Information and Inference

Jonathan Lawry

Department of Engineering Mathematics
University of Bristol

Knowledge Elicitation

- Knowledge based systems are often dependent on formal representations of knowledge elicited from domain experts.
- For example, discussions with a doctor may result in the following knowledge base:
- Symptom D *strongly* suggests disease B, Symptom D is *rather uncommon*, A patient of type E cannot have disease B, *About* 50% of patients are of type E, Condition F is *mainly found* in patients of type E etc
- Suppose then we encounter a patient which has condition F and also exhibits symptom D.
- How likely is it that this patient has disease B?

Probabilistic Formulation

- Assume that belief is quantified by a probability measure.
- After further interrogation the doctor gives the following representation of the above statements.
- $P(B|D) = 0.8$, $P(D) = 0.3$, $P(B|E) = 0$, $P(E) = 0.5$, $P(E|F) = 0.7$
- The query requires us to evaluate: $P(B|D \cap F)$
- But what in general can we infer from a knowledge base of this form and what extra assumptions do we need to get precise probability values?

Linear Knowledge

- We assume that a knowledge base can be represented by a set of linear equations on a probability measure P

$$K = \sum_{i=1}^{n_j} a_{i,j} P(A_{i,j}) = b_j : j = 1, \ldots, m$$

- Inference correstponds to:
- **1:** Identify the set of probability measures on W which satisfy K.
- **2:** Pick one of them according to some set of *common sense* principles.

Example

- A system has 4 possible states $W = \{w_1, w_2, w_3, w_4\}$.
- High Pressure: $HP = \{w_1, w_2\}$, High Temperature: $HT = \{w_1, w_3\}$, Failure: $F = \{w_1, w_4\}$
- Let $K = \{P(HP) = 0.8, P(HT) = 0.7\}$
- What is the probability of failure given K?
- Let $p(w_i) = p_i$ for $i = 1, \ldots, 4$ then we have the following contraints:
- $p_1 + p_2 = 0.8$, $p_1 + p_3 = 0.7$, $p_1 + p_2 + p_3 + p_4 = 1$
- Therefore, $p_2 = 0.8 - p_1$, $p_3 = 0.7 - p_1$ and $p_4 = p_1 - 0.5$ where $0.5 \leq p_1 \leq 0.7$.
- $P(F) = p_1 + p_4 = 2p_1 - 0.5$ and hence $0.5 \leq P(F) \leq 0.9$
- Are some values more reasonable than others?

Inference Process

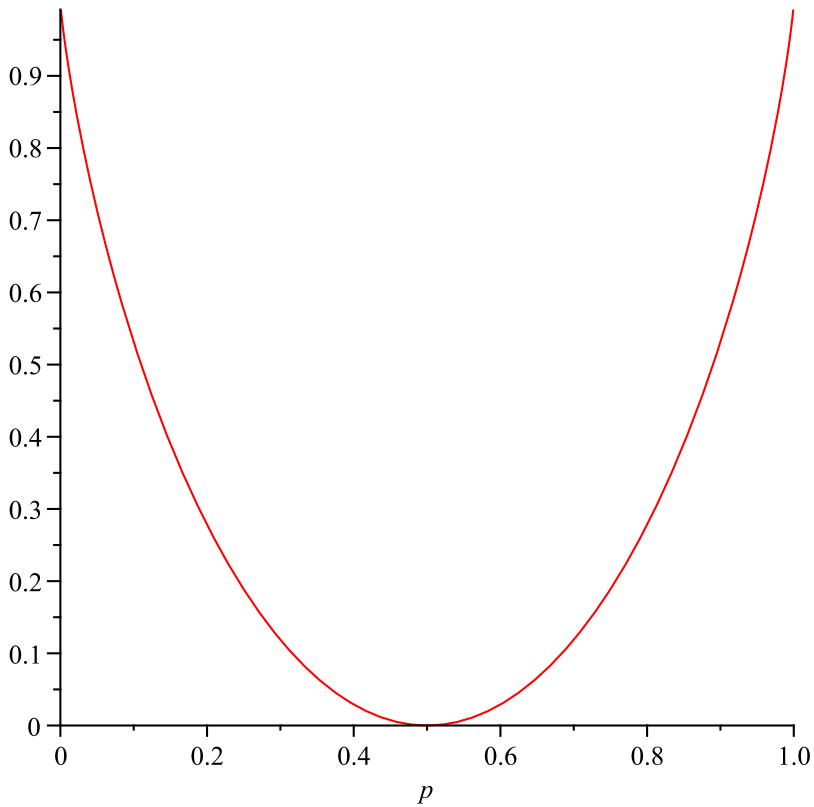
- For knowledge base K let $V(K) \subseteq V$ be the set of probability distributions satisfying K.
- For the above example:
  $V(K) = \{\langle p_1, 0.8 - p_1, 0.7 - p_1, p_1 - 0.5\rangle : 0.5 \leq p_1 \leq 0.7\}$
- An inference process is a function N which given a linear knowledge base K picks a probability measure $N(K)$ satisfying K.
- Alternatively, we can think of N as picking a single point in $V(K)$.
- For the above example, an inference processes effectively selects a value for p_1 between 0.5 and 0.7.

Measuring Information

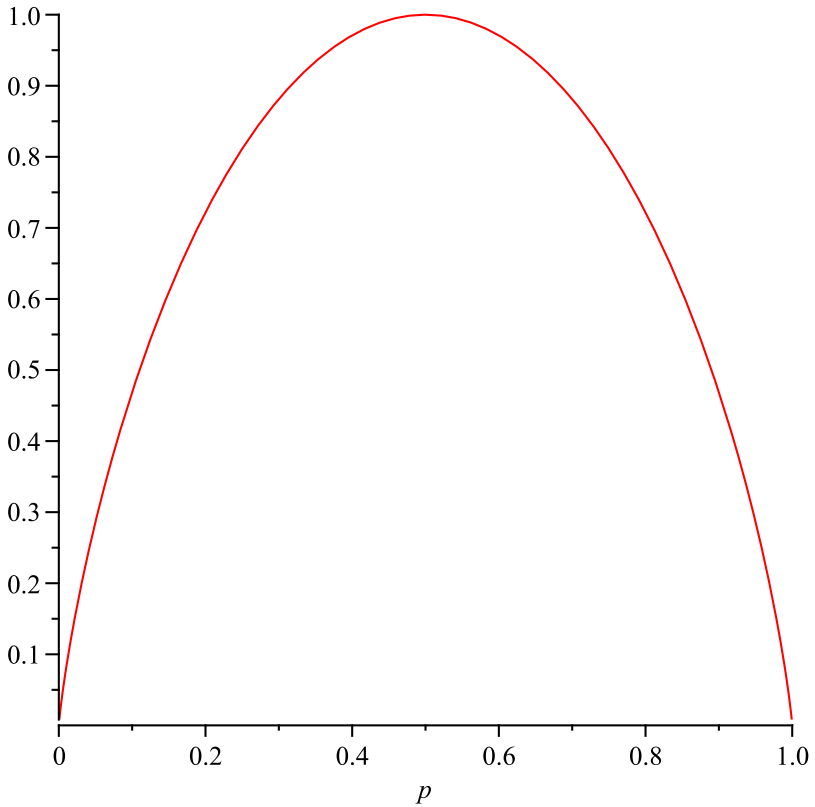
- Not all probability distributions have the same information content.
- Consider a coin tossing experiment so that $W = \{H, T\}$. In this case $V = \{\langle p, 1-p\rangle : 0 \leq p \leq 1\}$
- Distribution $\langle 1, 0\rangle$ tells us that the next throw is *certain* to be a head.
- Distribution $\langle 0, 1\rangle$ tells us that the next throw is *certain* to be a tail.
- Distribution $\langle \frac{1}{2}, \frac{1}{2}\rangle$ only tells us that we are *completely uncertain* as to the outcome of the next throw.

Information and Entropy

Information content for different p



Entropy is absense of information



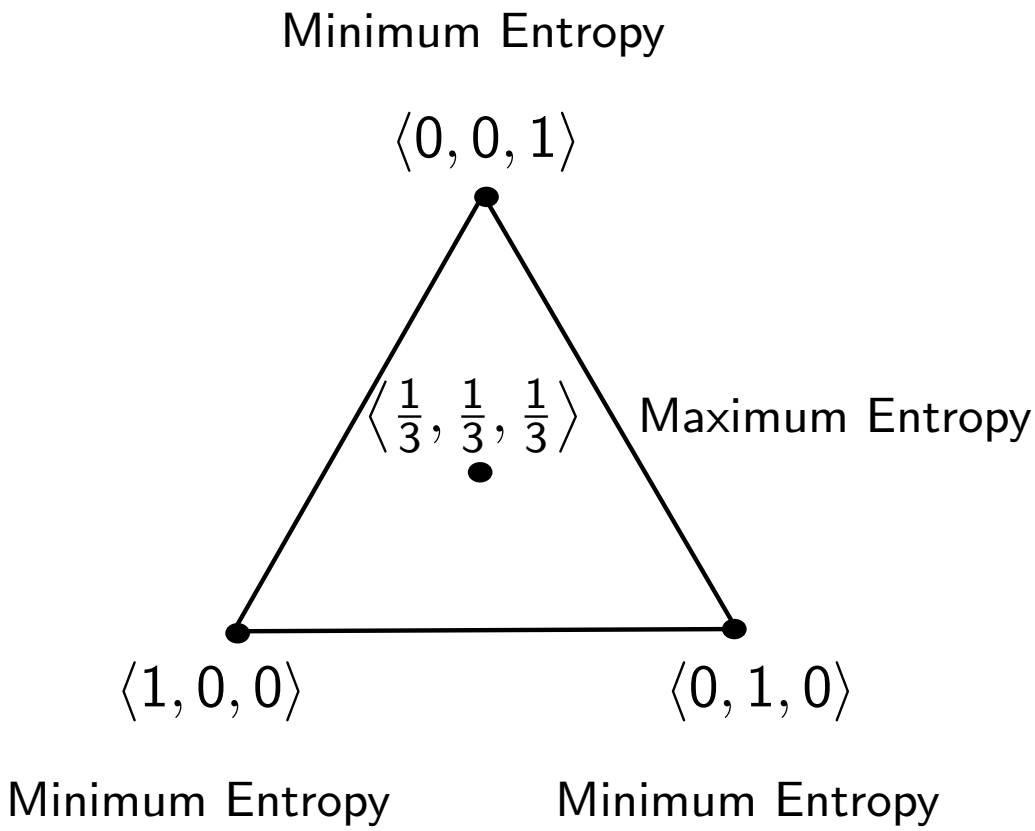
Shannon's Entropy Measure

- Let $W = \{w_1, \ldots, w_n\}$ and $P(w_i) = p_i$ then the information content of this distribution is:

$$H := -\sum_{i=1}^{n} p_i \log_2(p_i)$$

- H is minimal when $p_j = 1$ for some $j \in \{1, \ldots, n\}$ and $p_i = 0$ for all $i \neq j$.
- H is maximal when $p_i = \frac{1}{n}$ for $i = 1, \ldots, n$. i.e. the uniform probability distribution on W.
- Jaynes has argued that this provides some justification for the assumption of a uniform prior distribution.

Geometric Representation



ME Inference Process

- Select the probability distribution consistent with K which has maximum entropy.
- Alternatively select the element of $V(K)$ with the highest H value.
- **Idea:** Minimize the additional assumptions being made in the inference process beyond those explicitly contained in K.
- Since K is a linear set of contraints then $V(K)$ is convex. Therefore, H restricted to $V(K)$ has a unique maximum value.

Entropy Example

- Recall
  $V(K) = \{\langle p_1, 0.8 - p_1, 0.7 - p_1, p_1 - 0.5\rangle : 0.5 \leq p_1 \leq 0.7\}$
- *Ent* restricted to $V(K)$ is:

$$H = -p_1 \log_2(p_1) - (0.8 - p_1)\log_2(0.8 - p_1)$$
$$-(0.7 - p_1)\log_2(0.7 - p_1) + (p_1 - 0.5)\log_2(p_1 - 0.5)$$

- Taking the derivative with respect to p_1 gives:

$$\frac{\partial H}{\partial p_1} = -\log_2(p_1) - \log_2(p_1 - 0.5)$$
$$+\log_2(0.8 - p_1) + \log_2(0.7 - p_1)$$

- Setting $\frac{\partial H}{\partial p_1} = 0$ gives $p_1 = 0.56$ so that $P(F) = 0.62$.

Centre of Mass

- Assume that we can define a *second order* probability distribution on the probability distributions on W consistent with K.
- In other word, we give each element of $V(K)$ a probability and then take the expected value.
- Suppose we give every element of $V(K)$ equal probability (i.e. define a uniform second order prior) then we obtain the probability distribution

$$\hat{p}_i = \frac{\int_{V(K)} p_i dV(K)}{\int_{V(K)} dV(K)} \text{ for } i = 1, \ldots, n$$

- $\langle \hat{p}_1, \ldots, \hat{p}_n \rangle$ is the centre of mass of $V(K)$

CM Example

- For
  $V(K) = \{\langle p_1, 0.8 - p_1, 0.7 - p_1, p_1 - 0.5\rangle : 0.5 \leq p_1 \leq 0.7\}$
  we have that:

$$\hat{p_1} = \int_{V(K)} p_1 dV(K) = \frac{\int_{0.5}^{0.7} p_1 dp_1}{\int_{0.5}^{0.7} dp_1} = \frac{0.5 + 0.7}{2} = 0.6$$

- This gives $p_4 = 0.6 - 0.5 = 0.1$ and hence
  $P(F) = 0.6 + 0.1 = 0.7$

Example: CM vs ME

- Let W, HT and HP be as defined previously.
- Let K be any knowledge base which simply specifies the values of $P(HT)$ and $P(HP)$.
- Then the ME solution always gives $p_1 = P(HT \cap HP) = P(HT) \times P(HP)$.
- The CM solution always gives $P(HT \cap HP) = \frac{\max(0, P(HT)+P(HP)-1)+\min(P(HT),P(HP))}{2}$
- Notice that for any probability measure P on W $\max(0, P(HT) + P(HP) - 1) \leq P(HT \cap HP) \leq \min(P(HT), P(HP))$