Chapter 12

Information Theory

So far in this course, we have discussed notions such as "complexity" and "information", but we have not grounded these ideas in formal detail. In this part of the course, we will introduce key concepts from the area of *information theory*, such as formal definitions of information and entropy, as well as how these definitions relate to concepts previously discussed in the course.

As we will see, many of the ideas introduced in previous chapters—such as the notion of maximum likelihood—can be re-framed through the lens of information theory. Moreover, we will discuss how machine learning methods can be derived from a information-theoretic perspective, based on the idea of maximizing the amount of information gain from training data.

In this chapter, we begin with key definitions from information theory, as well as how they relate to previous concepts in this course. In this next chapter, we will introduce a new supervised learning technique—termed decision trees—which can be derived from an information-theoretic perspective.

12.1 Entropy and Information

Suppose we have a discrete random variable x. The key intuition of information theory is that we want to quantify how much information this random variable conveys. In particular, we want to know how much information we obtain when we observe a particular value for this random variable. This idea is often motivated via the notion of *surprise*: the more surprising an observation is, the more information it contains. Put in another way, if we observe a random event that is completely predictable, then we do not gain any information, since we could already predict the result. On the other hand, if we observe a random event that is unpredictable and surprising, then we gain a substantial amount of information, since we could not predict the event beforehand.

These ideas can be formalized by considering the underlying distribution $P(x)$ over the discrete random variable x. Without loss of generality, we can assume that the support of this distribution is a set of integers $\mathcal{X} \subseteq \mathbb{Z}$ We can

then characterize the information $h(x = k)$ obtained by sampling a specific value $k \in \mathcal{X}$ from this distribution by taking the (base-2) logarithm of the probability of this event:

$$h(x = k) = -\log(P(x = k)), \tag{12.1}$$

where we take the negative since probabilities will always be in $[0, 1]$. In other words, the information we gain when we sample a specific integer k from this distribution is proportional to the (negative) log-probability of this event happening. The logarithm here is a natural choice, since it guarantees that the information obtained by sampling two independent random variables x and y is additive, i.e., for independent events we have that

$$P(x = k_1, y = k_2) = P(x = k_1)P(y = k_2), \tag{12.2}$$

which implies that the information content is

$$\log P(x = k_1, y = k_2) = -\log(P(x = k_1)) - \log(P(y = k_2)). \tag{12.3}$$

Thus, the negative logarithm of the probability of a *single* event gives us the amount of information that single event. We can then generalize this to the compute the amount of information in the random variable (i.e., the full distribution) by considering the *expected* information that we will receive by observing this variable.

$$H(x) = \mathbb{E}[h(x)] = -\sum_{k \in \mathcal{Z}} P(x = k) \log(P(x = k)). \tag{12.4}$$

The term $H(x)$ is often known as the *entropy*. Note that $\lim_{z \to 0} z \log(z) = 0$ so we can handle cases where the probability of an event is 0 by assigning such events zero information.

Some examples of entropy As a first example, suppose we have a random variable x that takes four possible values $\mathcal{X} = \{1, 2, 3, 4\}$, each with equal probability (i.e., $P(x = k) = \frac{1}{4}$ for all events). In this case, we would have that

$$H(x) = -4\frac{1}{4}\log\left(\frac{1}{4}\right) = 2. \tag{12.5}$$

As a second example, we could consider a case where the events have non-uniform probabilities, given by $\{\frac{1}{8}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}\}$. In this case, we would have that

$$H(x) = -2\frac{1}{8}\log\left(\frac{1}{8}\right) - \frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{1}{2}\log\left(\frac{1}{2}\right) = 1.75. \tag{12.6}$$

In this case, we see that the second distribution, which is non-uniform with a sharper peak contains less information. Typically, distributions that have sharper peaks have lower entropy, i.e., they are more predicable. On the other hand, distributions that are relatively flat or uniform have higher entropy.

12.1.1 Entropy and encoding

One useful interpretation of entropy is that it gives a lower bound on the expected number of bits needed to transmit the state of a random variable. For example, if we consider the uniform distribution over four events from Equation 12.5, we would need two bits to encode this event. This is intuitive because a length-two binary string can encode four unique events. However, we can leverage the lower entropy of non-uniform distributions—such as the distribution in Equation 12.6—to use more efficient encodings with variable lengths. In particular, for the distribution in Equation 12.6 we could use the following codes $\{0, 10, 110, 111\}$ for events $\{1, 2, 3, 4\}$, respectively, and the expected length of the message would be

$$2 \times \frac{1}{8} \times 3 - \frac{1}{4} \times 2 - \frac{1}{2} = 1.75, \tag{12.7}$$

since we use shorter codes to encode the more frequent events. Note that we cannot use shorter code strings than this because we need to be able to disambiguate a concatenation of codes from a single long code (e.g., 010 corresponds to the sequence of events $1, 2$ while 110 corresponds to the single event 3).

12.2 Relative Entropy and Mutual Information

Entropy is a useful notion for quantifying the amount of information within a particular distribution. However, we can also use information theory to quantify the relative amounts of information between different distributions. In this section, we will again consider discrete distributions over integers for simplicity. However, all these results naturally generalize to continuous distributions.

12.2.1 Conditional entropy

The first key concept we must introduce is the notion of the conditional entropy between two distributions. In particular, assume that we have a joint distribution $P(x, y)$ and that we observe a realization of the random variable x from this distribution. The conditional entropy tells us how much information this observation from x gives us about the random variable y. The conditional entropy can be computed as follows

$$H(y|x) = \sum_{k_1 \in \mathcal{X}, k_2 \in \mathcal{Y}} P(x = k_1, y = k_2) \log \left(P(y = k_2 | x = k_1) \right). \tag{12.8}$$

Intuitively, the conditional entropy tells us how many bits it takes to send a message about y, assuming that both the sender and receiver have already observed the random variable x. For example, in the case where y is a deterministic function of x then the $H(y|x) = 0$. Alternatively, if x and y are independent, then $H(y|x) = H(y)$.

12.2.2 Relative entropy (KL-divergence)

One of the most popular concepts from information theory that is used in machine learning is *relative entropy*, which is commonly called the Kullback-Leibler (KL) divergence. The KL divergence is a measure of how much one distribution differs from another. Assuming we have two distributions P and Q, the KL-divergence of Q relative to P is equal to

$$KL(P||Q) = -\sum_{k \in \mathcal{X}} P(x=k) \log \left(\frac{Q(x=k)}{P(x=k)} \right) \tag{12.9}$$

$$= -\mathbb{E}_{k \sim P(x)} \left[\log \left(\frac{Q(x=k)}{P(x=k)} \right) \right] \tag{12.10}$$

One way of interpreting the KL divergence is that it measures the additional information we need to encode x if we use a code based on the distribution $Q(x)$ when in fact x is distributed according to $P(x)$. Note that the KL-divergence is not symmetric.

Cross-entropy and KL-divergence

We can use the KL-divergence to give more formal intuition for the cross-entropy loss function in logistic regression. Recall that the cross-entropy loss is defined as

$$L(\hat{y}, y) = -y \log(\hat{y}) - (1-y) \log(1-\hat{y}). \tag{12.11}$$

We can in fact interpret this as the KL-divergence between the estimated distribution, given by $\hat{y}$, and the true distribution, which is given by y. Note that the true distribution always puts all its probability mass on the correct answer.

Maximum likelihood and KL-divergence

We can also relate the KL-divergence to the notion of maximum likelihood optimization more generally. Suppose that we are trying to learn some parameters Θ that maximize the likelihood of a dataset $\mathcal{D} = \{k_1, k_2, ..., k_n\}$, which is sampled from some true distribution $P(x)$. We can view our parameters as specifying some estimated distribution $P_\Theta(x)$ which is attempting to approximate the true distribution $P(x)$. This goal can be formalized as minimizing the KL-divergence of our learned distribution from the true distribution:

$$KL(P||P_\Theta) = -\mathbb{E}_{k \sim P(x)} \left[\log \left(\frac{P_\theta(x=k)}{P(x=k)} \right) \right] \tag{12.12}$$

$$= -\mathbb{E}_{k \sim P(x)} \left[\log \left(P_\Theta(x=k) \right) \right] - \mathbb{E}_{k \sim P(x)} \left[\log \left(P(x=k) \right) \right]. \tag{12.13}$$

Now, from the perspective of optimizing parameters, we can ignore the term

$$\mathbb{E}_{k \sim P(x)} \left[\log \left(P(x=k) \right) \right],$$

since it does not depend on the parameters Θ. Moreover, in practice, we must approximate the expectation in the KL-divergence by taking an average over an i.i.d. sampled dataset $\mathcal{D}$, which gives

$$-\mathbb{E}_{k\sim P(x)}\left[\log\left(P_\Theta(x=k)\right)\right] \approx \frac{-1}{|\mathcal{D}|}\sum_{k\in\mathcal{D}}\log\left(P_\Theta(x=k)\right), \tag{12.14}$$

which is exactly the negative log-likelihood of the data. Thus, we can interpret maximum likelihood as the goal of minimizing the empirical KL-divergence between an estimated distribution and a true distribution.

12.2.3 Mutual information

A final useful concept is the notion of the *mutual information* between two variables x and y. The mutual information $I(x, y)$ quantifies how much the two variables x and y depend on one another. It is defined as the KL-divergence between the joint distribution $P(x, y)$ and the product of the marginals $P(x)P(y)$

$$I(x, y) = KL(P(x, y)||P(x)P(y)). \tag{12.15}$$

In other words, the mutual information measures how much sampling x and y from the distribution $P(x, y)$ differs from sampling the two variables x and y independently from their own marginal distributions. For example, if these two variables are totally independent, then the mutual information would be zero, since in this case we would have that $P(x, y) = P(x)P(y)$.

We can also relate the mutual information to the notions of entropy and conditional entropy as follows

$$I(x, y) = H(x) - H(x|y) = H(y) - H(y|x). \tag{12.16}$$

In this view, the mutual information measures the difference between the entropy $H(x)$ and the conditional entropy $H(x|y)$. Again, we have that the x and y are totally independent, then $H(x) = H(x|y)$, which implies that the mutual information is zero.

The mutual information is a useful concept for measuring how related two variables are. Unlike measures such as statistical correlation, the mutual information does not make assumptions such as linearity. Unfortunately, however, estimating the mutual information for continuous variables can be computationally expensive.