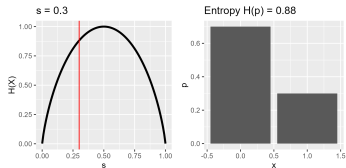


Introduction to Machine Learning

Information Theory

Entropy II



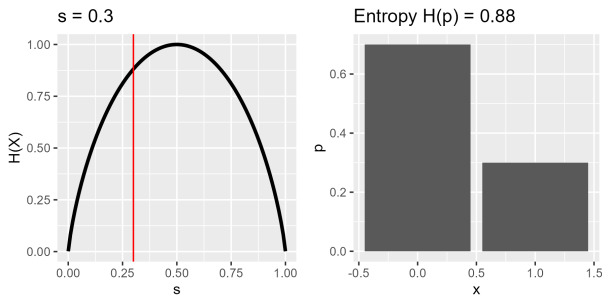
Learning goals

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

ENTROPY OF BERNOULLI DISTRIBUTION

Let X be Bernoulli / a coin with $\mathbb{P}(X = 1) = s$ and $\mathbb{P}(X = 0) = 1 - s$.

$$H(X) = -s \cdot \log_2(s) - (1 - s) \cdot \log_2(1 - s).$$



We note: If the coin is deterministic, so $s = 1$ or $s = 0$, then $H(s) = 0$; $H(s)$ is maximal for $s = 0.5$, a fair coin. $H(s)$ increases monotonically the closer we get to $s = 0.5$. This all seems plausible.

JOINT ENTROPY

- The **joint entropy** of two discrete random variables X and Y is:

$$H(X, Y) = H(p_{X,Y}) = - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2(p(x, y))$$

- Intuitively, the joint entropy is a measure of the total uncertainty in the two variables X and Y . In other words, it is simply the entropy of the joint distribution $p(x, y)$.
- There is nothing really new in this definition because $H(X, Y)$ can be considered to be a single vector-valued random variable.
- More generally:

$$H(X_1, X_2, \dots, X_n) = - \sum_{x_1 \in \mathcal{X}_1} \dots \sum_{x_n \in \mathcal{X}_n} p(x_1, x_2, \dots, x_n) \log_2(p(x_1, x_2, \dots, x_n))$$



ENTROPY IS ADDITIVE UNDER INDEPENDENCE

- 7 Entropy is additive for independent RVs.

Let X and Y be two independent RVs. Then:

$$\begin{aligned}H(X, Y) &= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log_2(p(x, y)) \\&= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_X(x) p_Y(y) \log_2(p_X(x) p_Y(y)) \\&= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_X(x) p_Y(y) \log_2(p_X(x)) + p_X(x) p_Y(y) \log_2(p_Y(y)) \\&= - \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_X(x) p_Y(y) \log_2(p_X(x)) - \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p_X(x) p_Y(y) \log_2(p_Y(y)) \\&= - \sum_{x \in \mathcal{X}} p_X(x) \log_2(p_X(x)) - \sum_{y \in \mathcal{Y}} p_Y(y) \log_2(p_Y(y)) = H(X) + H(Y)\end{aligned}$$



THE UNIQUENESS THEOREM

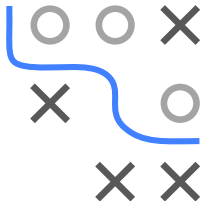
► Khinchin 1957 showed that the only family of functions satisfying

- $H(p)$ 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

Can be solved via Lagrangian multipliers (here with base e)

$$L(p(x), (\lambda_m)_{m=0}^M) = - \sum_{x \in \mathcal{X}} p(x) \log(p(x)) + \lambda_0 \left(\sum_{x \in \mathcal{X}} p(x) - 1 \right) + \sum_{m=1}^M \lambda_m \left(\sum_{x \in \mathcal{X}} g_m(x) p(x) - \alpha_m \right)$$

Finding critical points $p^*(x)$:

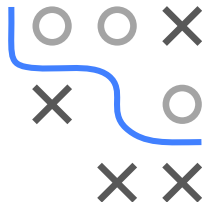
$$\frac{\partial L}{\partial p(x)} = -\log(p(x)) - 1 + \lambda_0 + \sum_{m=1}^M \lambda_m g_m(x) \stackrel{!}{=} 0 \iff p^*(x) = \exp(\lambda_0 - 1) \exp\left(\sum_{m=1}^M \lambda_m g_m(x)\right)$$

This is a maximum as $-1/p(x) < 0$. Since probs must sum to 1 we get

$$1 \stackrel{!}{=} \sum_{x \in \mathcal{X}} p^*(x) = \frac{1}{\exp(1 - \lambda_0)} \sum_{x \in \mathcal{X}} \exp\left(\sum_{m=1}^M \lambda_m g_m(x)\right) \Rightarrow \exp(1 - \lambda_0) = \sum_{x \in \mathcal{X}} \exp\left(\sum_{m=1}^M \lambda_m g_m(x)\right)$$

Plugging $\exp(1 - \lambda_0)$ into $p^*(x)$ we obtain the constrained maxent distribution:

$$p^*(x) = \frac{\exp \sum_{m=1}^M \lambda_m g_m(x)}{\sum_{x \in \mathcal{X}} \exp \sum_{m=1}^M \lambda_m g_m(x)}$$



THE MAXIMUM ENTROPY PRINCIPLE

We now have: functional form of our distribution, up to M unknowns, the λ_m . But also: M equations, the moment conditions. So we can solve.

Example: Consider discrete RV representing a six-sided die roll and the moment condition $\mathbb{E}(X) = 4.8$. What is the maxent distribution?

- Condition means $g_1(x) = x$, $\alpha_1 = 4.8$. Then for some λ solution is

$$p^*(x) = \frac{\exp(\lambda g(x))}{\sum_{j=1}^6 \exp(\lambda g(x_j))} = \frac{\exp(\lambda x)}{\sum_{j=1}^6 \exp(\lambda x_j)}$$

- Inserting into moment condition and solving (numerically) for λ :

$$4.8 \stackrel{!}{=} \sum_{j=1}^6 x_j p^*(x_j) = \frac{e^\lambda + \dots + 6(e^\lambda)^6}{e^\lambda + \dots + (e^\lambda)^6} \Rightarrow \lambda \approx 0.5141$$

x	1	2	3	4	5	6
$p^*(x)$	3.22%	5.38%	9.01%	15.06%	25.19%	42.13%

