Chapter 1

Probability, Inference, and Learning

1.1 Introduction

In this chapter, we will study the role of probability in inference, codifying relationships, and machine learning. When considering these problems, we deal with uncertainty, and that's were probability comes in. In other words, we are interested in probability because it allows us to model uncertainty (or equivalently, belief and knowledge). Sources of uncertainty, for example in machine learning, include:

- Noise: aggregate contribution of factors that we do not (wish to) consider (models focus on the most important quantities).
- Finite sample size: finite size of data makes it impossible to determine relationships (i.e., probability distributions) as some configuration may never happen or happen few times in finite data.

1.2 Relationships and joint probability distributions

Is there any relationship between the arrival times of two people working at a business (opening at 9:00 am), both living in the same area? If so, how can we represent this relationship? How can we make prediction about one being late given the other is late (e.g., if we need at least one person be present)?

In the same way that we can encode our information about a random quantity as a distribution, we can encode information about random quantities, as well as their relationships, as joint distributions.

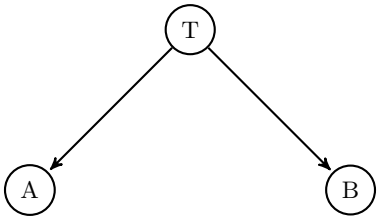
In our example, there's obviously a relationship, that is, the arrival times are not independent. For example, both are affected by traffic. Let

$$
\begin{aligned}
&T_0 : \text{normal traffic}\\
&T_1 : \text{heavy traffic}\\
&A_0 : \text{Alice is on time}\\
&A_1 : \text{Alice is late}\\
&B_0, B_1 \text{ for Bob}
\end{aligned}
$$

and assume

$$
\begin{aligned}
\Pr(T_0) &= 0.65,\\
\Pr(A_0|T_0) &= 0.9,\\
\Pr(B_0|T_0) &= 0.82,\\
\Pr(A_0|T_1) &= 0.5,\\
\Pr(B_0|T_1) &= 0.15.
\end{aligned}
$$

Finally, conditioned on the traffic situation, Alice and Bob's arrival times are independent. This information completely determines all probabilities. As we will see in much grater depth later, the fact that the Alice and Bob's arrival times are only related through traffic can be shown *graphically* as



Causal reasoning:

$$\Pr(A_0) = \Pr(T_0)\Pr(A_0|T_0) + \Pr(T_1)\Pr(A_0|T_1) = (0.65 \times 0.9) + (0.35 \times 0.5) = 0.76$$
$$\Pr(B_0) = \Pr(T_0)\Pr(B_0|T_0) + \Pr(T_1)\Pr(B_0|T_1) = (0.65 \times 0.82) + (0.35 \times 0.15) = 0.5855$$

Evidential reasoning (inverse probabilities, uses Bayes rule):

$$\Pr(T_0|A_0) = \Pr(A_0|T_0)\Pr(T_0)/\Pr(A_0) = 0.65 \times 0.9/0.76 = 0.7697$$
$$\Pr(T_0|B_0) = \Pr(B_0|T_0)\Pr(T_0)/\Pr(B_0) = 0.65 \times 0.82/0.5855 = 0.9103$$

The common cause makes the events A_i and B_i dependent. Recall that two events E_1 and E_2 are independent, denoted $E_1 \perp\!\!\!\perp E_2$ if $\Pr(E_1E_2) = \Pr(E_1)\Pr(E_2)$, or, if $\Pr(E_2) \neq 0$, $\Pr(E_1|E_2) = \Pr(E_1)$. We have

$$\Pr(A_0|B_0) = \Pr(A_0B_0)/\Pr(B_0)$$
$$\Pr(A_0B_0) = (0.65 \times 0.82 \times 0.9) + (0.35 \times 0.15 \times 0.5) = 0.506$$
$$\Pr(A_0|B_0) = 0.506/0.586 = 0.863 \neq \Pr(A_0)$$
$$\Pr(B_0|A_0) = 0.506/0.76 = 0.6658 \neq \Pr(B_0)$$

So $A_0 \not\perp\!\!\!\perp B_0$.

However, they are *conditionally independent*, by assumption

$$\Pr(A_0B_0|T_0) = \Pr(A_0|T_0)\Pr(B_0|T_0),$$

which is denoted as $A_0 \perp\!\!\!\perp B_0|T_0$.

What is the source of uncertainty in this problem? Since we have assumed the distribution is known, finite sample size is not an issue. The source is noise. For example, if we had information about other factors affecting Bob, e.g., how reliable his car is, if he needs to drop off his kids, etc., we could reduce the amount of noise and make better predictions.

1.3 Inference and decision making

Let us consider a problem about **inferring** unknown values and making decisions and use probability to solve it, using both frequentist and Bayesian views. Suppose that the probability that someone with a given allele of a gene will develop a certain disease is θ. We are interested in determining θ. In particular, we may be interested in comparing this with the fraction of people in the general population with that disease, say 0.01. Different interpretations lead to different approaches to problems. But to decide, both frequentists and Bayesians need data.

Data ($\mathcal{D}$): Among a sample of 100 people with this allele, 2 had the disease.

- A Frequentist thinks of θ as unknown non-random parameter. She starts by asking "What is the probability of the observation as a function of θ?" We can view each of the 100 people chosen to be an independent Bernoulli trial with probability θ. So the distribution is Binomial and the probability of the observation as a function of θ is

$$L(\theta) = \binom{100}{2}\theta^2(1-\theta)^{98}.$$

  Probability of the observation as a function of the parameter is called the *likelihood function*. So what value for θ makes the most sense? Since the observation has actually happened, we would expect it to have a high probability so we find θ that maximizes the likelihood. This method is called *maximum likelihood estimation*, and we'll discuss it in much more detail later. In this case, we estimate θ to be

$$\hat{\theta} = \arg\max_{\theta} L(\theta) = \frac{2}{100},$$

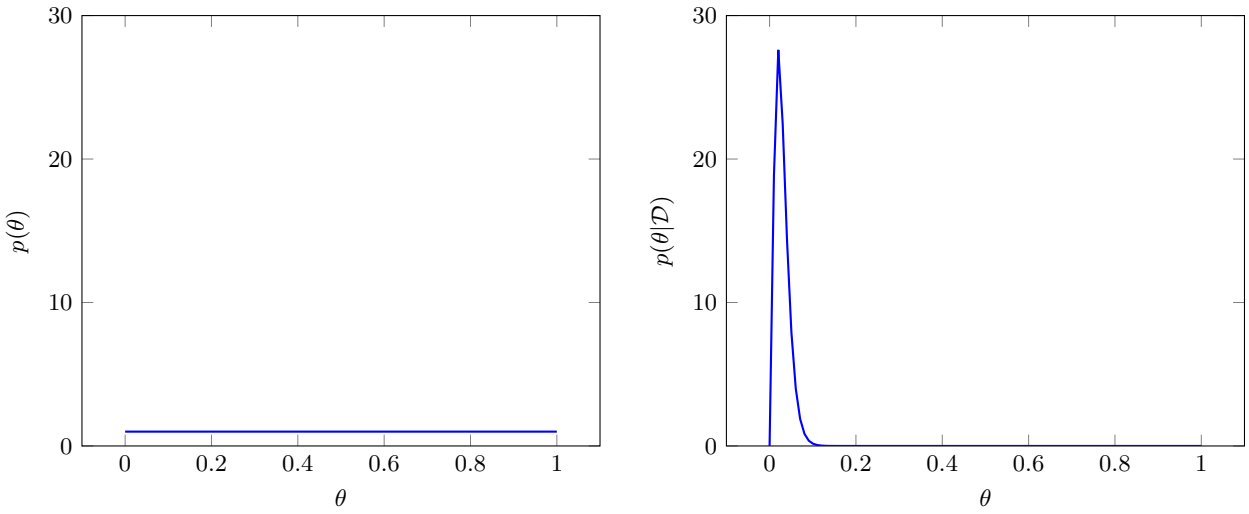
  which is a reasonable estimate. But how close is the estimate to the true value? For frequentists, this is a tricky question to answer probabilistically since the true value and the estimate are both deterministic at this point. With some clever reasoning (some would say mental gymnastics), frequentists come up with *confidence intervals* and *confidence levels* to quantify the accuracy of estimators.

- A Bayesian thinks of θ as random and assigns to it a distribution, called the *prior*, before seeing the data. Thinking of θ as random is imaginative (some would say questionable) since there is no repeatable experiment and there is a single value that is true. One way to justify randomness of θ is to think of our universe being drawn from a set of possible universes. Regardless, the Bayesian view is used widely in practice.

  Our Bayesian statistician then looks at the data and updates her distribution for θ, thus obtaining the *posterior* distribution. Assume that before seeing the data, we believe that the distribution for θ is uniform, i.e., $p(\theta) \sim \text{Uni}[0,1] = \text{Beta}(1,1)$. This means that while we do not know what θ is, we believe it is equally likely to be any value between 0 and 1. When we see the data, we can update this belief,

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \qquad \text{(Bayes' rule)}$$

  It turns out $p(\theta|\mathcal{D}) \sim \text{Beta}(3, 99)$.



In contrast to the frequentist view, the Bayesian view is consistent and flexible. For example, we can show that

$$p(\theta > 0.01|\mathcal{D}) = 0.92.$$

What is the source of uncertainty in this problem? It is the finite sample size. If we know the status of a very large number of people with the allele, we would know the distribution/ the value of θ.

1.4 Machine Learning and Probability

Let us consider the generic form of supervised machine learning problems, which have the following components:

- **Data**: $\mathcal{D} = \{(x_1, y_1), ..., (x_N, y_N)\}, x_i \in \mathcal{X}, y_i \in \mathcal{Y}$. $\mathcal{X}$ is called the feature space, and $\mathcal{Y}$ is called the label space. As an example, each x_i could be a vector providing information about a house, e.g., (location, lot size, square footage, number of bedrooms, ...), and y_i can be the sale price of the house.
- **Assumption**: (x_i, y_i) are iid samples of random variables X and Y. The joint distribution (X, Y) is (partially) unknown.
- **Goal**: Find the "best" function f to predict y corresponding to a given x. In other words, the function f produces an estimate $\hat{y} = f(x)$ of y given data x. Continuing our example, y would be the true but unknown price of the house with features x, and $f(x)$ would be a prediction (similar to what Zillow does).
- **Evaluation**: How do we define "*best*"? For a given data point (x, y), evaluate the success of f using a loss function $L(y, f(x))$, e.g., $L(y, f(x)) = |y - f(x)|$. Ideally, we would like to minimize the expected loss over all possible outcomes weighted by their probabilities, so we define

  $$\mathcal{L}(f) = \mathbb{E}[L(Y, f(X))], \tag{1.1}$$

  also known as the **population risk**, where the expectation is over the distribution $p(x, y)$ of (X, Y). Our goal then becomes finding

  $$f^{**} = \arg\min_f \mathcal{L}(f) = \arg\min_f \mathbb{E}[L(Y, f(X))]. \tag{1.2}$$

- **Learning Algorithm**: The algorithm that finds f^{**}, or tries to.

The expectation in (1.2) is computed using the joint distribution $p(x, y)$. Here is where we face our main machine learning challenge: ***What we have is the data set $\mathcal{D}$ consisting of samples from $p(x, y)$, but what we need to find f^{**} is the joint distribution $p(x, y)$.*** We can address this mismatch in two ways, either through the Empirical Risk Minimization framework discussed in §1.4.1, or through estimating the unknown distribution $p(x, y)$ using $\mathcal{D}$ as discussed in §1.4.2.

Before proceeding further, let us consider two common problems in supervised learning:

- **Regression**: $\mathcal{Y}$ consists of **scalars or vectors of reals**. For example, predicting stock price based on financial information, or determining the score someone will assign a movie based on previous scores. A common loss function is the **quadratic** or **squared error** loss function:

  $$L(y, f(x)) = (y - f(x))^2. \tag{1.3}$$

  It can be shown that for this loss, *if the distribution is known*,

  $$f^{**}(x) = \mathbb{E}[Y|X = x]. \tag{1.4}$$

- **Classification**: $\mathcal{Y}$ consists of **classes or categories**. For example, speech recognition, hand writing recognition, the presence or absence of a disease. A common loss function is the **0-1 loss**:

  $$L(y, f(x)) = \begin{cases} 1, & \text{if } y \neq f(x). \\ 0, & \text{if } y = f(x). \end{cases} \tag{1.5}$$

  In this case, *if the distribution is known*, then the best classifier is

  $$f^{**}(x) = \arg\max_{y \in \mathcal{Y}} p(y|x). \tag{1.6}$$

We emphasize again that to solve the problem optimally as in (1.4) and (1.6), we need to know the joint distribution of x and y or the conditional distribution of y given x.

1.4.1 Empirical Risk Minimization (ERM)

Since we usually do not know the distribution but have access to data $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, we cannot directly minimize the expected loss as in (1.2). Instead we can minimize the **empirical risk**, i.e., the loss on observed data points,

$$f^{**} = \arg\min_f \mathbb{E}[L(Y, f(X)] \quad \rightarrow \quad f_N^{**} = \arg\min_f \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)). \tag{1.7}$$

So instead of the best possible solution based on the distribution, f^{**}, we should try to find f_N^{**} based on N data points. But finding f_N^{**} is still problematic, as it only provides a way for us to determine the value of $f(x)$ for $x \in \{x_1, \ldots, x_N\}$. In other words, it is not able to extrapolate or generalize.

A common solution, which is also helpful from a practical point of view, is to restrict the choices for f to a set $\mathcal{H}$, called the **hypothesis set**. This leads to the ERM formulation of the learning problem

$$f^* = \arg\min_{f \in \mathcal{H}} \mathbb{E}[L(Y, f(X)] \quad \rightarrow \quad f_N^* = \arg\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)). \tag{1.8}$$

For example, we may choose $\mathcal{H}$ to be the set of linear or sigmoid functions. By restricting predictors to the hypothesis set $\mathcal{H}$, we have introduced our prior knowledge, or *bias* towards the learning task.

1.4.2 Density estimation

As mentioned, distribution estimation, aka density estimation, is another way to use data for prediction. Here we discuss only *parametric density estimation*, where we can (or choose to) represent the joint distribution of (X, Y) using a probabilistic model with some unknown parameters, for example, a graphical model with known structure and unknown parameters. There are also nonparametric ways of estimating distributions.

Let us consider maximum likelihood, which is one method for parameter estimation. Suppose the distribution has a set of unknown parameters θ and we represent the distribution as p_θ. So what should we choose as the value of θ? If an outcome has a small probability, the chance it appears in our dataset $\mathcal{D}$ is small. So those outcomes observed in $\mathcal{D}$ must have large probability. Hence, we must choose θ such that the probability assigned to $\mathcal{D}$ is large, that is,

$$\begin{aligned}
\hat{\theta} &= \arg\max_\theta p_\theta(\mathcal{D}) \\
&= \arg\max_\theta \prod_{i=1}^{N} p_\theta(x_i, y_i) \\
&= \arg\max_\theta \sum_{i=1}^{N} \log p_\theta(x_i, y_i),
\end{aligned}$$

where in the last step, we use the monotonously of the log function to convert the product to a simpler-to-deal-with summation. We'll cover this in more detail later. For now, let us assume we can find $\hat{\theta}$, and in turn, $p_{\hat{\theta}}(x, y)$ as our estimate of the joint distribution $p(x, y)$.

With $p_{\hat{\theta}}(x, y)$ in hand, we can solve (1.2) as

$$\hat{f}_N = \arg\min_f \mathbb{E}_{\hat{\theta}}[L(Y, f(X)],$$

where $\mathbb{E}_{\hat{\theta}}$ is expectation computed using the estimated distribution $p_{\hat{\theta}}$. As we have seen in (1.4), for quadratic and 0-1 losses, we respectively have

$$\hat{f}_N(x) = \mathbb{E}_{\hat{\theta}}[Y|X = x],$$
$$\hat{f}_N(x) = \arg\max_{y \in \mathcal{Y}} p_{\hat{\theta}}(y|x).$$

1.5 Information theory and machine learning

Information theory deals with quantifying information and the rules that govern its transmission, storage, and transformation from one form to another. It has applications in communications, data storage, machine learning, and biology. In machine learning it can be used to help better understand relationships between knowns and unknowns, design loss functions, and establish fundamental limits on how well we can do with a certain amount of data (regardless of the type of algorithm and computational resources).

1.5.1 Quantifying uncertainty

Let X be a Bernoulli random variable that is equal to 1 if it is raining in Seattle and 0 otherwise. Similarly, let Y be indicate whether its raining in Phoenix. How much information do X and Y provide us? Alternatively, before they are revealed, how uncertain are we about X and about Y? Can we measure the information content of a random variable, or equivalently, our uncertainty about them.

Let's look at specific outcomes for each variable:

- $X = 1$: It's raining in Seattle. This is a statement with a fair amount of information as rain in Seattle is almost 50/50.
- $Y = 0$: It's not raining in Phoenix. This statement doesn't provide a lot of information as this outcome is expected and has a high probability.
- $Y = 1$: It's raining in Phoenix. This provides a lot of information as this outcomes is unlikely and surprising.

So as a function of probability, the amount of information of a given statement decreases as the probability increases. If the probability of an outcome is p, what is a good function describing the amount of information we gain from learning that the outcome has occurred? It turns out a good choice is $I(p) = \log \frac{1}{p}$, which is called the *self-information* function and shown in Figure 1.1 when the base of the log is 2. Then the information content of the statement '$X = x_i$' is

$$I(p(x_i)) = \log \frac{1}{p(x_i)}.$$

And the amount of information *on average* for a random variable X that takes values in the set $\mathcal{X} = \{x_1, \ldots, x_m\}$ is

$$H(X) = \mathbb{E}\left[\log \frac{1}{p(X)}\right] = \sum_{i=1}^{m} p(x_i) \log \frac{1}{p(x_i)},$$

where for continuous RVs, the sum must be replaced with an integral. This is called the *entropy*. If the log is base 2, then the unit is a *bit*.

If there are m different possible outcomes, then the maximum value that entropy can take is $\log m$. So

$$0 \leq H(X) \leq \log m.$$

An important special case is the binary entropy function $H_b(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$ for experiments

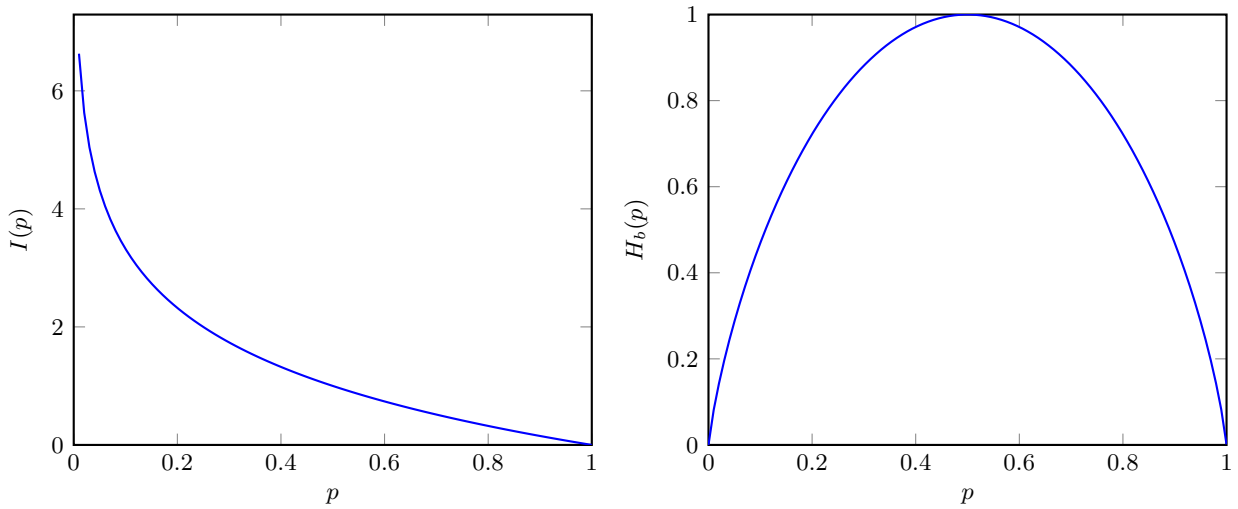


Figure 1.1: Self-information (left) for an event with probability p and binary entropy (right) for a Bernoulli RV with probability of success equal to p.

with two outcomes with probabilities p and $1-p$. For example,

$$H(\text{Fair coin}) = H_b\left(\frac{1}{2}\right) = 1,$$
$$H(\text{6 on a die}) = H_b\left(\frac{1}{6}\right) = 0.65,$$
$$H(\text{Rainy day in Seattle}) = H_b\left(\frac{150}{365}\right) = 0.977,$$
$$H(\text{Rainy day in Phoenix}) = H_b\left(\frac{33}{365}\right) = 0.43784,$$
$$H(\text{Rainy day in the Sahara}) = H_b\left(\frac{1}{365}\right) = 00.027267.$$

The plot for binary entropy is given in Figure 1.1. The maximum entropy is 1 bit. This makes sense since we can represent the outcome with 1 bit. Random variables with equal chances of 0 and 1 have the highest entropy (and maximum uncertainty). Those with predictable outcomes have lower entropies.

Entropy was introduced by Shannon in his article "A mathematical theory of communication" in 1948. It is also the minimum amount of "bandwidth" you need to transmit the outcome of the experiment. He also popularized the term *bit* (Binary digit).

> "My greatest concern was what to call it. I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage." – Claude Shannon, Scientific American (1971), volume 225, page 180.

1.5.2 Relative entropy

Let X be a random variable with set of possible values denoted as $\mathcal{X}$ and its distribution as $p(x)$. Let q be another distribution also over $\mathcal{X}$. For example, let X be a random Latin letter with p given by the

English letter frequencies and q by the French letter frequencies. For example, we have $p(E) = 12.6\%$ and $q(E) = 15.1\%$.

The *relative entropy*, or the *Kullback–Leibler divergence*, between two distributions p and q is defined as

$$D_{KL}(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}. \tag{1.9}$$

The divergence can be viewed as the difference between the entropy of X when self-information is computed based on an approximate distribution and when it is based on the "true" distribution since

$$\begin{aligned} D_{KL}(p||q) &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)} - \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)} - H(X). \end{aligned}$$

Relative entropy provides a measure of difference between two distributions. It is always non-negative and equals 0 if and only if $q = p$. In machine learning, it is used to measure how good our estimated distribution q is to the true distribution p. It is not symmetric, so $D_{KL}(p||q)$ is not necessarily equal to $D_{KL}(q||p)$.

A related quantity is cross-entropy, which is also used as a loss function,

$$H(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)}.$$

So $D_{KL}(p||q) = H(p||q) + H(X)$.

1.5.3 Conditional entropy and mutual entropy*

We can also measure the information in multiple random variables using entropy. The information in both X and Y is denoted $H(X, Y)$ and is defined as

$$H(X, Y) = \mathbb{E}\left[\log \frac{1}{p(X, Y)}\right] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)}.$$

If we know Y, how much information is left in X? This is denoted $H(X|Y)$. If, for example $X = Y + 2$, then $H(X|Y) = 0$ since if we know Y, we also know X. Conditional entropy is defined as

$$H(X|Y) = \sum_{y \in \mathcal{Y}} p(y) H(X|Y = y) = E\left[\log \frac{1}{p(X|Y)}\right] = H(X, Y) - H(Y)$$

Mutual information, $I(X; Y)$, represents the amount of information that one random variable has about the other, and is defined as

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X).$$

While this quick overview is sufficient for our purposes in this course, if you are interested, you can check out the slides for this Short Lecture on Information Theory, or the course Mathematics of Information.