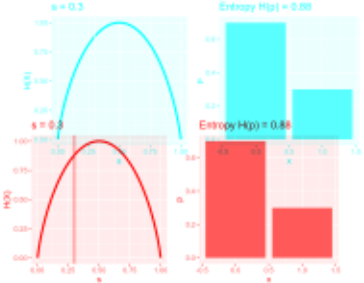
Introduction to Machine Learning

Information Theory

Entropy II



Learning goals

- Further properties of entropy and joint entropy
- Understand that uniqueness theorem justifies choice of entropy formula
- Maximum entropy principle

Entropy I

Learning goals

- Further properties of entropy and joint entropy
- Understand that uniqueness theorem justifies choice of entropy formula
- Maximum entropy principle

THE UNIQUENESS THEOREM

Khinchin, 1957 showed that the only family of functions satisfying

- $H(p)$ is continuous in probabilities $p(x)$
- adding or removing an event with $p(x) = 0$ does not change it
- is additive for independent RVs
- is maximal for a uniform distribution.

is of the following form:

$$H(p) = -\lambda \sum_{x \in \mathcal{X}} p(x) \log p(x)$$

where λ is a positive constant. Setting $\lambda = 1$ and using the binary logarithm gives us the Shannon entropy.

THE MAXIMUM ENTROPY PRINCIPLE

Assume we know M properties about a discrete distribution $p(x)$ on $\mathcal{X}$, stated as "moment conditions" for functions $g_m(\cdot)$ and scalars α_m:

$$\mathbb{E}[g_m(X)] = \sum_{x \in \mathcal{X}} g_m(x)p(x) = \alpha_m \text{ for } m = 0, \dots, M$$

Maximum entropy principle Jaynes 2003: Among all feasible distributions satisfying the constraints, choose the one with maximum entropy!

- Motivation: ensure no unwarranted assumptions on $p(x)$ are made beyond what we know.
- MEP follows similar logic to Occam's razor and principle of insufficient reason