Day 3: Probabilistic models of cognition

1 Working with probability distributions

1. Marginalization: $P(a) = \sum_b P(a, b)$
2. Conditional probability: $P(a|b) = \frac{P(a,b)}{P(b)}$
3. Chain rule: $P(a|b)P(b) = P(a, b)$
4. **Bayes rule**: $P(h|d) = \frac{P(d|h)P(h)}{P(d)}$
5. Bayes rule with background knowledge: $P(h|d, b) = \frac{P(d|h,b)P(h|b)}{P(d|b)}$

2 Bayesian concept learning

Notation and problem formulation

- $\mathcal{H} = \{h_1, \ldots, h_M\}$ is a hypothesis space of concepts.
- $X = \{x_1, \ldots, x_n\}$ is a set of n positive examples of some concept C that belongs to $\mathcal{H}$
- A Bayesian learner's beliefs about the identity of the unknown concept C are captured by

$$P(h|X) = \frac{P(X|h)P(h)}{P(X)} \propto P(X|h)P(h) \tag{1}$$

Prior $P(h)$

Hypothesis space for number game

Mathematical properites:

- Odd numbers
- Even numbers
- Square numbers
- Cube numbers
- Primes
- Multiples of n: $3 \leq n \leq 12$
- Powers of n: $2 \leq n \leq 10$
- Numbers ending in n: $0 \leq n \leq 9$

Magnitude properties:

- Intervals between n and m: $1 \leq n \leq 100$; $n \leq m \leq 100$

- Total probability assigned to mathematical concepts is λ.
- Total probability assigned to magnitude concepts is $1 - \lambda$.
- Total probability assigned to all other concepts is 0.

Possible likelihoods $P(X|h)$

- Strong sampling:

$$p(X|h) = \begin{cases} \left[\frac{1}{\text{size}(h)}\right]^n & \text{if all } x_i \text{ are in } h \\ 0 & \text{otherwise} \end{cases}$$

- Weak sampling:

$$p(X|h) = \begin{cases} 1 & \text{if labels for all } x_i \text{ are consistent with } h \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Prediction by hypothesis averaging

- Let $\mathcal{H}_X$ be the set of all hypotheses that are consistent with the data X
- A Bayesian learner will make a prediction about an unlabeled item y by using

$$P(y \in C|X) = \sum_{h \in \mathcal{H}} P(y \in C|h) P(h|X) = \sum_{h \in \mathcal{H}_y} P(h|X) \tag{3}$$

3 Bayesian networks

A Bayesian network (or Bayes net or directed graphical model) specifies a joint distribution $P(v_1, \ldots, v_n)$.

The network includes:

- A directed acyclic graph G with a node for each variable V_i. You should aim to use graphs where an edge from V_i to V_j means that V_i has a direct causal influence on V_j.
- A conditional probability distribution $P(v_i|\text{pa}(V_i))$ that specifies how the value of V_i depends on the values of its parent nodes $\text{Pa}(V_i)$.

The joint distribution can be represented as

$$P(v_1, \ldots, v_n) = \prod_i P(v_i|\text{pa}(V_i)) \tag{4}$$

Why work with Bayesian networks?

- Bayesian networks help modelers define high dimensional distributions.
- Bayesian networks provide a concise way of representing probability distributions.
- Bayesian networks often support efficient inference.
- Bayesian networks are modular and therefore easy to extend.
- Bayesian networks can be used to define causal models that reason about interventions and counterfactuals.

4 Inference by Sampling

Inference by sampling from the prior

For the food web problem, consider how we generalize to an unobserved node in the food web: e.g.

$$P(\text{humans}|obs) = \sum_h P(\text{humans}|h)P(h|obs) \tag{5}$$

$$\propto \sum_h P(\text{humans}|h)P(obs|h)P(h) \tag{6}$$

where the last step follows from Bayes rule.

Equation 6 can be approximated by drawing M samples $\{h^1, \ldots, h^M\}$ from the **prior** distribution $P(h)$:

$$\sum_h P(\text{humans}|h)P(obs|h)P(h) \approx \frac{1}{M}\sum_{i=1}^{M} P(\text{humans}|h^i)P(obs|h^i) \tag{7}$$

Sampling from the prior is often straightforward, but Equation 7 tends to work only for fairly small problems.

Inference by sampling from the posterior

Equation 5 can be approximated by drawing M samples $\{h^1, \ldots, h^M\}$ from the **posterior** distribution $P(h|obs)$:

$$\sum_h P(\text{humans}|h)P(h|obs) \approx \frac{1}{M}\sum_{i=1}^{M} P(\text{humans}|h^i) \tag{8}$$

Sampling from the posterior is more difficult, but can be achieved using MCMC (Markov Chain Monte Carlo) sampling as implemented by packages like JAGS. MCMC can be successfully applied to relatively large problems.