief that the coin is 'fair'
    - 5% belief the coin is biased to land heads (with $\theta = 0.8$)
    - 15% belief the coin is biased to land tails (with $\theta = 0.1$)

Experiments

$N_H = 2$, $N_T = 8$

$$p(\theta = 0.1 \mid \mathcal{V}) = k \times 0.15 \times 0.1^2 \times 0.9^8 = k \times 6.46 \times 10^{-4}$$
$$p(\theta = 0.5 \mid \mathcal{V}) = k \times 0.8 \times 0.5^2 \times 0.5^8 = k \times 7.81 \times 10^{-4}$$
$$p(\theta = 0.8 \mid \mathcal{V}) = k \times 0.05 \times 0.8^2 \times 0.2^8 = k \times 8.19 \times 10^{-8}$$

$$k \times 6.46 \times 10^{-4} + k \times 7.81 \times 10^{-4} + k \times 8.19 \times 10^{-8} = 1$$
$$k = 1/0.0014$$
$$p(\theta = 0.1 \mid \mathcal{V}) \approx 0.4525,\ p(\theta = 0.5 \mid \mathcal{V}) \approx 0.5475,\ p(\theta = 0.8 \mid \mathcal{V}) \approx 0.0001$$

$N_H = 20,\ N_T = 80$

$$p(\theta = 0.1 \mid \mathcal{V}) \approx 1 - 1.93 \times 10^{-6}$$
$$p(\theta = 0.5 \mid \mathcal{V}) \approx 1.93 \times 10^{-6}$$
$$p(\theta = 0.8 \mid \mathcal{V}) \approx 2.13 \times 10^{-35}$$

- 9.1.2 Making decisions
  - If we correctly state the bias of the coin we gain 10 points; being incorrect, loses 20 points.
  - Let θ^0 be the true value for the bias
  - Suppose that we state the bias as θ
  - The points that we gain is
$$U(\theta, \theta^0) = 10\, \mathbb{I}[\theta = \theta^0] - 20\, \mathbb{I}[\theta \neq \theta^0]$$
  - The expected utility of the decision
$$U(\theta) = U(\theta, \theta^0 = 0.1)\, p(\theta^0 = 0.1|\mathcal{V}) \\ + U(\theta, \theta^0 = 0.5)\, p(\theta^0 = 0.5|\mathcal{V}) + U(\theta, \theta^0 = 0.8)\, p(\theta^0 = 0.8|\mathcal{V})$$

$N_H = 2$, $N_T = 8$

$$U(\theta = 0.1) = -6.4270$$
$$U(\theta = 0.5) = -3.5770$$
$$U(\theta = 0.8) = -19.999$$

$N_H = 20$, $N_T = 80$

$$U(\theta = 0.1) = 9.9999$$
$$U(\theta = 0.5) \approx -20.0$$
$$U(\theta = 0.8) \approx -20.0$$

- 9.1.3 A continuum of parameters
  - Equation

$$p(\theta|\mathcal{V}) \propto p(\theta)\theta^{N_H}(1-\theta)^{N_T}$$

  - θ is a continuous variable
  - The prior $p(\theta) = ?$

- Using a flat prior

$$p(\theta) = k \text{ for some constant } k$$

$$\int_0^1 p(\theta)d\theta = 1 \implies k = 1$$

$$p(\theta|\mathcal{V}) \propto p(\theta)\theta^{N_H}(1-\theta)^{N_T}$$

$$p(\theta|\mathcal{V}) = \frac{1}{c}\theta^{N_H}(1-\theta)^{N_T} \text{ where } c = \int_0^1 \theta^{N_H}(1-\theta)^{N_T}d\theta$$

$$\operatorname*{argmax}_\theta p(\theta|\mathcal{V}) = \frac{N_H}{N}$$

- Using a conjugate prior

$$p(\theta|\mathcal{V}) \propto p(\theta)\theta^{N_H}(1-\theta)^{N_T}$$

The conjugate of $\theta^{N_H}(1-\theta)^{N_T}$ is a Beta distribution

$$p(\theta) = \frac{1}{k}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

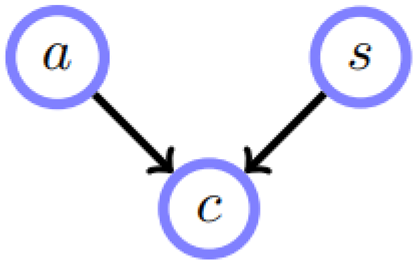
$$p(\theta|\mathcal{V}) = \frac{1}{c}\theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{N_H}(1-\theta)^{N_T}$$
$$= \frac{1}{c}\theta^{N_H+\alpha-1}(1-\theta)^{N_T+\beta-1}$$

$$\underset{\theta}{\operatorname{argmax}}\, p(\theta|\mathcal{V}) = \frac{N_H+\alpha-1}{N+\alpha+\beta-2}$$

9.3 Maximum Likelihood Training of Belief Networks

- Lung cancer
  - Relationship between
    - Exposure to asbestos (a)
    - Being a smoker (s)
    - The incidence of lung cancer (c)

$$p(a, s, c) = p(c|a, s)p(a)p(s)$$
$$\mathrm{dom}(a) = \{0,1\}$$
$$\mathrm{dom}(s) = \{0,1\}$$
$$\mathrm{dom}(c) = \{0,1\}$$



a	s	c
1	1	1
1	0	0
0	1	1
0	1	0
1	1	1
0	0	0
1	0	1

$$x^i = (a^i, s^i, c^i)$$
$$i = 1, \dots, 7$$

Problem.

$\mathcal{X} = \{x^1, \dots, x^7\}$

Infer $\theta = p(c = 1 | a = 0, s = 1)$ by ML

Solution.

Let $\mathcal{X}_0 = \{x^3, x^4\}$. Then

$$\operatorname*{argmax}_\theta p(\mathcal{X}|\theta) = \operatorname*{argmax}_\theta p(\mathcal{X}_0|\theta).$$

By direct computation

$$\begin{aligned} p(\mathcal{X}_0|\theta) &= p(x^3|\theta)p(x^4|\theta) \\ &= \theta(1-\theta)p(a=0)p(s=1) \end{aligned}$$

Hence the ML solution is

$$\theta = 0.5$$

a	s	c
1	1	1
1	0	0
0	1	1
0	1	0
1	1	1
0	0	0
1	0	1

Problem.

Infer $p(c|a,s)$ by ML

$$p(c=1|a=0,s=0)$$
$$p(c=1|a=0,s=1)$$
$$p(c=1|a=1,s=0)$$
$$p(c=1|a=1,s=1)$$

Systematic solution?

8.7.3 Maximum likelihood and the empirical distribution

Let $\mathcal{X} = \{x^1, \dots, x^N\}$ be a data set and q the empirical distribution

A distribution p_0 minimizes $\mathrm{KL}(q|p) \Leftrightarrow p_0$ is obtained by maximum likelihood

Proof. Consider the equation

$$\begin{aligned}\mathrm{KL}(q|p) &= \langle \log q(x) \rangle_{q(x)} - \langle \log p(x) \rangle_{q(x)} \\ &= -\langle \log p(x) \rangle_{q(x)} + \text{const.} \\ &= -\frac{1}{N}\sum_{n=1}^{N} \log p(x^n) + \text{const.}\end{aligned}$$

$$\boxed{\langle f(x) \rangle_{q(x)} = \frac{1}{N}\sum_{n=1}^{N} f(x^n)}$$

Hence

$$\begin{aligned}\underset{p}{\operatorname{argmin}}\, \mathrm{KL}(q|p) &= \underset{p}{\operatorname{argmax}} \sum_{n=1}^{N} \log p(x^n) \\ &= \underset{p}{\operatorname{argmax}} \prod_{n=1}^{N} p(x^n)\end{aligned}$$

In the last term $\prod_{n=1}^{N} p(x^n)$ is the likelihood

- Maximum likelihood corresponds to counting

For a BN we have

$$p(x) = \prod_{i=1}^{K} p(x_i|pa(x_i))$$

We want to find the distribution $p(x)$ that minimizes the KL divergence $\mathrm{KL}(q|p)$ for the empirical distribution $q(x)$

$$\mathrm{KL}(q|p) = -\left\langle \sum_{i=1}^{K} \log p\big(x_i|\mathrm{pa}(x_i)\big) \right\rangle_{q(x)} + \text{const.}$$

$$= -\sum_{i=1}^{K} \big\langle \log p\big(x_i|\mathrm{pa}(x_i)\big) \big\rangle_{q(x_i,\mathrm{pa}(x_i))} + \text{const.}$$

$$\begin{aligned}\mathrm{KL}(q|p) &= \sum_{i=1}^{K}\left[\left\langle \log q\big(x_i|\mathrm{pa}(x_i)\big)\right\rangle_{q(x_i,\mathrm{pa}(x_i))} - \left\langle \log p\big(x_i|\mathrm{pa}(x_i)\big)\right\rangle_{q(x_i,\mathrm{pa}(x_i))}\right] + \mathrm{const.}\\ &= \sum_{i=1}^{K}\mathrm{KL}\Big(q\big(x_i|\mathrm{pa}(x_i)\big)\Big|q\big(x_i|\mathrm{pa}(x_i)\big)\Big) + \mathrm{const.}\end{aligned}$$

Therefore, if $p(x)$ minimizes $\mathrm{KL}(q|p)$, then

$$p(x_i|\mathrm{pa}(x_i)) = q(x_i|\mathrm{pa}(x_i)).$$

The distribution p is determined by counting

$$p(x_i = \mathrm{s}|\mathrm{pa}(x_i) = \mathrm{t}) = \frac{\sum_{n=1}^{N}\mathbb{I}[x_i^n = \mathrm{s}, \mathrm{pa}(x_i^n) = \mathrm{t}]}{\sum_{n=1}^{N}\mathbb{I}\big[\mathrm{pa}\big(x_i^n\big) = \mathrm{t}\big]}$$

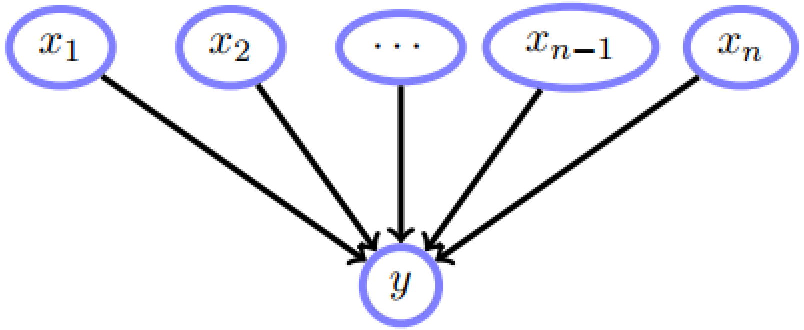
a	s	c
1	1	1
1	0	0
0	1	1
0	1	0
1	1	1
0	0	0
1	0	1

$p(c|a,s)$ by ML is obtained by counting (empirical distribution)

$$p(c=1|a=0,s=0)=0/1$$
$$p(c=1|a=0,s=1)=1/2$$
$$p(c=1|a=1,s=0)=1/2$$
$$p(c=1|a=1,s=1)=2/2$$

- Conditional probability functions

Binary variable y with n binary parental variables $x_1, \dots, x_n$
Find $p(y|x)$



There are 2^n entries in the CPT of $p(y|x)$
For very large 2^n, we assume that p is a function f

$$p(y=1|\mathbf{x},\mathbf{w}) = f(\mathbf{x},\mathbf{w})$$

where $w = (w_1, \dots, w_n) \in \mathbb{R}^n$

For example

$$f(\mathbf{x},\mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^{\mathrm{T}}\mathbf{x}}}$$

a	s	c
1	1	1
1	0	0
0	1	1
0	1	0
1	1	1
0	0	0
1	0	1

$$x^i = \left(a^i, s^i, c^i\right)$$
$$i = 1, \ldots, 7$$

Assume that $p(c = 1|a, s) = e^{\theta_1 + \theta_2 a + \theta_3 s}$

The likelihood is

$$p(\mathcal{X}|\theta_1, \theta_2) = \prod_{i=1}^{7} p\left(a^i, s^i, c^i \middle| \theta_1, \theta_2\right)$$
$$= \prod_{i=1}^{7} p\left(c^i \middle| a^i, s^i, \theta_1, \theta_2\right) p\left(a^i | \theta_1, \theta_2\right) p(s^i | \theta_1, \theta_2)$$

The log likelihood is

$$\log p(\mathcal{X}|w) = \sum_{i=1}^{7} \log p\left(c^i \middle| a^i, s^i, \theta_1, \theta_2\right) + \text{const.}$$
$$= \sum_{i=1}^{7} \mathbb{I}\left[c^i = 1\right](\theta_1 + \theta_2 a^i + \theta_3 s^i) + \sum_{i=1}^{7} \mathbb{I}\left[c^i = 1\right] \log\left(1 - e^{\theta_1 + \theta_2 a^i + \theta_3 s^i}\right)$$
$$+ \text{const.}$$

Maximize this value

9.4 Bayesian Belief Network Training

a	s	c
1	1	1
1	0	0
0	1	1
0	1	0
1	1	1
0	0	0
1	0	1

$$v^n = (a^n, s^n, c^n)$$
$$\mathcal{V} = \{v^1, \dots, v^7\}$$

$$\theta_a = p(a = 1),\ \theta_s = p(s = 1),\ \theta_c^{0,1} = p(c = 1 | a = 0, s = 1)$$
$$\theta_c = \left(\theta_c^{0,0}, \theta_c^{0,1}, \theta_c^{1,0}, \theta_c^{1,1}\right)$$