2 — Probability, Entropy, and Inference

Table of contents

2.1 Probabilities and Ensembles

An **ensemble** X is a framework for a random variable, defined by a set of possible outcomes, or alphabet, $\mathcal{A}_X = \{a_1, a_2, ..., a_I\}$, and their corresponding probabilities, $\mathcal{P}_X = \{p_1, p_2, ..., p_I\}$. These probabilities must be non-negative ($p_i \geq 0$) and sum to unity ($\sum P(x = a_i) = 1$). The probability of a subset of outcomes T is the sum of the probabilities of all outcomes within that subset.

A **joint ensemble** XY involves outcomes that are ordered pairs (x, y), described by a **joint probability** $P(x, y)$. From this, a **marginal probability** $P(x)$ can be obtained by summing over all possibilities of y: $P(x) = \sum_y P(x, y)$.

The **conditional probability** $P(x|y)$ represents the probability of outcome x given that y has occurred. It is defined as:

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

This definition is valid only if $P(y) \neq 0$. These concepts lead to three foundational rules of probability theory:

1. **The Product Rule:** This rule, also known as the chain rule, restates the definition of conditional probability: $P(x, y) = P(x|y)P(y) = P(y|x)P(x)$.
2. **The Sum Rule:** This rule shows how to find a marginal probability from conditional probabilities: $P(x) = \sum_y P(x, y) = \sum_y P(x|y)P(y)$.
3. **Bayes' Theorem:** This theorem is derived directly from the product rule and is essential for inference. It inverts the conditional probability:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

   The denominator $P(x)$ can be expanded using the sum rule, $P(x) = \sum_{y'} P(x|y')P(y')$.

Finally, two variables X and Y are **independent** if and only if their joint probability is the simple product of their marginal probabilities: $P(x, y) = P(x)P(y)$.

2.2 The Meaning of Probability

Probability can be interpreted in two main ways. The **frequentist** interpretation confines probability to describing the long-run frequencies of outcomes in repeatable random experiments.

The **Bayesian** interpretation is more general, using probability to describe a **degree of belief** in any proposition, even one that is not a repeatable random variable. Examples include the probability that a particular suspect is guilty or that a historical event occurred. This is also known as the subjective interpretation, as the probabilities depend on assumptions. The **Cox axioms** provide a formal foundation for this view, demonstrating that if degrees of belief satisfy basic consistency rules, they can be mapped directly to the laws of probability.

2.3 Forward and Inverse Probabilities

Probability calculations can be categorized as either forward or inverse. * **Forward probability** problems involve a **generative model** where the parameters are known, and the goal is to calculate the probability distribution of the data or outcomes that result from the process. * **Inverse probability** problems reverse this: given the observed data, the goal is to infer the conditional probability of the unobserved variables or parameters of the generative model.

Inverse probability problems invariably require the use of **Bayes' theorem**]. In this context, the terms in Bayes' theorem are given specific names:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

- **Prior Probability** $P(\theta)$: This is the initial probability or degree of belief in the parameter θ *before* any data is observed.
- **Likelihood** $P(D|\theta)$: This is the probability of the observed data D given a specific value of the parameter θ. It is treated as a function of the parameter θ, *not* as a probability distribution over θ.
- **Posterior Probability** $P(\theta|D)$: This is the updated probability of the parameter θ *after* observing the data D.
- **Evidence** $P(D)$: This is the marginal probability of the data, $P(D) = \sum_{\theta'} P(D|\theta')P(\theta')$, which serves as a normalization constant.

When using this framework to make predictions, the correct Bayesian approach is not to simply choose the single most plausible hypothesis. Instead, one must **marginalize** (use the sum rule) over all possible values of the parameter, weighting each one by its posterior probability.

This approach adheres to the **likelihood principle**, which states that all inferences should depend *only* on the likelihood function $P(d_1|\theta)$ for the data d_1 that was *actually observed*. The probabilities of other outcomes that *could have* happened but did not are irrelevant to the inference.

2.4 Definition of Entropy and Related Functions

The **Shannon information content** $h(x)$ of an outcome x quantifies its surprisal and is defined as:

$$h(x) = \log_2 \frac{1}{P(x)}$$

This quantity is measured in **bits**. An improbable outcome (low $P(x)$) has high information content, while a common outcome (high $P(x)$) has low information content].

The **entropy** $H(X)$ of the entire ensemble is the *average* or *expected value* of the Shannon information content. It is also known as the uncertainty of the ensemble.

$$H(X) = \sum_{x \in \mathcal{A}_X} P(x) \log_2 \frac{1}{P(x)}$$

, 479]

Entropy is always non-negative ($H(X) \geq 0$) and is at its maximum when the probability distribution $P(x)$ is uniform. The **joint entropy** $H(X, Y)$ is calculated using the same formula but with the joint probability $P(x, y)$. Entropy is additive for independent variables: $H(X, Y) = H(X) + H(Y)$ if and only if X and Y are independent.

2.5 Decomposability of the Entropy

The entropy of an ensemble can be computed recursively by decomposing the process into stages. The total entropy is the sum of the entropies of each stage, where the entropy of any subsequent stage is weighted by the probability that that stage is even reached.

For example, consider an ensemble with probabilities $\mathbf{p} = \{1/2, 1/4, 1/4\}$. This can be viewed as a two-stage process: 1. A first choice between the first outcome and the other two, with probabilities $\{1/2, 1/2\}$. The entropy of this stage is $H(1/2, 1/2) = 1$ bit. 2. *If* the first outcome is not chosen (which happens with probability $1/2$), a second choice is made between the remaining two outcomes, which have conditional probabilities $\{1/2, 1/2\}$. The entropy of this stage is also $H(1/2, 1/2) = 1$ bit.

The total entropy is the sum of the entropy from stage 1 and the weighted entropy from stage 2: $H(X) = 1 + (1/2) \times 1 = 1.5$ bits.

2.6 Gibbs' Inequality

The **relative entropy**, or **Kullback-Leibler (KL) divergence** $D_{KL}(P||Q)$, measures the inefficiency of assuming a distribution is Q when the true distribution is P. It is defined as:

$$D_{KL}(P||Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)}$$

Gibbs' inequality states that the relative entropy is always non-negative:

$$D_{KL}(P||Q) \geq 0$$

Equality holds if and only if $P(x) = Q(x)$ for all x]. The KL divergence is not a true distance measure because it is not symmetric; that is, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$.

2.7 Jensen's Inequality for Convex Functions

A function $f(x)$ is **convex** (or "convex-smile") if every chord drawn between two points on the function's graph lies on or above the function itself. Examples include x^2, e^x, and $x \log x$.

Jensen's inequality states that for a convex function f and a random variable x, the expectation of the function is greater than or equal to the function of the expectation:

$$E[f(x)] \geq f(E[x])$$

If the function is **concave** (or "concave-frown"), the inequality is reversed: $E[f(x)] \leq f(E[x])$. For strictly convex functions, equality holds only if the random variable x is a constant.