Extending Maximum Entropy Techniques to Entropy Constraints

Gang Xiang[1] and Vladik Kreinovich[2]

[1]Philips Healthcare
El Paso, Texas 79902, USA
gxiang@sigmaxi.net

[2]Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
contact email vladik@utep.edu

1. Probabilities are Usually Imprecise: A Reminder

- Often, we have only *partial* (imprecise) information about the probabilities:
  - Sometimes, we have *crisp* (interval) bounds on probabilities (and/or other statistical characteristics).
  - Sometimes, we have *fuzzy* bounds, i.e., different interval bounds with different degrees of certainty.
- In this case, for each statistical characteristic, it is desirable to find:
  - the worst possible value of this characteristic,
  - the best possible value of this characteristic, and
  - the "typical" ("most probable") value of this characteristic.

2. Maximum Entropy (MaxEnt) Approach

- By the "typical" value of a characteristic, we mean its value for a "typical" distribution.
- Usually, as such a "typical" distribution, we select the one with the largest value of the *entropy* S.
- *Meaning:* $S =$ average # of "yes"-"no" questions (bits) that we need to ask to determine the exact value x_i.
- When we have n different values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$, the entropy $S(p)$ is defined as

$$S \stackrel{\text{def}}{=} -\sum_{i=1}^{n} p_i \cdot \log_2(p_i).$$

- For pdf $\rho(x)$, $S \stackrel{\text{def}}{=} -\int \rho(x) \cdot \log_2(\rho(x))\, dx$.
- S is related to the average number of questions needed to determine x with a given accuracy $\varepsilon > 0$.

3. MaxEnt Approach: Successes and Limitations

- *Successes:* when we know values of ranges and moments.
- *Example 1:* if we only know that $x \in [\underline{x}, \overline{x}]$, we get a uniform distribution on this interval.
- *Example 2:* if we only know the first 2 moments, we get a Gaussian distribution.
- *Problem:* sometimes, we also know the value S_0 of the entropy itself.
- *Why this is a problem:*
  - all distributions satisfying this constraint $S = S_0$ have the same entropy;
  - hence the MaxEnt approach cannot select a one.
- *What we do:* we show how to handle this constraint.

4. Main Idea and Its Consequences

- *Fact:* the actual *probabilities* $p_1, \ldots, p_n$ are only approximately equal to frequencies: $p_i \approx f_i$.
- *Idea:* instead of selecting "typical" probabilities, let us select "typical" frequencies.
- *Hence:* since $p_i \approx f_i$, we have $S(p_i) \approx S(f_i) = S_0$.
- *Idea:* select f_i *and* consistent p_i for which the entropy $S(p)$ is the largest possible.
- *Asymptotically:* each $\delta_i \stackrel{\text{def}}{=} p_i - f_i$ is normal, with mean 0 and $\sigma_i^2 = \dfrac{f_i \cdot (1 - f_i)}{N}$, where N denotes sample size.
- *Thus:* by χ^2, $\sum\limits_{i=1}^{n} \dfrac{\delta_i^2}{\sigma_i^2} = \sum\limits_{i=1}^{n} \dfrac{\delta_i^2}{f_i \cdot (1 - f_i)/N} \approx n.$
- *Resulting problem:* find f_i and p_i that maximize $S(p)$ under the above condition.

5. Analysis of the Problem

- *Problem:* under $\sum_{i=1}^{n} \frac{\delta_i^2}{f_i \cdot (1-f_i)/N} = \frac{n}{N}$, maximize $S(p) = S(f_1+\delta_1,\ldots,f_n+\delta_n) = -\sum_{i=1}^{n}(f_i+\delta_i)\cdot \log_2(f_i+\delta_i)$.
- *For large N:* δ_i are small, so

$$S(f_1+\delta_1,\ldots,f_n+\delta_n) = S(f_1,\ldots,f_n) + \sum_{i=1}^{n} \frac{\partial S}{\partial f_i}\cdot \delta_i.$$

- Here, $\frac{\partial S}{\partial f_i} = -\log_2(f_i) - \log_2(e)$, so Lagrange multiplier method leads to maximizing

$$S_0 - \sum_{i=1}^{n}(\log_2(f_i)+\log_2(e))\cdot \delta_i + \lambda \cdot \sum_{i=1}^{n} \frac{\delta_i^2}{f_i\cdot(1-f_i)}.$$

6. Analysis of the Problem (cont-d)

- *Reminder:* we maximize

$$S_0 - \sum_{i=1}^{n} (\log_2(f_i) + \log_2(e)) \cdot \delta_i + \lambda \cdot \sum_{i=1}^{n} \frac{\delta_i^2}{f_i \cdot (1 - f_i)}.$$

- *Analysis:* equating derivatives to 0, we get δ_i in terms of λ, then λ in terms of f_i, so the maximum is

$$S_2 \stackrel{\text{def}}{=} \sum_{i=1}^{n} (\log_2(f_i) + \log_2(e))^2 \cdot f_i \cdot (1 - f_i).$$

- *Result:*
  - if we have several distributions with the same values of the entropy S,
  - we should select the one with the largest value of the new characteristic S_2.

7. Continuous Case: Main Idea

- *Situation:* we have a continuous distribution, with pdf

$$\rho(x) \stackrel{\text{def}}{=} \lim_{\Delta x \to 0} \frac{p([x, x + \Delta x])}{\Delta x}.$$

- *Idea:*
  - we divide the interval of possible values of x into intervals $[x_i, x_i + \Delta x]$ of small width Δx;
  - we consider the discrete distribution with these intervals as possible values.
- *Fact:* when Δx is small, by the definition of the pdf, we have $p_i \approx \rho(x_i) \cdot \Delta x$.
- *Limit:* then, we take the limit $\Delta x \to 0$.
- *Example:* this is how we go from the discrete entropy $S(p_1, \ldots, p_n)$ to the entropy $S(\rho)$ of the continuous one.

8. How We Go From $S(p_1, \ldots, p_n)$ to $S(\rho)$: Reminder

- *Reminder:* $S \stackrel{\text{def}}{=} -\sum_{i=1}^{n} p_i \cdot \log_2(p_i)$.
- *Idea:* take $p_i = \rho(x_i) \cdot \Delta x$ and take a limit $\Delta x \to 0$:

$$S = -\sum_{i=1}^{n} \rho(x_i) \cdot \Delta x \cdot \log_2(\rho(x_i) \cdot \Delta x) =$$

$$-\sum_{i=1}^{n} \rho(x_i) \cdot \Delta x \cdot \log_2(\rho(x_i)) - \sum_{i=1}^{n} \rho(x_i) \cdot \Delta x \cdot \log_2(\Delta x).$$

- So, $S \sim -\int \rho(x) \cdot \log_2(\rho(x))\, dx - \log_2(\Delta x)$.
- *Fact:* the second term in this sum does not depend on the probability distribution at all.
- *Corollary:* maximizing the entropy S is equivalent to maximizing the integral in the above expression.
- *Obervation:* the integral $-\int \rho(x) \cdot \log_2(\rho(x))\, dx$ is exactly the entropy of the continuous distribution.

9. Continuous Analog of S_2

- *Reminder:* $S_2 \stackrel{\text{def}}{=} \sum_{i=1}^{n} (\log_2(f_i) + \log_2(e))^2 \cdot f_i \cdot (1 - f_i)$.
- *Idea:* take $f_i = \rho(x_i) \cdot \Delta x$ and take a limit $\Delta x \to 0$.
- *Asymptotically:* $S_2 = \int (\log_2(\rho(x)))^2 \cdot \rho(x)\, dx -$

  $2 \cdot (\log_2(\Delta x) + \log_2(e)) \cdot S + (\log_2(\Delta x) + \log_2(e))^2$.
- The 2nd and 3rd terms depend only on the step size Δx and on the entropy S – but not explicitly on $\rho(x)$.
- *Reminder:* we assume that S is known.
- *Corollary:* maximizing the value S_2 is equivalent to maximizing the integral in the above expression.
- *Conclusion:* select the distribution with the largest

$$S_2(\rho) \stackrel{\text{def}}{=} \int (\log_2(\rho(x)))^2 \cdot \rho(x)\, dx.$$

10. Meaning of Entropy S: Reminder

- *Idea:* the average number $E[q]$ of "yes"-"no" questions needed to locate x with a given accuracy $\varepsilon > 0$.
- *Simple case:* k alternatives, probabilities unknown.
- *Fact:* after q yes-no questions, we have 2^q combinations of answers $(a_1, \ldots, a_n)$, so $2^q \geq k$ and $q \geq \log_2(k)$.
- We can ask the questions about all b binary digits of $x = 1, \ldots, k$, so we need $q \leq b \approx \log_2(k)$ questions.
- For each interval of width ε, we have $p \approx \rho(x) \cdot \varepsilon$ hence $N \cdot p$ elements.
- To locate x, we locate a group of $N \cdot p$ elements out of N; there are $k = 1/p$ such groups.
- We need $q = \log_2(k) = -\log_2(p) = -\log(\rho(x) \cdot \varepsilon)$ questions.
- Thus, $E[q] = -\int \log(\rho(x) \cdot \varepsilon) \cdot \rho(x)\, dx = S - \log_2(\varepsilon)$.

11. Meaning of S_2

- S relates to the average number $E[q]$ of "yes"-"no" questions needed to locate x with accuracy $\varepsilon > 0$.
- *Our case:* all distribution have the same entropy.
- *Corollary:* they have the same mean $E[q]$.
- *Difference:* they may have different st. dev. $\sigma[q]$.
- *General idea:* possible values of q are in the interval $[E[q] - k_0 \cdot \sigma[q], E[q] - k_0 \cdot \sigma[q]]$, with $k_0 = 2, 3, 6$.
- *Corollary:* for fixed $E[q]$, the largest q is possible when $\sigma[q]$ is the largest.
- *Observation:* S_2 is indeed related to the st. dev. $\sigma[q]$ of the number q of "yes"-"no" questions:

$$S_2 \to \max \Leftrightarrow \sigma[q] \to \max\,.$$

12. Conclusions and Future Work

- In many practical situations, we have incomplete information about the probabilities.
- In this case, among all possible probability distributions, it is desirable to select the most "typical" one.
- Traditionally, we select the distribution which has the largest possible value of the entropy S.
- This approach has many successful applications, but it does not work when we know the entropy S_0.
- We show that in such situations, we should maximize a special characteristic $S_2 = \int (\log_2(\rho(x)))^2 \cdot \rho(x)\,dx$.
- Remaining open questions:
  - what if we know the values of S_2 as well?
  - how to extend S_2 to the cases of interval and fuzzy uncertainty?

13. Acknowledgment

- This work was also supported in part
  - by the National Science Foundation grants HRD-0734825 and DUE-0926721,
  - by Grant 1 T36 GM078000-01 from the National Institutes of Health, and
  - by the Science and Technology Centre in Ukraine (STCU) Grant 5015, funded by European Union.
- The authors are thankful to Ron Yager for valuable discussions.

Title Page
◀◀ ▶▶
◀ ▶
Page 14 of 14
Go Back
Full Screen
Close
Quit