6 Information Theory

In this chapter, we introduce a few basic concepts from the field of **information theory**. More details can be found in other books such as [Mac03; CT06], as well as the sequel to this book, [Mur22].

6.1 Entropy

The **entropy** of a probability distribution can be interpreted as a measure of uncertainty, or lack of predictability, associated with a random variable drawn from a given distribution, as we explain below.

We can also use entropy to define the **information content** of a data source. For example, suppose we observe a sequence of symbols $X_n \sim p$ generated from distribution p. If p has high entropy, it will be hard to predict the value of each osbervation X_n. Hence we say that the dataset $\mathcal{D} = (X_1, \ldots, X_n)$ has high information content. By contrast, if p is a degenerate distribution with 0 entropy (the minimal value), then every X_n will be the same, so $\mathcal{D}$ does not contain much information. (All of this can be formalized in terms of data compression, as we discuss in the sequel to this book.)

6.1.1 Entropy for discrete random variables

The entropy of a discrete random variable X with distribution p over K states is defined by

$$\mathbb{H}(X) \triangleq -\sum_{k=1}^{K} p(X=k) \log_2 p(X=k) = -\mathbb{E}_X[\log p(X)] \tag{6.1}$$

(Note that we use the notation $\mathbb{H}(X)$ to denote the entropy of the rv with distribution p, just as people write $\mathbb{V}[X]$ to mean the variance of the distribution associated with X; we could alternatively write $\mathbb{H}(p)$.) Usually we use log base 2, in which case the units are called **bits** (short for binary digits). For example, if $X \in \{1, \ldots, 5\}$ with histogram distribution $p = [0.25, 0.25, 0.2, 0.15, 0.15]$, we find $H = 2.29$ bits. If we use log base e, the units are called **nats**.

The discrete distribution with **maximum entropy** is the uniform distribution. Hence for a K-ary random variable, the entropy is maximized if $p(x=k) = 1/K$; in this case, $\mathbb{H}(X) = \log_2 K$. To see this, note that

$$\mathbb{H}(X) = -\sum_{k=1}^{K} \frac{1}{K} \log(1/K) = -\log(1/K) = \log(K) \tag{6.2}$$

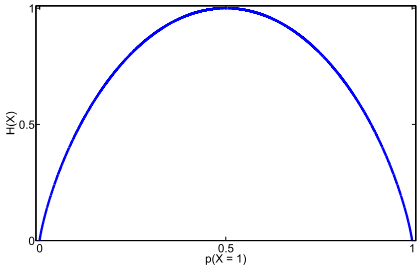
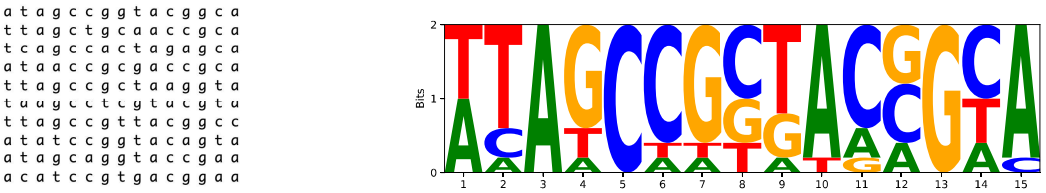


Figure 6.1: Entropy of a Bernoulli random variable as a function of θ. The maximum entropy is $\log_2 2 = 1$. Generated by code at figures.probml.ai/book1/6.1.



(a) *(b)*

Figure 6.2: (a) Some aligned DNA sequences. Each row is a sequence, each column is a location within the sequence. (b) The corresponding **position weight matrix** *represented as a sequence logo. Each column represents a probablity distribution over the alphabet $\{A, C, G, T\}$ for the corresponding location in the sequence. The size of the letter is proportional to the probability. The height of column t is given by $2 - H_t$, where $0 \leq H_t \leq 2$ is the entropy (in bits) of the distribution $\mathbf{p}_t$. Thus deterministic distributions (with an entropy of 0, corresponding to highly conserved locations) have height 2, and uniform distributions (with an entropy of 2) have height 0. Generated by code at figures.probml.ai/book1/6.2.*

Conversely, the distribution with minimum entropy (which is zero) is any delta-function that puts all its mass on one state. Such a distribution has no uncertainty.

For the special case of binary random variables, $X \in \{0, 1\}$, we can write $p(X = 1) = \theta$ and $p(X = 0) = 1 - \theta$. Hence the entropy becomes

$$\mathbb{H}(X) = -[p(X = 1) \log_2 p(X = 1) + p(X = 0) \log_2 p(X = 0)] \tag{6.3}$$

$$= -[\theta \log_2 \theta + (1 - \theta) \log_2 (1 - \theta)] \tag{6.4}$$

This is called the **binary entropy function**, and is also written $\mathbb{H}(\theta)$. We plot this in Figure 6.1. We see that the maximum value of 1 bit occurs when the distribution is uniform, $\theta = 0.5$. A fair coin requires a single yes/no question to determine its state.

As an interesting application of entropy, consider the problem of representing **DNA sequence motifs**, which is a distribution over short DNA strings. We can estimate this distribution by aligning a set of DNA sequences (e.g., from different species), and then estimating the empirical distribution of each possible nucleotide from the 4 letter alphabet $X \sim \{A, C, G, T\}$ at each location t in the ith sequence as follows:

$$\mathbf{N}_t = \left(\sum_{i=1}^{N} \mathbb{I}\left(X_{it} = A\right), \sum_{i=1}^{N} \mathbb{I}\left(X_{it} = C\right), \sum_{i=1}^{N} \mathbb{I}\left(X_{it} = G\right), \sum_{i=1}^{N} \mathbb{I}\left(X_{it} = T\right) \right) \tag{6.5}$$

$$\hat{\boldsymbol{\theta}}_t = \mathbf{N}_t / N, \tag{6.6}$$

This $\mathbf{N}_t$ is a length four vector counting the number of times each letter appears at each location amongst the set of sequences. This $\hat{\boldsymbol{\theta}}_t$ distribution is known as a motif. We can also compute the most probable letter in each location; this is called the **consensus sequence**.

One way to visually summarize the data is by using a **sequence logo**, as shown in Figure 6.2(b). We plot the letters A, C, G and T, with the most probable letter on the top; the height of the t'th bar is defined to be $0 \leq 2 - H_t \leq 2$, where H_t is the entropy of $\hat{\boldsymbol{\theta}}_t$ (note that 2 is the maximum possible entropy for a distribution over 4 letters). Thus tall bars correspond to nearly deterministic distributions, which are the locations that are conserved by evolution (e.g., because they are part of a gene coding region). In this example, we see that column 13 is all G's, and hence has height 2.

Estimating the entropy of a random variable with many possible states requires estimating its distribution, which can require a lot of data. For example, imagine if X represents the identity of a word in an English document. Since there is a long tail of rare words, and since new words are invented all the time, it can be difficult to reliably estimate $p(X)$ and hence $\mathbb{H}(X)$. For one possible solution to this problem, see [VV13].

6.1.2 Cross entropy

The **cross entropy** between distribution p and q is defined by

$$\mathbb{H}(p, q) \triangleq -\sum_{k=1}^{K} p_k \log q_k \tag{6.7}$$

One can show that the cross entropy is the expected number of bits needed to compress some data samples drawn from distribution p using a code based on distribution q. This can be minimized by setting $q = p$, in which case the expected number of bits of the optimal code is $\mathbb{H}(p, p) = \mathbb{H}(p)$ — this is known as **Shannon's source coding theorem** (see e.g., [CT06]).

6.1.3 Joint entropy

The joint entropy of two random variables X and Y is defined as

$$\mathbb{H}(X, Y) = -\sum_{x,y} p(x, y) \log_2 p(x, y) \tag{6.8}$$

For example, consider choosing an integer from 1 to 8, $n \in \{1, \ldots, 8\}$. Let $X(n) = 1$ if n is even, and $Y(n) = 1$ if n is prime:

n	1	2	3	4	5	6	7	8
X	0	1	0	1	0	1	0	1
Y	0	1	1	0	1	0	1	0

The joint distribution is

$p(X,Y)$	$Y=0$	$Y=1$
$X=0$	$\frac{1}{8}$	$\frac{3}{8}$
$X=1$	$\frac{3}{8}$	$\frac{1}{8}$

so the joint entropy is given by

$$\mathbb{H}(X,Y) = -\left[\frac{1}{8}\log_2\frac{1}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{1}{8}\log_2\frac{1}{8}\right] = 1.81 \text{ bits} \tag{6.9}$$

Clearly the marginal probabilities are uniform: $p(X=1) = p(X=0) = p(Y=0) = p(Y=1) = 0.5$, so $\mathbb{H}(X) = \mathbb{H}(Y) = 1$. Hence $\mathbb{H}(X,Y) = 1.81$ bits $< \mathbb{H}(X) + \mathbb{H}(Y) = 2$ bits. In fact, this upper bound on the joint entropy holds in general. If X and Y are independent, then $\mathbb{H}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y)$, so the bound is tight. This makes intuitive sense: when the parts are correlated in some way, it reduces the "degrees of freedom" of the system, and hence reduces the overall entropy.

What is the lower bound on $\mathbb{H}(X,Y)$? If Y is a deterministic function of X, then $\mathbb{H}(X,Y) = \mathbb{H}(X)$. So

$$\mathbb{H}(X,Y) \geq \max\{\mathbb{H}(X), \mathbb{H}(Y)\} \geq 0 \tag{6.10}$$

Intuitively this says combining variables together does not make the entropy go down: you cannot reduce uncertainty merely by adding more unknowns to the problem, you need to observe some data, a topic we discuss in Section 6.1.4.

We can extend the definition of joint entropy from two variables to n in the obvious way.

6.1.4 Conditional entropy

The **conditional entropy** of Y given X is the uncertainty we have in Y after seeing X, averaged over possible values for X:

$$\mathbb{H}(Y|X) \triangleq \mathbb{E}_{p(X)}\left[\mathbb{H}(p(Y|X))\right] \tag{6.11}$$

$$= \sum_x p(x)\,\mathbb{H}(p(Y|X=x)) = -\sum_x p(x)\sum_y p(y|x)\log p(y|x) \tag{6.12}$$

$$= -\sum_{x,y} p(x,y)\log p(y|x) = -\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)} \tag{6.13}$$

$$= -\sum_{x,y} p(x,y)\log p(x,y) - \sum_x p(x)\log\frac{1}{p(x)} \tag{6.14}$$

$$= \mathbb{H}(X,Y) - \mathbb{H}(X) \tag{6.15}$$

If Y is a deterministic function of X, then knowing X completely determines Y, so $\mathbb{H}(Y|X) = 0$. If X and Y are independent, knowing X tells us nothing about Y and $\mathbb{H}(Y|X) = \mathbb{H}(Y)$. Since $\mathbb{H}(X,Y) \leq \mathbb{H}(Y) + \mathbb{H}(X)$, we have

$$\mathbb{H}(Y|X) \leq \mathbb{H}(Y) \tag{6.16}$$

with equality iff X and Y are independent. This shows that, on average, conditioning on data never increases one's uncertainty. The caveat "on average" is necessary because for any *particular* observation (value of X), one may get more "confused" (i.e., $\mathbb{H}(Y|x) > \mathbb{H}(Y)$). However, in expectation, looking at the data is a good thing to do. (See also Section 6.3.8.)

We can rewrite Equation (6.15) as follows:

$$\mathbb{H}(X_1, X_2) = \mathbb{H}(X_1) + \mathbb{H}(X_2|X_1) \tag{6.17}$$

This can be generalized to get the **chain rule for entropy**:

$$\mathbb{H}(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} \mathbb{H}(X_i|X_1, \ldots, X_{i-1}) \tag{6.18}$$

6.1.5 Perplexity

The **perplexity** of a discrete probability distribution p is defined as

$$\text{perplexity}(p) \triangleq 2^{\mathbb{H}(p)} \tag{6.19}$$

This is often interpreted as a measure of predictability. For example, suppose p is a uniform distribution over K states. In this case, the perplexity is K. Obviously the lower bound on perplexity is $2^0 = 1$, which will be achieved if the distribution can perfectly predict outcomes.

Now suppose we have an empirical distribution based on data $\mathcal{D}$:

$$p_{\mathcal{D}}(x|\mathcal{D}) = \frac{1}{N} \sum_{n=1}^{N} \delta_{x_n}(x) \tag{6.20}$$

We can measure how well p predicts $\mathcal{D}$ by computing

$$\text{perplexity}(p_{\mathcal{D}}, p) \triangleq 2^{\mathbb{H}(p_{\mathcal{D}}, p)} \tag{6.21}$$

Perplexity is often used to evaluate the quality of statistical language models, which is a generative model for sequences of tokens. Suppose the data is a single long document x of length N, and suppose p is a simple unigram model. In this case, the cross entropy term is given by

$$H = -\frac{1}{N} \sum_{n=1}^{N} \log p(x_n) \tag{6.22}$$

and hence the perplexity is given by

$$\text{perplexity}(p_{\mathcal{D}}, p) = 2^H = 2^{-\frac{1}{N} \log(\prod_{n=1}^{N} p(x_n))} = \sqrt[N]{\prod_{n=1}^{N} \frac{1}{p(x_n)}} \tag{6.23}$$

This is sometimes called the **exponentiated cross entropy**. We see that this is the geometric mean of the inverse predictive probabilities.

In the case of language models, we usually condition on previous words when predicting the next word. For example, in a bigram model, we use a second order Markov model of the form $p(x_i|x_{i-1})$. We define the **branching factor** of a language model as the number of possible words that can follow any given word. We can thus interpret the perplexity as the weighted average branching factor. For example, suppose the model predicts that each word is equally likely, regardless of context, so $p(x_i|x_{i-1}) = 1/K$. Then the perplexity is $((1/K)^N)^{-1/N} = K$. If some symbols are more likely than others, and the model correctly reflects this, its perplexity will be lower than K. However, as we show in Section 6.2, we have $\mathbb{H}(p^*) \leq \mathbb{H}(p^*, p)$, so we can never reduce the perplexity below the entropy of the underlying stochastic process p^*.

See [JM08, p96] for further discussion of perplexity and its uses in language models.

6.1.6 Differential entropy for continuous random variables *

If X is a continuous random variable with pdf $p(x)$, we define the **differential entropy** as

$$h(X) \triangleq -\int_{\mathcal{X}} dx\, p(x) \log p(x) \tag{6.24}$$

assuming this integral exists. For example, suppose $X \sim U(0, a)$. Then

$$h(X) = -\int_0^a dx\, \frac{1}{a} \log \frac{1}{a} = \log a \tag{6.25}$$

Note that, unlike the discrete case, *differential entropy can be negative*. This is because pdf's can be bigger than 1. For example if $X \sim U(0, 1/8)$, we have $h(X) = \log_2(1/8) = -3$.

One way to understand differential entropy is to realize that all real-valued quantities can only be represented to finite precision. It can be shown [CT91, p228] that the entropy of an n-bit quantization of a continuous random variable X is approximately $h(X) + n$. For example, suppose $X \sim U(0, \frac{1}{8})$. Then in a binary representation of X, the first 3 bits to the right of the binary point must be 0 (since the number is $\leq 1/8$). So to describe X to n bits of accuracy only requires $n - 3$ bits, which agrees with $h(X) = -3$ calculated above.

6.1.6.1 Example: Entropy of a Gaussian

The entropy of a d-dimensional Gaussian is

$$h(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = \frac{1}{2} \ln |2\pi e \boldsymbol{\Sigma}| = \frac{1}{2} \ln[(2\pi e)^d |\boldsymbol{\Sigma}|] = \frac{d}{2} + \frac{d}{2} \ln(2\pi) + \frac{1}{2} \ln |\boldsymbol{\Sigma}| \tag{6.26}$$

In the 1d case, this becomes

$$h(\mathcal{N}(\mu, \sigma^2)) = \frac{1}{2} \ln \left[2\pi e \sigma^2\right] \tag{6.27}$$

6.1.6.2 Connection with variance

The entropy of a Gaussian increases monotonically as the variance increases. However, this is not always the case. For example, consider a mixture of two 1d Gaussians centered at -1 and +1. As we

move the means further apart, say to -10 and +10, the variance increases (since the average distance from the overall mean gets larger). However, the entropy remains more or less the same, since we are still uncertain about where a sample might fall, even if we know that it will be near -10 or +10. (The exact entropy of a GMM is hard to compute, but a method to compute upper and lower bounds is presented in [Hub+08].)

6.1.6.3 Discretization

In general, computing the differential entropy for a continuous random variable can be difficult. A simple approximation is to **discretize** or **quantize** the variables. There are various methods for this (see e.g., [DKS95; KK06] for a summary), but a simple approach is to bin the distribution based on its empirical quantiles. The critical question is how many bins to use [LM04]. Scott [Sco79] suggested the following heuristic:

$$B = N^{1/3} \frac{\max(\mathcal{D}) - \min(\mathcal{D})}{3.5\sigma(\mathcal{D})} \tag{6.28}$$

where $\sigma(\mathcal{D})$ is the empirical standard deviation of the data, and $N = |\mathcal{D}|$ is the number of datapoints in the empirical distribution. However, the technique of discretization does not scale well if X is a multi-dimensional random vector, due to the curse of dimensionality.

6.2 Relative entropy (KL divergence) *

Given two distributions p and q, it is often useful to define a **distance metric** to measure how "close" or "similar" they are. In fact, we will be more general and consider a **divergence measure** $D(p, q)$ which quantifies how far q is from p, without requiring that D be a metric. More precisely, we say that D is a divergence if $D(p, q) \geq 0$ with equality iff $p = q$, whereas a metric also requires that D be symmetric and satisfy the **triangle inequality**, $D(p, r) \leq D(p, q) + D(q, r)$. There are many possible divergence measures we can use. In this section, we focus on the **Kullback-Leibler divergence** or **KL divergence**, also known as the **information gain** or **relative entropy**, between two distributions p and q.

6.2.1 Definition

For discrete distributions, the KL divergence is defined as follows:

$$D_{\mathbb{KL}}\left(p \| q\right) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k} \tag{6.29}$$

This naturally extends to continuous distributions as well:

$$D_{\mathbb{KL}}\left(p \| q\right) \triangleq \int dx\, p(x) \log \frac{p(x)}{q(x)} \tag{6.30}$$

6.2.2 Interpretation

We can rewrite the KL as follows:

$$D_{\mathbb{KL}}\left(p \| q\right) = \underbrace{\sum_{k=1}^{K} p_k \log p_k}_{-\mathbb{H}(p)} \underbrace{- \sum_{k=1}^{K} p_k \log q_k}_{\mathbb{H}(p,q)} \tag{6.31}$$

We recognize the first term as the negative entropy, and the second term as the cross entropy. It can be shown that the cross entropy $\mathbb{H}(p, q)$ is a lower bound on the number of bits needed to compress data coming from distribution p if your code is designed based on distribution q; thus we can interpret the KL divergence as the "extra number of bits" you need to pay when compressing data samples if you use the incorrect distribution q as the basis of your coding scheme compared to the true distribution p.

There are various other interpretations of KL divergence. See the sequel to this book, [Mur22], for more information.

6.2.3 Example: KL divergence between two Gaussians

For example, one can show that the KL divergence between two multivariate Gaussian distributions is given by

$$\begin{aligned} &D_{\mathbb{KL}}\left(\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \| \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)\right) \\ &= \frac{1}{2}\left[\mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - D + \log\left(\frac{\det(\boldsymbol{\Sigma}_2)}{\det(\boldsymbol{\Sigma}_1)}\right)\right] \end{aligned} \tag{6.32}$$

In the scalar case, this becomes

$$D_{\mathbb{KL}}\left(\mathcal{N}(x|\mu_1, \sigma_1) \| \mathcal{N}(x|\mu_2, \sigma_2)\right) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{6.33}$$

6.2.4 Non-negativity of KL

In this section, we prove that the KL divergence is always non-negative.

To do this, we use **Jensen's inequality**. This states that, for any convex function f, we have that

$$f(\sum_{i=1}^{n} \lambda_i \boldsymbol{x}_i) \leq \sum_{i=1}^{n} \lambda_i f(\boldsymbol{x}_i) \tag{6.34}$$

where $\lambda_i \geq 0$ and $\sum_{i=1}^{n} \lambda_i = 1$. In words, this result says that f of the average is less than the average of the f's. This is clearly true for $n = 2$, since a convex function curves up above a straight line connecting the two end points (see Section 8.1.3). To prove for general n, we can use induction.

For example, if $f(x) = \log(x)$, which is a concave function, we have

$$\log(\mathbb{E}_x g(x)) \geq \mathbb{E}_x \log(g(x)) \tag{6.35}$$

We use this result below.

Theorem 6.2.1. *(Information inequality)* $D_{\mathbb{KL}}(p\|q) \geq 0$ *with equality iff* $p = q$.

Proof. We now prove the theorem following [CT06, p28]. Let $A = \{x : p(x) > 0\}$ be the support of $p(x)$. Using the concavity of the log function and Jensen's inequality (Section 6.2.4), we have that

$$-D_{\mathbb{KL}}(p\|q) = -\sum_{x\in A} p(x)\log\frac{p(x)}{q(x)} = \sum_{x\in A} p(x)\log\frac{q(x)}{p(x)} \tag{6.36}$$

$$\leq \log\sum_{x\in A} p(x)\frac{q(x)}{p(x)} = \log\sum_{x\in A} q(x) \tag{6.37}$$

$$\leq \log\sum_{x\in\mathcal{X}} q(x) = \log 1 = 0 \tag{6.38}$$

Since $\log(x)$ is a strictly concave function ($-\log(x)$ is convex), we have equality in Equation (6.37) iff $p(x) = cq(x)$ for some c that tracks the fraction of the whole space $\mathcal{X}$ contained in A. We have equality in Equation (6.38) iff $\sum_{x\in A} q(x) = \sum_{x\in\mathcal{X}} q(x) = 1$, which implies $c = 1$. Hence $D_{\mathbb{KL}}(p\|q) = 0$ iff $p(x) = q(x)$ for all x. □

This theorem has many important implications, as we will see throughout the book. For example, we can show that the uniform distribution is the one that maximizes the entropy:

Corollary 6.2.1. *(Uniform distribution maximizes the entropy)* $\mathbb{H}(X) \leq \log|\mathcal{X}|$, *where* $|\mathcal{X}|$ *is the number of states for* X, *with equality iff* $p(x)$ *is uniform.*

Proof. Let $u(x) = 1/|\mathcal{X}|$. Then

$$0 \leq D_{\mathbb{KL}}(p\|u) = \sum_x p(x)\log\frac{p(x)}{u(x)} = \log|\mathcal{X}| - \mathbb{H}(X) \tag{6.39}$$

□

6.2.5 KL divergence and MLE

Suppose we want to find the distribution q that is as close as possible to p, as measured by KL divergence:

$$q^* = \arg\min_q D_{\mathbb{KL}}(p\|q) = \arg\min_q \int p(x)\log p(x)dx - \int p(x)\log q(x)dx \tag{6.40}$$

Now suppose p is the empirical distribution, which puts a probability atom on the observed training data and zero mass everywhere else:

$$p_{\mathcal{D}}(x) = \frac{1}{N}\sum_{n=1}^{N}\delta(x - x_n) \tag{6.41}$$

Using the sifting property of delta functions we get

$$D_{\mathbb{KL}}\left(p_{\mathcal{D}} \| q\right) = -\int p_{\mathcal{D}}(x) \log q(x) dx + C \tag{6.42}$$

$$= -\int \left[\frac{1}{N}\sum_n \delta(x - x_n)\right] \log q(x) dx + C \tag{6.43}$$

$$= -\frac{1}{N}\sum_n \log q(x_n) + C \tag{6.44}$$

where $C = \int p(x) \log p(x) dx$ is a constant independent of q. This is called the **cross entropy** objective, and is equal to the average negative log likelihood of q on the training set. Thus we see that minimizing KL divergence to the empirical distribution is equivalent to maximizing likelihood.

This perspective points out the flaw with likelihood-based training, namely that it puts too much weight on the training set. In most applications, we do not really believe that the empirical distribution is a good representation of the true distribution, since it just puts "spikes" on a finite set of points, and zero density everywhere else. Even if the dataset is large (say 1M images), the universe from which the data is sampled is usually even larger (e.g., the set of "all natural images" is much larger than 1M). We could smooth the empirical distribution using kernel density estimation (Section 16.3), but that would require a similar kernel on the space of images. An alternative, algorithmic approach is to use **data augmentation**, which is a way of perturbing the observed data samples in way that we believe reflects plausible "natural variation". Applying MLE on this augmented dataset often yields superior results, especially when fitting models with many parameters (see Section 19.1).

6.2.6 Forward vs reverse KL

Suppose we want to approximate a distribution p using a simpler distribution q. We can do this by minimizing $D_{\mathbb{KL}}\left(q \| p\right)$ or $D_{\mathbb{KL}}\left(p \| q\right)$. This gives rise to different behavior, as we discuss below.

First we consider the **forwards KL**, also called the **inclusive KL**, defined by

$$D_{\mathbb{KL}}\left(p \| q\right) = \int p(x) \log \frac{p(x)}{q(x)} dx \tag{6.45}$$

Minimizing this wrt q is known as an **M-projection** or **moment projection**.

We can gain an understanding of the optimal q by considering inputs x for which $p(x) > 0$ but $q(x) = 0$. In this case, the term $\log p(x)/q(x)$ will be infinite. Thus minimizing the KL will force q to include all the areas of space for which p has non-zero probability. Put another way, q will be **zero-avoiding** or **mode-covering**, and will typically over-estimate the support of p. Figure 6.3(a) illustrates mode covering where p is a bimodal distribution but q is unimodal.

Now consider the **reverse KL**, also called the **exclusive KL**:

$$D_{\mathbb{KL}}\left(q \| p\right) = \int q(x) \log \frac{q(x)}{p(x)} dx \tag{6.46}$$

Minimizing this wrt q is known as an **I-projection** or **information projection**.

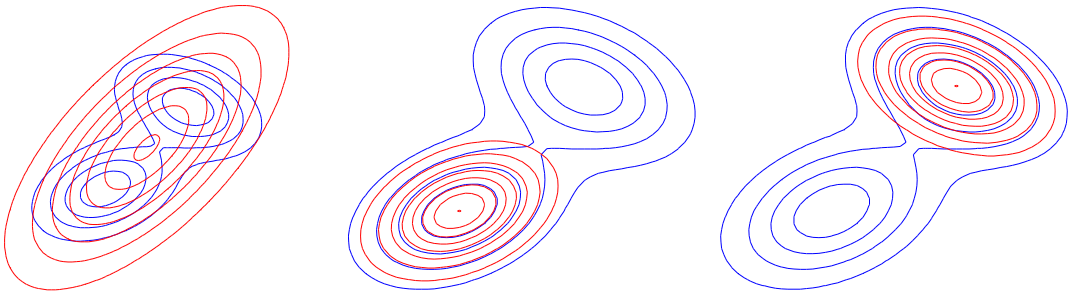


Figure 6.3: Illustrating forwards vs reverse KL on a bimodal distribution. The blue curves are the contours of the true distribution p. The red curves are the contours of the unimodal approximation q. (a) Minimizing forwards KL, $D_{\mathbb{KL}}(p\|q)$, wrt q causes q to "cover" p. (b-c) Minimizing reverse KL, $D_{\mathbb{KL}}(q\|p)$ wrt q causes q to "lock onto" one of the two modes of p. Adapted from Figure 10.3 of [Bis06]. Generated by code at figures.probml.ai/book1/6.3.

We can gain an understanding of the optimal q by consider inputs x for which $p(x) = 0$ but $q(x) > 0$. In this case, the term $\log q(x)/p(x)$ will be infinite. Thus minimizing the exclusive KL will force q to exclude all the areas of space for which p has zero probability. One way to do this is for q to put probability mass in very few parts of space; this is called **zero-forcing** or **mode-seeking** behavior. In this case, q will typically under-estimate the support of p. We illustrate mode seeking when p is bimodal but q is unimodal in Figure 6.3(b-c).

6.3 Mutual information *

The KL divergence gave us a way to measure how similar two distributions were. How should we measure how dependant two random variables are? One thing we could do is turn the question of measuring the dependence of two random variables into a question about the similarity of their distributions. This gives rise to the notion of **mutual information** (MI) between two random variables, which we define below.

6.3.1 Definition

The mutual information between rv's X and Y is defined as follows:

$$\mathbb{I}(X;Y) \triangleq D_{\mathbb{KL}}(p(x,y)\|p(x)p(y)) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)} \tag{6.47}$$

(We write $\mathbb{I}(X;Y)$ instead of $\mathbb{I}(X,Y)$, in case X and/or Y represent sets of variables; for example, we can write $\mathbb{I}(X;Y,Z)$ to represent the MI between X and (Y,Z).) For continuous random variables, we just replace sums with integrals.

It is easy to see that MI is always non-negative, even for continuous random variables, since

$$\mathbb{I}(X;Y) = D_{\mathbb{KL}}(p(x,y)\|p(x)p(y)) \geq 0 \tag{6.48}$$

We achieve the bound of 0 iff $p(x, y) = p(x)p(y)$.

6.3.2 Interpretation

Knowing that the mutual information is a KL divergence between the joint and factored marginal distributions tells us that the MI measures the information gain if we update from a model that treats the two variables as independent $p(x)p(y)$ to one that models their true joint density $p(x, y)$.

To gain further insight into the meaning of MI, it helps to re-express it in terms of joint and conditional entropies, as follows:

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X) \tag{6.49}$$

Thus we can interpret the MI between X and Y as the reduction in uncertainty about X after observing Y, or, by symmetry, the reduction in uncertainty about Y after observing X. Incidentally, this result gives an alternative proof that conditioning, on average, reduces entropy. In particular, we have $0 \leq \mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y)$, and hence $\mathbb{H}(X|Y) \leq \mathbb{H}(X)$.

We can also obtain a different interpretation. One can show that

$$\mathbb{I}(X;Y) = \mathbb{H}(X,Y) - \mathbb{H}(X|Y) - \mathbb{H}(Y|X) \tag{6.50}$$

Finally, one can show that

$$\mathbb{I}(X;Y) = \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) \tag{6.51}$$

See Figure 6.4 for a summary of these equations in terms of an **information diagram**. (Formally, this is a signed measure mapping set expressions to their information-theoretic counterparts [Yeu91].)

6.3.3 Example

As an example, let us reconsider the example concerning prime and even numbers from Section 6.1.3. Recall that $\mathbb{H}(X) = \mathbb{H}(Y) = 1$. The conditional distribution $p(Y|X)$ is given by normalizing each row:

	Y=0	Y=1
X=0	$\frac{1}{4}$	$\frac{3}{4}$
X=1	$\frac{3}{4}$	$\frac{1}{4}$

Hence the conditional entropy is

$$\mathbb{H}(Y|X) = -\left[\frac{1}{8}\log_2\frac{1}{4} + \frac{3}{8}\log_2\frac{3}{4} + \frac{3}{8}\log_2\frac{3}{4} + \frac{1}{8}\log_2\frac{1}{4}\right] = 0.81 \text{ bits} \tag{6.52}$$

and the mutual information is

$$\mathbb{I}(X;Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X) = (1 - 0.81) \text{ bits} = 0.19 \text{ bits} \tag{6.53}$$

You can easily verify that

$$\mathbb{H}(X,Y) = \mathbb{H}(X|Y) + \mathbb{I}(X;Y) + \mathbb{H}(Y|X) \tag{6.54}$$

$$= (0.81 + 0.19 + 0.81) \text{ bits} = 1.81 \text{ bits} \tag{6.55}$$

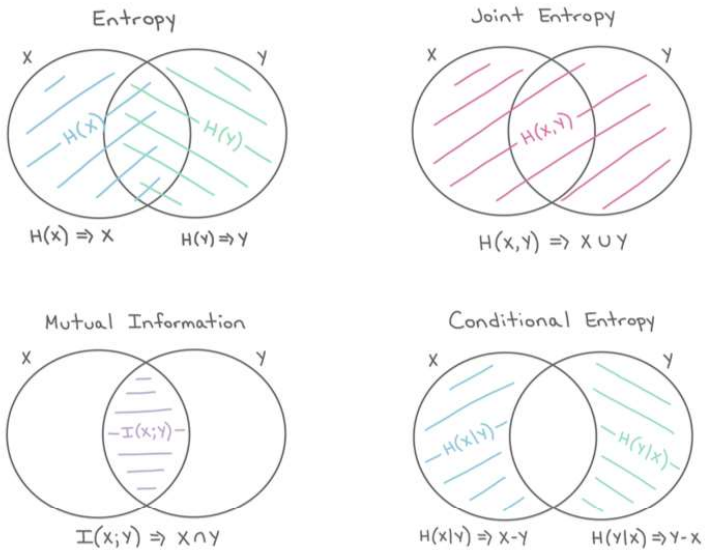


Figure 6.4: The marginal entropy, joint entropy, conditional entropy and mutual information represented as information diagrams. Used with kind permission of Katie Everett.

6.3.4 Conditional mutual information

We can define the **conditional mutual information** in the obvious way

$$\begin{aligned}
\mathbb{I}(X;Y|Z) &\triangleq \mathbb{E}_{p(Z)}\left[\mathbb{I}(X;Y)|Z\right] && (6.56)\\
&= \mathbb{E}_{p(x,y,z)}\left[\log \frac{p(x,y|z)}{p(x|z)p(y|z)}\right] && (6.57)\\
&= \mathbb{H}(X|Z) + \mathbb{H}(Y|Z) - \mathbb{H}(X,Y|Z) && (6.58)\\
&= \mathbb{H}(X|Z) - \mathbb{H}(X|Y,Z) = \mathbb{H}(Y|Z) - \mathbb{H}(Y|X,Z) && (6.59)\\
&= \mathbb{H}(X,Z) + \mathbb{H}(Y,Z) - \mathbb{H}(Z) - \mathbb{H}(X,Y,Z) && (6.60)\\
&= \mathbb{I}(Y;X,Z) - \mathbb{I}(Y;Z) && (6.61)
\end{aligned}$$

The last equation tells us that the conditional MI is the extra (residual) information that X tells us about Y, excluding what we already knew about Y given Z alone.

We can rewrite Equation (6.61) as follows:

$$\mathbb{I}(Z,Y;X) = \mathbb{I}(Z;X) + \mathbb{I}(Y;X|Z) \tag{6.62}$$

Generalizing to N variables, we get the **chain rule for mutual information**:

$$\mathbb{I}(Z_1,\ldots,Z_N;X) = \sum_{n=1}^{N} \mathbb{I}(Z_n;X|Z_1,\ldots,Z_{n-1}) \tag{6.63}$$

6.3.5 MI as a "generalized correlation coefficient"

Suppose that (x, y) are jointly Gaussian:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix}\right) \tag{6.64}$$

We now show how to compute the mutual information between X and Y.

Using Equation (6.26), we find that the entropy is

$$h(X, Y) = \frac{1}{2} \log\left[(2\pi e)^2 \det \Sigma\right] = \frac{1}{2} \log\left[(2\pi e)^2 \sigma^4 (1 - \rho^2)\right] \tag{6.65}$$

Since X and Y are individually normal with variance σ^2, we have

$$h(X) = h(Y) = \frac{1}{2} \log\left[2\pi e \sigma^2\right] \tag{6.66}$$

Hence

$$\begin{aligned} I(X, Y) &= h(X) + h(Y) - h(X, Y) && (6.67)\\ &= \log[2\pi e\sigma^2] - \frac{1}{2}\log[(2\pi e)^2\sigma^4(1-\rho^2)] && (6.68)\\ &= \frac{1}{2}\log[(2\pi e\sigma^2)^2] - \frac{1}{2}\log[(2\pi e\sigma^2)^2(1-\rho^2)] && (6.69)\\ &= \frac{1}{2}\log\frac{1}{1-\rho^2} = -\frac{1}{2}\log[1-\rho^2] && (6.70) \end{aligned}$$

We now discuss some interesting special cases.

1. $\rho = 1$. In this case, $X = Y$, and $I(X, Y) = \infty$, which makes sense. Observing Y tells us an infinite amount of information about X (as we know its real value exactly).
2. $\rho = 0$. In this case, X and Y are independent, and $I(X, Y) = 0$, which makes sense. Observing Y tells us nothing about X.
3. $\rho = -1$. In this case, $X = -Y$, and $I(X, Y) = \infty$, which again makes sense. Observing Y allows us to predict X to infinite precision.

Now consider the case where X and Y are scalar, but not jointly Gaussian. In general it can be difficult to compute the mutual information between continuous random variables, because we have to estimate the joint density $p(X, Y)$. For scalar variables, a simple approximation is to **discretize** or **quantize** them, by dividing the ranges of each variable into bins, and computing how many values fall in each histogram bin [Sco79]. We can then easily compute the MI using the empirical pmf.

Unfortunately, the number of bins used, and the location of the bin boundaries, can have a significant effect on the results. One way to avoid this is to use K-nearest neighbor distances to estimate densities in a non-parametric, adaptive way. This is the basis of the **KSG estimator** for MI proposed in [KSG04]. This is implemented in the sklearn.feature_selection.mutual_info_regression function. For papers related to this estimator, see [GOV18; HN19].

6.3.6 Normalized mutual information

For some applications, it is useful to have a normalized measure of dependence, between 0 and 1. We now discuss one way to construct such a measure.

First, note that

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) \leq \mathbb{H}(X) \tag{6.71}$$

$$= \mathbb{H}(Y) - \mathbb{H}(Y|X) \leq \mathbb{H}(Y) \tag{6.72}$$

so

$$0 \leq \mathbb{I}(X;Y) \leq \min(\mathbb{H}(X), \mathbb{H}(Y)) \tag{6.73}$$

Therefore we can define the **normalized mutual information** as follows:

$$NMI(X,Y) = \frac{\mathbb{I}(X;Y)}{\min(\mathbb{H}(X), \mathbb{H}(Y))} \leq 1 \tag{6.74}$$

This normalized mutual information ranges from 0 to 1. When $NMI(X,Y) = 0$, we have $\mathbb{I}(X;Y) = 0$, so X and Y are independent. When $NMI(X,Y) = 1$, and $\mathbb{H}(X) < \mathbb{H}(Y)$, we have

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(X) \implies \mathbb{H}(X|Y) = 0 \tag{6.75}$$

and so X is a deterministic function of Y. For example, suppose X is a discrete random variable with pmf $[0.5, 0.25, 0.25]$. We have $MI(X,X) = 1.5$ (using log base 2), and $H(X) = 1.5$, so the normalized MI is 1, as is to be expected.

For continuous random variables, it is harder to normalize the mutual information, because of the need to estimate the differential entropy, which is sensitive to the level of quantization. See Section 6.3.7 for further discussion.

6.3.7 Maximal information coefficient

As we discussed in Section 6.3.6, it is useful to have a normalized estimate of the mutual information, but this can be tricky to compute for real-valued data. One approach, known as the maximal information coefficient (MIC) [Res+11], is to define the following quantity:

$$\text{MIC}(X,Y) = \max_G \frac{\mathbb{I}((X,Y)|_G)}{\log ||G||} \tag{6.76}$$

where G is the set of 2d grids, and $(X,Y)|_G$ represents a discretization of the variables onto this grid, and $||G||$ is $\min(G_x, G_y)$, where G_x is the number of grid cells in the x direction, and G_y is the number of grid cells in the y direction. (The maximum grid resolution depends on the sample size n; they suggest restricting grids so that $G_x G_y \leq B(n)$, where $B(n) = n^\alpha$, where $\alpha = 0.6$.) The denominator is the entropy of a uniform joint distribution; dividing by this ensures $0 \leq \text{MIC} \leq 1$.

The intuition behind this statistic is the following: if there is a relationship between X and Y, then there should be some discrete gridding of the 2d input space that captures this. Since we don't know the correct grid to use, MIC searches over different grid resolutions (e.g., 2x2, 2x3, etc), as well as over locations of the grid boundaries. Given a grid, it is easy to quantize the data and compute

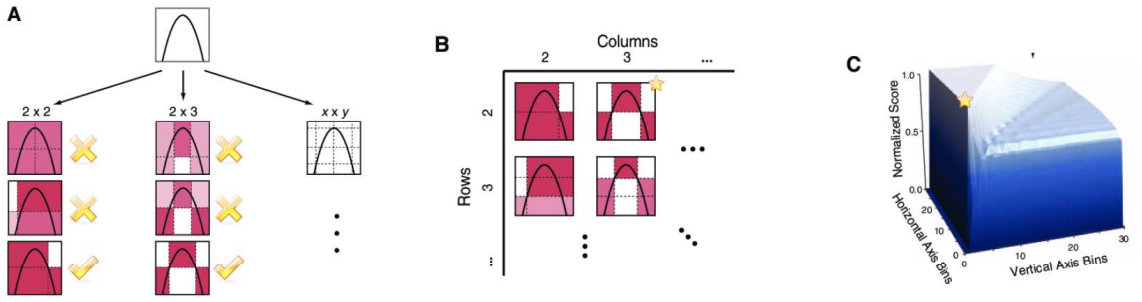


Figure 6.5: Illustration of how the maximal information coefficient (MIC) is computed. (a) We search over different grid resolutions, and grid cell locations, and compute the MI for each. (b) For each grid resolution (k, l), we define set $M(k, l)$ to be the maximum MI for any grid of that size, normalized by $\log(\min(k, l))$. *(c) We visualize the matrix* $\mathbf{M}$. *The maximum entry (denoted by a star) is defined to be the MIC. From Figure 1 of [Res+11]. Used with kind permission of David Reshef.*

MI. We define the **characteristic matrix** $M(k, l)$ to be the maximum MI achievable by any grid of size (k, l), normalized by $\log(\min(k, l))$. The MIC is then the maximum entry in this matrix, $\max_{kl \leq B(n)} M(k, l)$. See Figure 6.5 for a visualization of this process.

In [Res+11], they show that this quantity exhibits a property known as **equitability**, which means that it gives similar scores to equally noisy relationships, regardless of the type of relationship (e.g., linear, non-linear, non-functional).

In [Res+16], they present an improved estimator, called **MICe**, which is more efficient to compute, and only requires optimizing over 1d grids, which can be done in $O(n)$ time using dynamic programming. They also present another quantity, called **TICe** (total information content), that has higher power to detect relationships from small sample sizes, but lower equitability. This is defined to be $\sum_{kl \leq B(n)} M(k, l)$. They recommend using TICe to screen a large number of candidate relationships, and then using MICe to quantify the strength of the relationship. For an efficient implementation of both of these metrics, see [Alb+18].

We can interpret MIC of 0 to mean there is no relationship between the variables, and 1 to represent a noise-free relationship of any form. This is illustrated in Figure 6.6. Unlike correlation coefficients, MIC is not restricted to finding linear relationships. For this reason, the MIC has been called "a correlation for the 21st century" [Spe11].

In Figure 6.7, we give a more interesting example, from [Res+11]. The data consists of 357 variables measuring a variety of social, economic, health and political indicators, collected by the World Health Organization (WHO). On the left of the figure, we see the correlation coefficient (CC) plotted against the MIC for all 63,546 variable pairs. On the right of the figure, we see scatter plots for particular pairs of variables, which we now discuss:

- The point marked C (near 0,0 on the plot) has a low CC and a low MIC. The corresponding scatter plot makes it clear that there is no relationship between these two variables (percentage of lives lost to injury and density of dentists in the population).
- The points marked D and H have high CC (in absolute value) and high MIC, because they

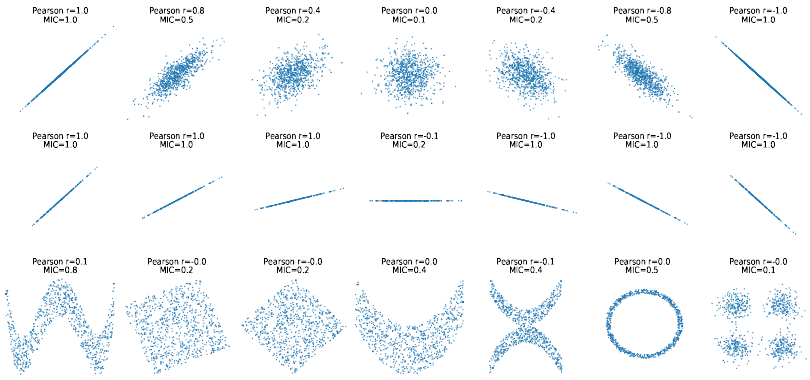


Figure 6.6: Plots of some 2d distributions and the corresponding estimate of correlation coefficient R^2 and the maximal information coefficient (MIC). Compare to Figure 3.1. Generated by code at figures.probml.ai/book1/6.6.

represent nearly linear relationships.

- The points marked E, F, and G have low CC but high MIC. This is because they correspond to non-linear (and sometimes, as in the case of E and F, non-functional, i.e., one-to-many) relationships between the variables.

6.3.8 Data processing inequality

Suppose we have an unknown variable X, and we observe a noisy function of it, call it Y. If we process the noisy observations in some way to create a new variable Z, it should be intuitively obvious that we cannot increase the amount of information we have about the unknown quantity, X. This is known as the **data processing inequality**. We now state this more formally, and then prove it.

Theorem 6.3.1. *Suppose $X \to Y \to Z$ forms a Markov chain, so that $X \perp Z|Y$. Then $\mathbb{I}(X;Y) \geq \mathbb{I}(X;Z)$.*

Proof. By the chain rule for mutual information (Equation (6.62)), we can expand the mutual information in two different ways:

$$\mathbb{I}(X;Y,Z) = \mathbb{I}(X;Z) + \mathbb{I}(X;Y|Z) \tag{6.77}$$

$$= \mathbb{I}(X;Y) + \mathbb{I}(X;Z|Y) \tag{6.78}$$

Since $X \perp Z|Y$, we have $\mathbb{I}(X;Z|Y) = 0$, so

$$\mathbb{I}(X;Z) + \mathbb{I}(X;Y|Z) = \mathbb{I}(X;Y) \tag{6.79}$$

Since $\mathbb{I}(X;Y|Z) \geq 0$, we have $\mathbb{I}(X;Y) \geq \mathbb{I}(X;Z)$. Similarly one can prove that $\mathbb{I}(Y;Z) \geq \mathbb{I}(X;Z)$. □

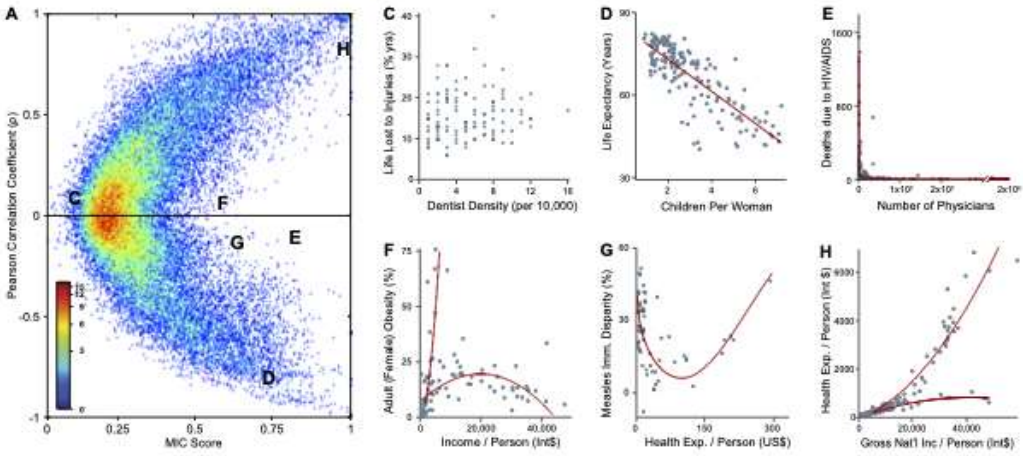


Figure 6.7: Left: Correlation coefficient vs maximal information criterion (MIC) for all pairwise relationships in the WHO data. Right: scatter plots of certain pairs of variables. The red lines are non-parametric smoothing regressions fit separately to each trend. From Figure 4 of [Res+11]. Used with kind permission of David Reshef.

6.3.9 Sufficient Statistics

An important consequence of the DPI is the following. Suppose we have the chain $\theta \to \mathcal{D} \to s(\mathcal{D})$. Then

$$\mathbb{I}(\theta; s(\mathcal{D})) \leq \mathbb{I}(\theta; \mathcal{D}) \tag{6.80}$$

If this holds with equality, then we say that $s(\mathcal{D})$ is a **sufficient statistic** of the data $\mathcal{D}$ for the purposes of inferring θ. In this case, we can equivalently write $\theta \to s(\mathcal{D}) \to \mathcal{D}$, since we can reconstruct the data from knowing $s(\mathcal{D})$ just as accurately as from knowing θ.

An example of a sufficient statistic is the data itself, $s(\mathcal{D}) = \mathcal{D}$, but this is not very useful, since it doesn't summarize the data at all. Hence we define a **minimal sufficient statistic** $s(\mathcal{D})$ as one which is sufficient, and which contains no extra information about θ; thus $s(\mathcal{D})$ maximally compresses the data $\mathcal{D}$ without losing information which is relevant to predicting θ. More formally, we say s is a *minimal* sufficient statistic for $\mathcal{D}$ if for all sufficient statistics $s'(\mathcal{D})$ there is some function f such that $s(\mathcal{D}) = f(s'(\mathcal{D}))$. We can summarize the situation as follows:

$$\theta \to s(\mathcal{D}) \to s'(\mathcal{D}) \to \mathcal{D} \tag{6.81}$$

Here $s'(\mathcal{D})$ takes $s(\mathcal{D})$ and adds redundant information to it, thus creating a one-to-many mapping.

For example, a minimal sufficient statistic for a set of N Bernoulli trials is simply N and $N_1 = \sum_n \mathbb{I}(X_n = 1)$, i.e., the number of successes. In other words, we don't need to keep track of the entire sequence of heads and tails and their ordering, we only need to keep track of the total number of heads and tails. Similarly, for inferring the mean of a Gaussian distribution with known variance we only need to know the empirical mean and number of samples.

6.3.10 Fano's inequality *

A common method for **feature selection** is to pick input features X_d which have high mutual information with the response variable Y. Below we justify why this is a reasonable thing to do. In particular, we state a result, known as **Fano's inequality**, which bounds the probability of misclassification (for any method) in terms of the mutual information between the features X and the class label Y.

Theorem 6.3.2. *(Fano's inequality) Consider an estimator $\hat{Y} = f(X)$ such that $Y \to X \to \hat{Y}$ forms a Markov chain. Let E be the event $\hat{Y} \neq Y$, indicating that an error occured, and let $P_e = P(Y \neq \hat{Y})$ be the probability of error. Then we have*

$$\mathbb{H}(Y|X) \leq \mathbb{H}\left(Y|\hat{Y}\right) \leq \mathbb{H}(E) + P_e \log |\mathcal{Y}| \tag{6.82}$$

Since $\mathbb{H}(E) \leq 1$, as we saw in Figure 6.1, we can weaken this result to get

$$1 + P_e \log |\mathcal{Y}| \geq \mathbb{H}(Y|X) \tag{6.83}$$

and hence

$$P_e \geq \frac{\mathbb{H}(Y|X) - 1}{\log |\mathcal{Y}|} \tag{6.84}$$

Thus minimizing $\mathbb{H}(Y|X)$ (which can be done by maximizing $\mathbb{I}(X;Y)$) will also minimize the lower bound on P_e.

Proof. (From [CT06, p38].) Using the chain rule for entropy, we have

$$\mathbb{H}\left(E, Y|\hat{Y}\right) = \mathbb{H}\left(Y|\hat{Y}\right) + \underbrace{\mathbb{H}\left(E|Y, \hat{Y}\right)}_{=0} \tag{6.85}$$

$$= \mathbb{H}\left(E|\hat{Y}\right) + \mathbb{H}\left(Y|E, \hat{Y}\right) \tag{6.86}$$

Since conditioning reduces entropy (see Section 6.2.4), we have $\mathbb{H}\left(E|\hat{Y}\right) \leq \mathbb{H}(E)$. The final term can be bounded as follows:

$$\mathbb{H}\left(Y|E, \hat{Y}\right) = P(E=0)\,\mathbb{H}\left(Y|\hat{Y}, E=0\right) + P(E=1)\,\mathbb{H}\left(Y|\hat{Y}, E=1\right) \tag{6.87}$$

$$\leq (1 - P_e)0 + P_e \log |\mathcal{Y}| \tag{6.88}$$

Hence

$$\mathbb{H}\left(Y|\hat{Y}\right) \leq \underbrace{\mathbb{H}\left(E|\hat{Y}\right)}_{\leq \mathbb{H}(E)} + \underbrace{\mathbb{H}\left(Y|E, \hat{Y}\right)}_{P_e \log |\mathcal{Y}|} \tag{6.89}$$

Finally, by the data processing inequality, we have $\mathbb{I}(Y;\hat{Y}) \leq \mathbb{I}(Y;X)$, so $\mathbb{H}(Y|X) \leq \mathbb{H}\left(Y|\hat{Y}\right)$, which establishes Equation (6.82). □

6.4 Exercises

Exercise 6.1 [Expressing mutual information in terms of entropies *]
Prove the following identities:

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{6.90}$$

and

$$H(X,Y) = H(X|Y) + H(Y|X) + I(X;Y) \tag{6.91}$$

Exercise 6.2 [Relationship between $D(p||q)$ and χ^2 statistic]
(Source: [CT91, Q12.2].)
Show that, if $p(x) \approx q(x)$, then

$$D_{\mathbb{KL}}(p\|q) \approx \frac{1}{2}\chi^2 \tag{6.92}$$

where

$$\chi^2 = \sum_x \frac{(p(x) - q(x))^2}{q(x)} \tag{6.93}$$

Hint: write

$$p(x) = \Delta(x) + q(x) \tag{6.94}$$

$$\frac{p(x)}{q(x)} = 1 + \frac{\Delta(x)}{q(x)} \tag{6.95}$$

and use the Taylor series expansion for $\log(1+x)$.

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} \cdots \tag{6.96}$$

for $-1 < x \leq 1$.

Exercise 6.3 [Fun with entropies *]
(Source: Mackay.) Consider the joint distribution $p(X,Y)$

		x			
		1	2	3	4
y	1	1/8	1/16	1/32	1/32
	2	1/16	1/8	1/32	1/32
	3	1/16	1/16	1/16	1/16
	4	1/4	0	0	0

a. What is the joint entropy $H(X,Y)$?
b. What are the marginal entropies $H(X)$ and $H(Y)$?
c. The entropy of X conditioned on a specific value of y is defined as

$$H(X|Y=y) = -\sum_x p(x|y) \log p(x|y) \tag{6.97}$$

Compute $H(X|y)$ for each value of y. Does the posterior entropy on X ever increase given an observation of Y?

d. The conditional entropy is defined as

$$H(X|Y) = \sum_y p(y) H(X|Y=y) \tag{6.98}$$

Compute this. Does the posterior entropy on X increase or decrease when averaged over the possible values of Y?

e. What is the mutual information between X and Y?

Exercise 6.4 [Forwards vs reverse KL divergence]

(Source: Exercise 33.7 of [Mac03].) Consider a factored approximation $q(x, y) = q(x)q(y)$ to a joint distribution $p(x, y)$. Show that to minimize the forwards KL $D_{\mathbb{KL}}(p\|q)$ we should set $q(x) = p(x)$ and $q(y) = p(y)$, i.e., the optimal approximation is a product of marginals

Now consider the following joint distribution, where the rows represent y and the columns x.

	1	2	3	4
1	1/8	1/8	0	0
2	1/8	1/8	0	0
3	0	0	1/4	0
4	0	0	0	1/4

Show that the reverse KL $D_{\mathbb{KL}}(q\|p)$ for this p has three distinct minima. Identify those minima and evaluate $D_{\mathbb{KL}}(q\|p)$ at each of them. What is the value of $D_{\mathbb{KL}}(q\|p)$ if we set $q(x, y) = p(x)p(y)$?