Electrical and Computer Engineering Department
Tarbiat Modares University

Probability and Information Theory

Foad Ghaderi, PhD

Probability Theory

Probability theory is a fundamental tool of many disciplines of science and engineering.

What is the use case of probability theory in artificial intelligence applications?

- The laws of probability tell us how AI systems should reason, so we design our algorithms to compute or approximate various expressions derived using probability theory.
- We can use probability and statistics to theoretically analyze the behavior of proposed AI systems.

Probability Theory

Why Probability?

Entities:

- Deterministic
- Stochastic

Possible sources of uncertainty

- Inherent stochasticity in the system (Card game)
- Incomplete observability (Monty Hall problem)
- Incomplete modeling (Discretized space for robot exploration)

Probability Theory

Why Probability?

- In many cases, it is more practical to use a simple but uncertain rule rather than a complex but certain one, even if the true rule is deterministic and our modeling system has the fidelity to accommodate a complex rule.

Probability

- Frequentist (drawing a certain hand of cards in a game of poker)
- Bayesian (a doctor says that the patient has a 40% chance of having the flu)

  we use probability to represent a *degree of belief*

Random Variables

A *random variable* is a variable that can take on different values randomly.

Discrete random variable
has a finite or countably infinite number of states
Continuous random variable
associated with a real value.

On its own, a random variable is just a description of the states that are possible; it must be coupled with a probability distribution that specifies how likely each of these states are.

Probability Distributions

We typically denote the random variable itself with a lower case letter in plain typeface, and the values it can take on with lower case script letters. For example, x_1 and x_2 are both possible values that the random variable x can take on.

For vector-valued variables, we would write the random variable as $\mathbf{x}$ and one of its values as $\boldsymbol{x}$.

Probability Distributions

A ***probability distribution*** is a description of how likely a random variable or set of random variables is to take on each of its possible states.

Discrete Variables and Probability Mass Functions

A probability distribution over discrete variables may be described using a *probability mass function* (PMF).

Probability Distributions

Discrete Variables and Probability Mass Functions

A function P must satisfy the following properties:

- The domain of P must be the set of all possible states of x.
- $\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1$. An impossible event has probability 0 and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no state can have a greater chance of occurring.
- $\sum_{x \in \mathrm{x}} P(x) = 1$. We refer to this property as being *normalized*. Without this property, we could obtain probabilities greater than one by computing the probability of one of many events occurring.

Probability Distributions

Discrete Variables and Probability Mass Functions

Example: Uniform distribution

$$P(\mathrm{x} = x_i) = \frac{1}{k}$$

Probability Distributions

Continuous Variables and Probability Density Functions

When working with continuous random variables, we describe probability distributions using a *probability density function (PDF).*

A function p must satisfy the following properties:

- The domain of p must be the set of all possible states of x.
- $\forall x \in \mathrm{x}$, $p(x) \geq 0$. Note that we do not require $p(x) \leq 1$.
- $\int p(x)dx = 1$

Probability Distributions

Continuous Variables and Probability Density Functions

A probability density function $p(x)$ does not give the probability of a specific state directly, instead the probability of landing inside an infinitesimal region with volume δx is given by $p(x)\delta x$.

Example: Uniform distribution

$$u(x; a, b) = \frac{1}{b - a}$$

Marginal Probability

The probability distribution over the subset is known as the *marginal probability* distribution.

Example, suppose we have discrete random variables x and y, and we know *P*(x, y). We can find P(x) with the *sum rule*:

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y)$$

$$p(x) = \int p(x, y) dy$$

Conditional Probability

The probability of some event, given that some other event has happened

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}$$

The conditional probability is only defined when $P(\mathrm{x} = x) > 0$.

The Chain Rule of Conditional Probabilities

Any joint probability distribution over many random variables may be decomposed into conditional distributions over only one variable:

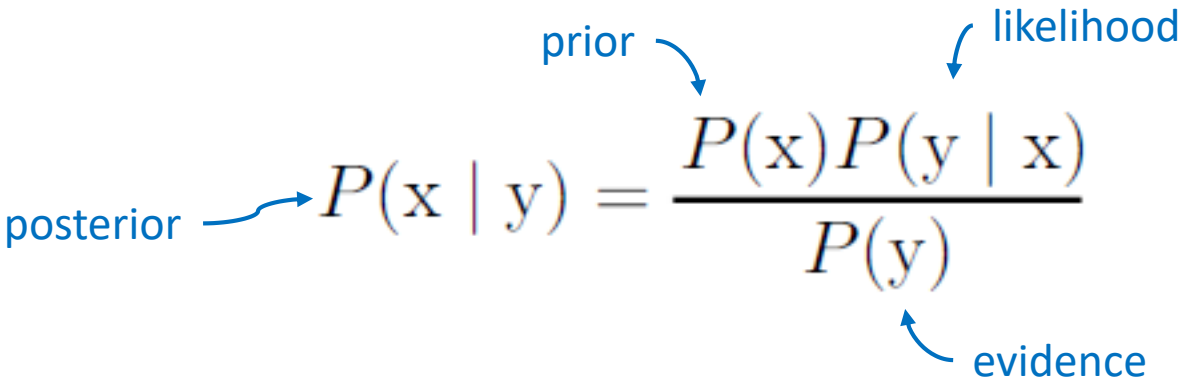
$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)})$$

Example:

$$\begin{aligned} P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b}, \mathrm{c}) \\ P(\mathrm{b}, \mathrm{c}) &= P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}) \\ P(\mathrm{a}, \mathrm{b}, \mathrm{c}) &= P(\mathrm{a} \mid \mathrm{b}, \mathrm{c}) P(\mathrm{b} \mid \mathrm{c}) P(\mathrm{c}) \end{aligned}$$

Bayes' Rule

We often find ourselves in a situation where we know $P(\mathrm{y} \mid \mathrm{x})$ and need to know $P(\mathrm{x} \mid \mathrm{y})$.



$$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}$$

$$P(\mathrm{y}) = \sum_x P(\mathrm{y} \mid x)P(x)$$

Information Theory

A branch of applied mathematics that revolves around quantifying how much information is present in a signal.

Learning that an unlikely event has occurred is more informative than learning that a likely event has occurred.

- "the sun rose this morning"
- "there was a solar eclipse this morning"

We would like to quantify information in a way that formalizes this intuition. Specifically,

- Likely events should have low information content.
- Less likely events should have higher information content.
- Independent events should have additive information.

Information Theory

Shannon entropy

We define the *self-information* of an event x = *x* to be $I(x) = -\log P(x)$. (unit is nat: One nat is the amount of information gained by observing an event of probability *1/e*.)

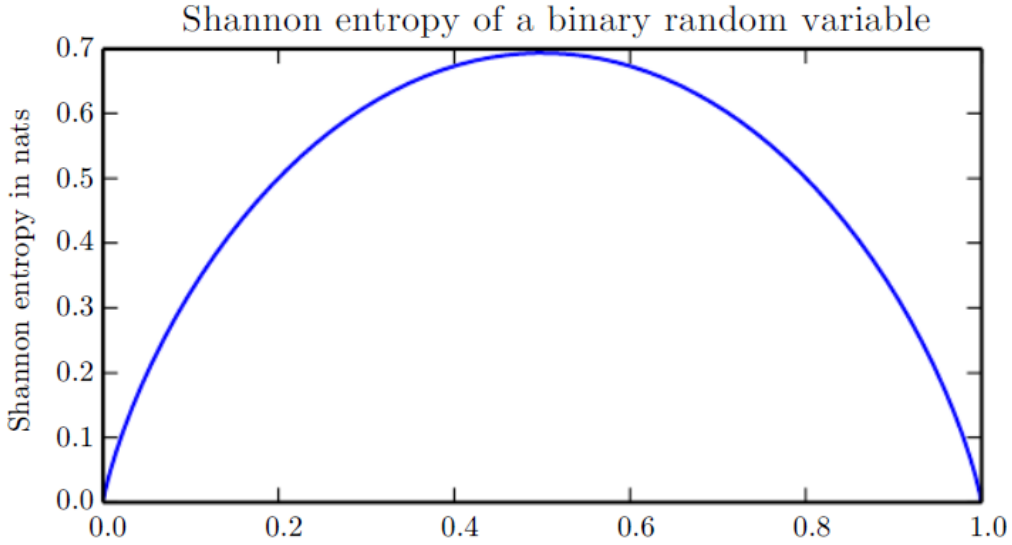
The amount of uncertainty in an entire probability distribution can be quantified using the *Shannon entropy:*

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)]$$

Information Theory

Shannon entropy

$$H(\mathrm{x}) = \mathbb{E}_{\mathbf{x}\sim P}[I(x)] = -\mathbb{E}_{\mathbf{x}\sim P}[\log P(x)]$$



p: the probability of a binary random variable being equal to 1.

The entropy is given by -(1-*p*) log(1−*p*)−*p* log *p*.

p is near 0, the distribution is nearly deterministic, because the random variable is nearly always 0.

p is near 1, the distribution is nearly deterministic, because the random variable is nearly always 1.

p = 0.5, the entropy is maximal, because the distribution is uniform over the two outcomes.

Information Theory

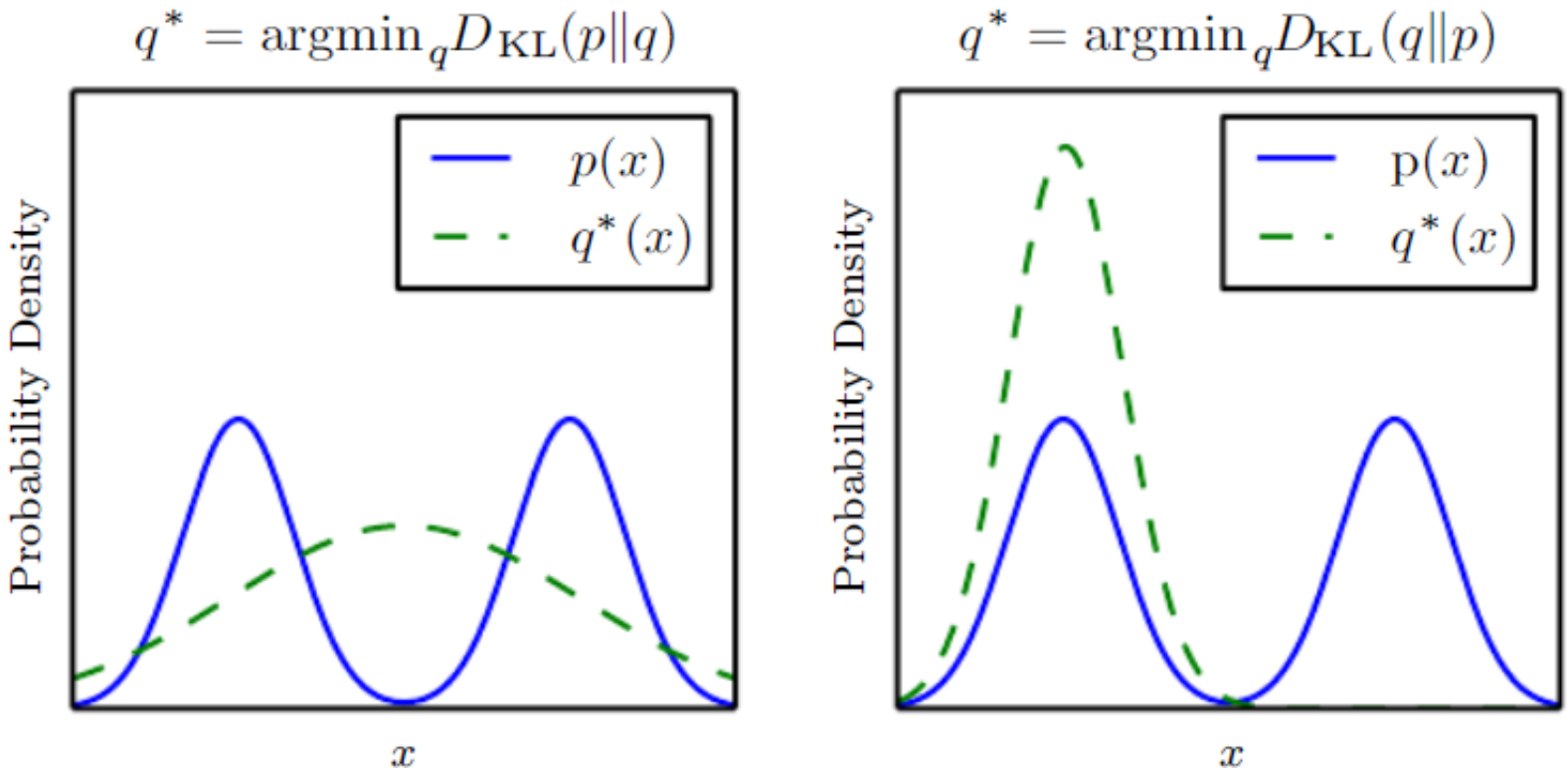
Kullback-Leibler (KL) divergence

If we have two separate probability distributions $P(\mathrm{x})$ and $Q(\mathrm{x})$ over the same random variable x , we can measure how different these two distributions are using the *Kullback-Leibler* (*KL*) *divergence*:

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right]$$

Information Theory

The KL divergence is asymmetric



Some applications require an approximation that usually places high probability anywhere that the true distribution places high probability,
while other applications require an approximation that rarely places high probability anywhere that the true distribution places low probability.