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

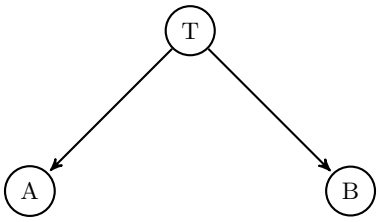
example, both are affected by traffic. Let

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

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:

$$\begin{aligned} \Pr(A_0) &= \Pr(T_0)\Pr(A_0|T_0) + \Pr(T_1)\Pr(A_0|T_1) = (0.65 \times 0.9) + (0.35 \times 0.5) = 0.76 \\ \Pr(B_0) &= \Pr(T_0)\Pr(B_0|T_0) + \Pr(T_1)\Pr(B_0|T_1) = (0.65 \times 0.82) + (0.35 \times 0.15) = 0.5855 \end{aligned}$$

Evidential reasoning (inverse probabilities, uses Bayes rule):

$$\begin{aligned} \Pr(T_0|A_0) &= \Pr(A_0|T_0)\Pr(T_0)/\Pr(A_0) = 0.65 \times 0.9/0.76 = 0.7697 \\ \Pr(T_0|B_0) &= \Pr(B_0|T_0)\Pr(T_0)/\Pr(B_0) = 0.65 \times 0.82/0.5855 = 0.9103 \end{aligned}$$

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

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

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

However, they are *conditionally independent*, by assumption

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

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 and we are interested to know if $\theta > 0.01$, where 0.01 is the fraction of people in the general population with that disease. 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 devises statistical tests to decide if $\theta > 0.01$. Clearly, 2 out of 100 is larger than would be expected by chance. So this may be because the allele and the disease are related. On the other hand, maybe the allele doesn't have anything to do with the disease, but we have been unlucky enough to pick two people with the disease. So how do we decide?

  Our statistician may consider how likely it is to see *similar or stronger evidence by chance*. This probability is called the *p-value*.

  If the probability of the disease is 0.01, what is the probability of seeing 2 or more sick people in a sample of size 100?

  $$p = 1 - \left(\binom{100}{0} 0.99^{100} + \binom{100}{1} 0.99^{99} 0.01^1 \right) = 1 - 0.37 - 0.37 = 0.26 > 0.05$$

  The smaller the p-value, the stronger the evidence. Typically, if the p-value is smaller than 0.05, we believe the evidence is strong enough to reject the hypothesis that the observation has occurred by chance.

- A Bayesian thinks of θ as random and assigns to it a distribution, called the *prior*, before seeing the data. She then looks at the data and updates her distribution for θ, thus obtaining the *posterior* distribution. (We'll learn more about Bayesian methods.)

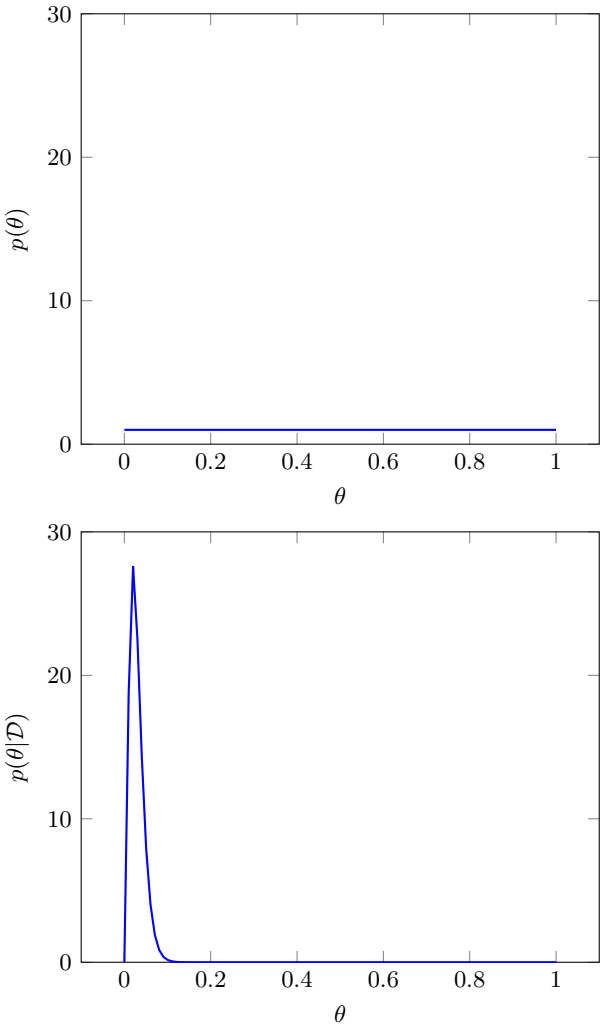
  Assume that before seeing the data, we believe that the distribution for θ is uniform, i.e., $p(\theta) \sim \text{Uni}[0, 1] = \text{Beta}(1, 1)$. This means that while we do not know what θ is, we believe it

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 $p(\theta|\mathcal{D}) \sim \text{Beta}(3, 99)$, and, as we will see,

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



What is the source of uncertainty in this case? Why can't we say for certain if $\theta > 0.01$? This is because of 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 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
$$\mathcal{L}(f) = \mathbb{E}[L(Y, f(X)], \tag{1.1}$$
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.

You may have noticed that $\mathcal{D}$ consists of samples from $p(x, y)$, but in (1.2), we need the joint distribution of X, Y. We can address this in two ways, either through the Empirical Risk Minimization framework discussed in §1.4.1, or through estimating the unknown distribution 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:
$$L(y, f(x)) = (y - f(x))^2. \tag{1.3}$$
For this choice, if the distribution is known, it can be shown that
$$\hat{y} = f(x) = \mathbb{E}[Y|X = x]. \tag{1.4}$$
- **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 $\hat{y} = \arg\max_{y \in \mathcal{Y}} p(y|x)$.

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 loss on observed data points,

$$f^* = \arg\min_f \mathbb{E}[L(Y, f(X)] \quad \rightarrow \quad f^* = \arg\min_f \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)). \tag{1.6}$$

This is, however, 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, called the **hypothesis set**. This leads to the ERM formulation of the learning problem

$$f^* = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)). \tag{1.7}$$

For example, we may choose $\mathcal{F}$ to be the set of linear or sigmoid functions.

The choice of $\mathcal{F}$ is critical to how well the predictor generalizes. On the one hand, it needs to be large enough to be able to produce a small loss. As an extreme example, setting $\mathcal{F}$ to contain only $f(x) = 0$ for all x is not a good choice. On the other hand, if $\mathcal{F}$ has too many degrees of freedom, we may get a predictor f that is tuned well to the dataset but does not generalize well, i.e., performs poorly for examples outside of the dataset. This is called **overfitting**. We check whether this is the case by setting aside part of the data, referred to as the **test set**, which is used only for evaluating performance but not for training. Data used for training is called the **training set**. (If we need to choose between different algorithms or tune hyper-parameters, we may further divide the training set to training and validation sets.)

1.4.2 Density estimation

As discussed, 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.

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,

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

Alternatively, we can formulate the problem as density estimation with maximum-likelihood loss to begin with. From the following equation, loss is minimized when log-likelihood is maximized.

$$\begin{aligned} L(P_\theta(x,y)) &= -\log P_\theta(x,y) \\ \mathcal{L}(\theta) &= -\mathbb{E}(\log\left(P_\theta(X,Y)\right) \end{aligned}$$

Again, before determining θ, we do not know the distribution and cannot evaluate the expected loss. So we minimize the empirical risk:

$$\begin{aligned} \mathcal{L}(\theta) &= -\sum_{i=1}^{n} \log P_\theta(x_i, y_i) \\ \hat{\theta} &= \arg\min_{\theta\in\Theta} \mathcal{L}(\theta), \end{aligned}$$

where Θ is the set of all valid parameters.

1.4.3 Decomposition of error for mean squared error

In (1.4), we claimed that for mean squared error and known joint distribution, the best predictor for Y given $X = x$ is $\mathbb{E}[Y|X = x]$. We start by proving this claim. First, let us consider: What is the best predictor for a (random) quantity Y when we know the distribution of Y but have no other information. Since we have no information, this predictor is a single constant value c and for the mean squared error we have

$$\begin{aligned} \mathbb{E}\left[(Y-c)^2\right] &= \mathbb{E}\Big[(Y-\mu+\mu-c)^2\Big] \\ &= \mathrm{Var}(Y) + 2\,\mathbb{E}[Y-\mu](\mu-c) + (\mu-c)^2 \\ &= \mathrm{Var}(Y) + (\mu-c)^2, \end{aligned}$$

where $\mu = \mathbb{E}[Y]$. This is minimized by letting $c = \mu = \mathbb{E}[Y]$.

Now let us consider the original problem: What is the best predictor $f(x)$ for Y if we know $X = x$ as well as the joint distribution of (X, Y)? Let $\bar{y}(x) = \mathbb{E}[Y|X = x]$. For the mean squared error for a given value of x, we have

$$\begin{aligned} \mathbb{E}\left[(Y-f(x))^2|X=x\right] &= \mathbb{E}\Big[(Y-\bar{y}(x)+\bar{y}(x)-f(x))^2|X=x\Big] \\ &= \mathbb{E}\Big[(Y-\bar{y}(x))^2|X=x\Big] + 2\,\mathbb{E}[Y-\bar{y}(x)|X=x](\bar{y}(x)-f(x)) + (\bar{y}(x)-f(x))^2 \\ &= \mathbb{E}\Big[(Y-\bar{y}(x))^2|X=x\Big] + (\bar{y}(x)-f(x))^2. \end{aligned}$$

Note that the error has two parts: an irreducible part, referred to as intrinsic error, which is not under our control, and a part that depends on the choice of the predictor. The intrinsic error results from the noise in our model and not lack of enough data. The reducible part, and thus the error, is minimized by setting $f(x) = \bar{y}(x) = \mathbb{E}[Y|X = x]$. However, doing so exactly is only possible if we have the distribution or an infinite amount of data. When f is determined based on a finite sample $\mathcal{D}$, the term $(\bar{y}(x) - f(x))^2$ can be decomposed into bias and variance components, which we will discuss later.

1.5 Quantifying uncertainty

Suppose we know the distribution for a random variable. How do we measure how uncertain we are? Alternatively how much information will we gain when we find out the outcome or how surprised will we be when we see the outcome?

First, we observe that the lower the probability of a statement, the higher the surprise/information content:

- The sun will rise tomorrow: Very likely, low information content
- It's raining in Seattle: Even chances, provides some information
- It's raining in the Sahara: Very unlikely, high information content

So we look for a function that decreases as the probability p of the event increases. 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* is

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

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

$$\begin{aligned} H(\text{Fair coin}) &= H_b(\frac{1}{2}) = \frac{1}{2}\log 2 + \frac{1}{2}\log 2 = 1, \\ H(\text{Sun coming up or not}) &= H_b(2^{-64}) = 2^{-64}\log 2^{64} + (1-2^{-64}) \log \frac{1}{1-2^{-64}} \\ &\simeq 65 \times 2^{-64} \simeq 2^{-58} \end{aligned}$$

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.

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).

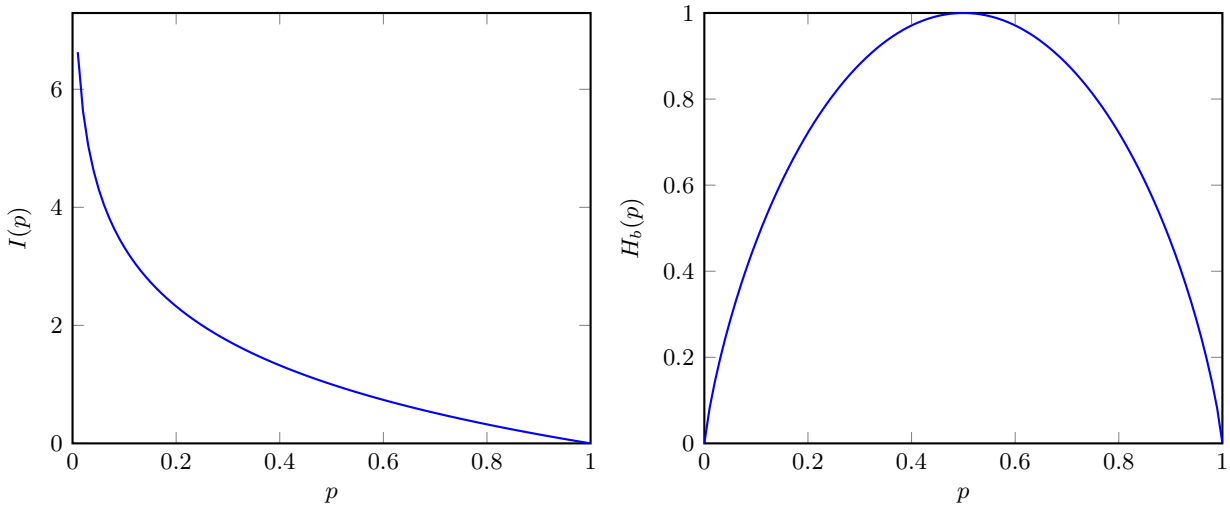


Figure 1.1: Self-information (left) for an event with probability p and binary entropy (right) for a Bernoulli RV with probability of success equal to p.

> "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.

1.6 Conditional entropy*

We can measure the information in multiple random variables also 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

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

Finally, relative entropy between two distributions p and q is defined as

$$KL(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)},$$

which can be viewed as a measure of difference between distributions.

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.