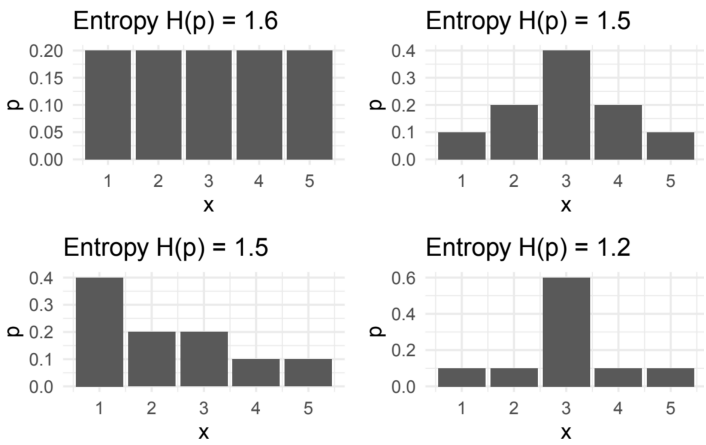
Introduction to Machine Learning

Entropy



Learning goals

- Know that the entropy expresses expected information for discrete RVs
- Entropy for a single RV
- Joint entropy for two RVs
- Understand that the uniqueness theorem justifies the choice for the formula of the entropy

INFORMATION THEORY

- **Information Theory** is a field of study based on probability theory.
- The foundation of the field was laid by Claude Shannon in 1948 and it has since found applications in areas as diverse as communication theory, computer science, optimization, cryptography, machine learning and statistical inference.
- In addition to quantifying information, it also deals with efficiently storing and transmitting the information.
- Information theory tries to quantify the "amount" of information gained or uncertainty reduced when a random variable is observed.

INFORMATION THEORY

- We introduce the basic concepts from a probabilistic perspective, without referring too much to communication, channels or coding.
- We will show some proofs, but not for everything. We recommend *Elements of Information Theory* by Cover and Thomas as a reference for more.
- The application of information theory to the concepts of statistics and ML can sometimes be confusing, we will try to make the connection as clear as possible.

ENTROPY

- We develop in this unit entropy as a measure of uncertainty.
- Entropy is often introduced in IT as a measure of expected information or in terms of bits needed for efficient coding, but for us in stats and ML the first type of intuition seems most useful.

For a discrete random variable X with domain $x \in \mathcal{X}$ and pmf $p(x)$:

$$H(X) := H(p) = -\mathbb{E}[\log_2(p(X))] = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

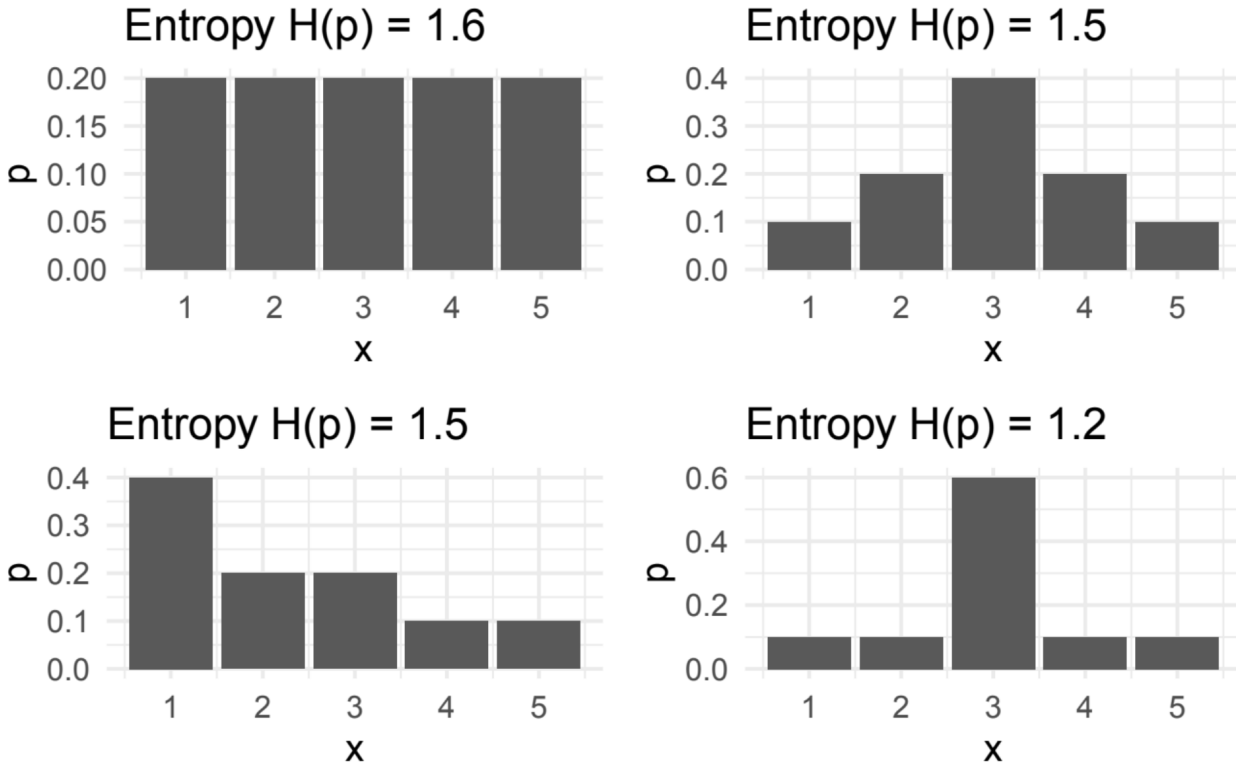
$$= \mathbb{E}\left[\log_2\left(\frac{1}{p(X)}\right)\right] = \sum_{x \in \mathcal{X}} p(x) \log_2 \frac{1}{p(x)}$$

- **Definition:** Base 2 means the information is measured in bits, but you can use any number > 1 as base of the logarithm.
- **Note:** Because $\lim_{p \to 0} p \log_2 p = 0$, if $p(x) = 0$, $p(x) \log_2 p(x)$ is taken to be zero for $x = 0$.

ENTROPY

- We have encountered various other measures of spread or uncertainty in statistics before.
- Furthermore, the formula does not seem to make intuitive sense?
- We will now do the following:
  - Approach it by various simple examples.
  - Note some basic properties - which seem desirable for an uncertainty measure.
  - Note that if we require these measures axiomatically, entropy drops out automatically as the result.
  - Note even more nice properties.

ENTROPY EXAMPLES

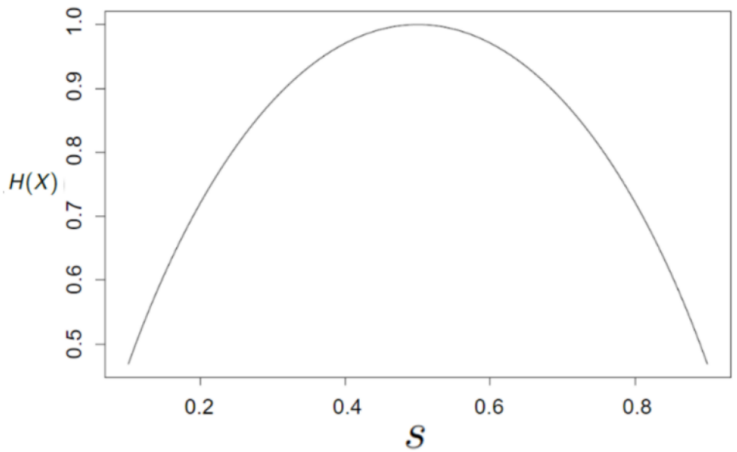


Naive observations: Uniform seems maximal, re-ordering does not matter, and the more peaked, the less entropy.

ENTROPY OF BERNOULLI DISTRIBUTION

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

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

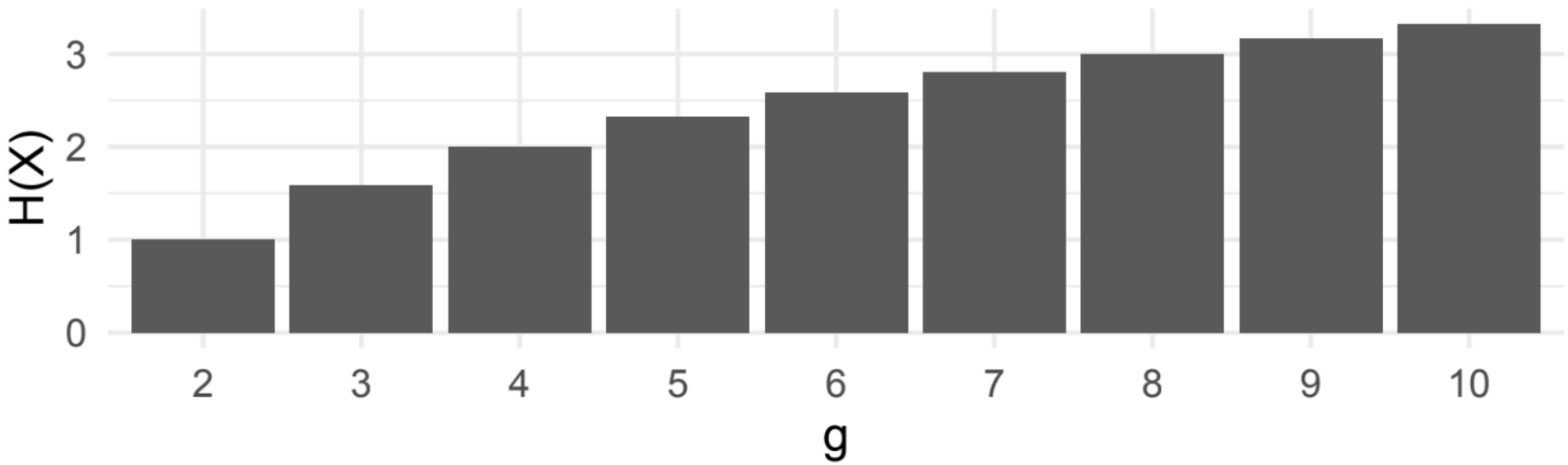


We note: If the coin is deterministic, so $s = 1$ or $s = 0$, then $H(s) = 0$; $H(s)$ is maximal for $s = 0.5$, a fair coin. $H(s)$ becomes monotonically larger the closer we get to $s = 0.5$. This all seems plausible.

ENTROPY OF UNIFORM DISTRIBUTIONS

Let X be a uniform, discrete RV with g outcomes (g-sided fair die).

$$H(X) = -\sum_{i=1}^{g} \frac{1}{g} \log_2 \left(\frac{1}{g} \right) = \log_2 g$$



The more sides a die has, the harder it is to predict the outcome. Unpredictability grows *monotonically* with the number of potential outcomes.

PROPERTIES OF DISCRETE ENTROPY

1. Entropy is non-negative, so $H(X) >= 0$.
2. If one event has probability $p(x) = 1$, then $H(X) = 0$.
3. Symmetry. If the values $p(x)$ in the pmf are re-ordered, entropy does not change.
4. Adding or removing an event with $p(x) = 0$ does not change entropy.
5. $H(X)$ is continuous in probabilities $p(x)$.
6. Entropy is additive for independent RVs.
7. Entropy is maximal for a uniform distribution, so for a domain with g elements: $H(X) \leq -g\frac{1}{g}\log_2(\frac{1}{g}) = log_2(g)$.

All properties except the last 2 follow trivially from the definition.

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:

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

ENTROPY IS ADDITIVE UNDER INDEPENDENCE

Let X and Y be two independent RVs. Then:

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

ENTROPY OF UNIFORM DISTRIBUTIONS

Claim: The entropy of a discrete random variable X which takes on values in $\{x_1, x_2, \ldots, x_g\}$ with associated probabilities $\{p_1, p_2, \ldots, p_g\}$ is maximal when the distribution over X is uniform.

Proof: The entropy $H(X)$ is $-\sum_{i=1}^{g} p_i \log_2 p_i$ and our goal is to find:

$$\underset{p_1,p_2,\ldots,p_g}{\operatorname{argmax}} \; -\sum_{i=1}^{g} p_i \log_2 p_i$$

subject to

$$\sum_{i=1}^{g} p_i = 1.$$

ENTROPY OF UNIFORM DISTRIBUTIONS

The Lagrangian $L(p_1, \ldots, p_g, \lambda)$ is :

$$L(p_1, \ldots, p_g, \lambda) = -\sum_{i=1}^{g} p_i \log_2(p_i) - \lambda \left(\sum_{i=1}^{g} p_i - 1 \right)$$

Solving for $\nabla L = 0$,

$$\frac{\partial L(p_1, \ldots, p_g, \lambda)}{\partial p_i} = 0 = -\log_2(p_i) - 1 - \lambda$$

$$\implies p_i = 2^{(-1-\lambda)} \implies p_i = \frac{1}{g},$$

where the last step follows from the fact that all p_i are equal and the constraint.

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.