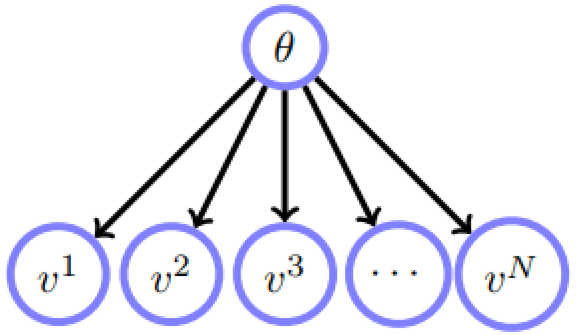
Chapter 9
Learning as Inference

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

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



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

$$\underset{\theta}{\operatorname{argmax}}\, p(\theta|\mathcal{V})$$

- 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

$$\begin{aligned} U(\theta) = {} & U(\theta, \theta^0 = 0.1)\, p(\theta^0 = 0.1|\mathcal{V}) \\ & + U(\theta, \theta^0 = 0.5)\, p(\theta^0 = 0.5|\mathcal{V}) + U(\theta, \theta^0 = 0.8)\, p(\theta^0 = 0.8|\mathcal{V}) \end{aligned}$$

$N_H = 2,\ N_T = 8$

$U(\theta = 0.1) = -6.4270$

$U(\theta = 0.5) = -3.5770$

$U(\theta = 0.8) = -19.999$

$N_H = 20,\ N_T = 80$

$U(\theta = 0.1) = 9.9999$

$U(\theta = 0.5) \approx -20.0$

$U(\theta = 0.8) \approx -20.0$

- 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

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

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

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