Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

13 Inference as Decision-Making

We now begin to uncover the role of information-theoretic measures in problems of inference. We do this in a manner that at first glance might seem somewhat surprising. Indeed, we introduced decision theory as a special case of statistical inference. Now we will turn the tables and interpret statistical inference as a special case of decision theory.

For the time being, we will emphasize a Bayesian viewpoint. For convenience, we will also restrict our attention to the case of discrete (and scalar) x and y, with $\mathcal{X}$ denoting, as usual, the alphabet of possible values that x can take on.

13.1 Generalized Bayesian Decision Theory

In our original development of Bayesian decision theory, our decision device took as input some data y, a cost criterion $C(\cdot,\cdot)$, an observation model $p_{\mathsf{y}|\mathsf{x}}(\cdot|\cdot)$, and a prior distribution (belief) $p_{\mathsf{x}}(\cdot)$, and produced a "hard" decision $\hat{\mathsf{x}}$. In particular, for every possible value of the data y, the decision device produces a number $\hat{x}(y)$.

We now consider a generalization in which our decision device produces not a number but a probability distribution $q(\cdot)$ that describes the relative likelihood of different elements of $\mathcal{X}$ based on the observed data $\mathsf{y} = y$. In other words, the decision device produces a "soft" decision.

Of course, ideally the decision device would produce the distribution

$$q(a) = \mathbb{1}_x(a) \triangleq \begin{cases} 1 & a = x \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

since this would identify the realized x with certainty. However, this is generally not possible since $q(\cdot)$ is constructed from the data and thus cannot depend explicitly on the unknown x.

Similar to hypothesis testing and estimation, we start by defining the cost criterion $C(\cdot,\cdot)$. There are lots of possible cost criteria one could use with such a generalized decision device. One example is the quadratic cost criterion

$$C(x,q) = A\sum_a (q(a) - \mathbb{1}_x(a))^2 + B(x), \tag{2}$$

where $A > 0$ and $B(\cdot)$ is arbitrary. This cost criterion identifies the closest match to (1) in a Euclidean sense. Another example is the log-loss criterion

$$C(x,q) = -A\log q(x) + B(x), \tag{3}$$

where, again, $A > 0$ and $B(\cdot)$ is arbitrary. The log-loss criterion emphasizes distributions $q(\cdot)$ that are as "peaky" as possible, and thus most concentrated.

Among all possible cost functions, some are of more interest to us than others. Below, we define two important desired properties of cost functions and show that the log-loss cost criterion possesses them both.

Definition 1. *A cost function $C(\cdot, \cdot)$ is* proper *if*

$$p_{\mathsf{x}|\mathsf{y}}(\cdot|y) = \underset{\{q(\cdot)\geq 0:\ \sum_a q(a)=1\}}{\arg\min} \mathbb{E}\left[C(x, q)|\mathsf{y} = y\right] \quad \textit{for all } y \in \mathcal{Y}. \tag{4}$$

A proper cost function leads to the true belief $p_{\mathsf{x}|\mathsf{y}}$ being reported, i.e., it corresponds to an "honest" Bayes decision device.

Claim 1. *The log-loss cost criterion is proper.*

Proof. We prove the claim through a simple application of the Gibbs' inequality:

$$\begin{aligned} \mathbb{E}\left[-A \log q(\mathsf{x}) + B(\mathsf{x})|\mathsf{y} = y\right] &= -A\,\mathbb{E}\left[\log q(\mathsf{x})|\mathsf{y} = y\right] + \mathbb{E}\left[B(\mathsf{x})|\mathsf{y} = y\right] && (5)\\ &\geq -A\,\mathbb{E}\left[\log p_{\mathsf{x}|\mathsf{y}}(x|y)\right] + \mathbb{E}\left[B(\mathsf{x})|\mathsf{y} = y\right], && (6) \end{aligned}$$

where (6) is true since the expectation is computed with respect to $p_{\mathsf{x}|\mathsf{y}}$. Equality holds if and only if $q(\cdot) \equiv p_{\mathsf{x}|\mathsf{y}}(\cdot|y)$, which proves that the log-loss is a proper cost criterion. □

This claim implies that if we adopt the log-loss criterion in our optimization, and succeed in computing the minimum in (4), we will obtain exactly the posterior probability distribution $p_{\mathsf{x}|\mathsf{y}}(\cdot|y)$.

Definition 2. *A cost function $C(\cdot, \cdot)$ is* local *if there exists a function $\phi : \mathcal{X} \times \mathbb{R} \mapsto \mathbb{R}$ such that $C(x, q) = \phi(x, q(x))$ for all $x \in \mathcal{X}$.*

A local cost function assesses the quality of the estimated belief q only in terms of the probability assigned to the actual outcome x, and is thus associated with "pure" inference.

Claim 2. *The log-loss cost criterion is local.*

Proof. It suffices to note that $\phi(\mu, \nu) = -A \log \nu + B(\mu)$. □

There are other proper cost functions, and other local cost functions. However, the log-loss cost function is special, as the following characterization theorem establishes.

Theorem 1. *When the alphabet $\mathcal{X}$ consists of at least three values ($|\mathcal{X}| \triangleq L \geq 3$), then the log-loss is the* only *smooth local, proper cost function.*

Proof. For convenience, let $\mathcal{X} = \{x_1, \ldots, x_L\}$. In turn, let $q_l \triangleq q(x_l)$ and $p_l \triangleq p_{\mathsf{x}|\mathsf{y}}(x_l|y)$ for $l = 1, \ldots, L$. Finally, let $\phi_l(\cdot) \triangleq \phi(x_l, \cdot)$ be our compact notation for a local cost function.

Proceeding, we seek constraints on $\phi_1, \ldots, \phi_L$ such that for any choice of $p_1, \ldots, p_L$, the cost function is proper, i.e.,

$$p_1, \ldots, p_L = \underset{\left\{q_1,\ldots,q_L \geq 0 \,:\, \sum_{l=1}^{L} q_l = 1\right\}}{\arg\min} \sum_{l=1}^{L} p_l \phi_l(q_l), \tag{7}$$

where we have used that

$$\mathbb{E}\left[\phi(\mathsf{x}, q(\mathsf{x}))|\mathsf{y} = y\right] = \sum_{l=1}^{L} p_l \phi_l(q_l).$$

Equivalently, we may rewrite (7) as

$$p_1, \ldots, p_L = \underset{q_1,\ldots,q_L}{\arg\min}\, \varphi, \quad \text{with } \varphi = \sum_{l=1}^{L} p_l \phi_l(q_l) + \lambda(p_1, \ldots, p_L) \left[\sum_{l=1}^{L} q_l - 1\right], \tag{8}$$

where $\lambda(p_1, \ldots, p_L) \geq 0$ denotes a Lagrange multiplier.

Since the ϕ_l are smooth, we may obtain the constraints from (8) by examining the stationary point of the objective function, i.e.,

$$\left.\frac{\partial \varphi}{\partial q_k}\right|_{q_l = p_l,\ l=1,\ldots,L} = p_k \dot{\phi}_k(p_k) + \lambda(p_1, \ldots, p_L) = 0, \quad k = 1, \ldots, L$$

with $\dot{\phi}_l(\cdot)$ denoting the derivative of $\phi_l(\cdot)$, whence

$$p_k \dot{\phi}_k(p_k) = -\lambda(p_1, \ldots, p_L), \quad k = 1, \ldots, L. \tag{9}$$

Since, of course, $\phi_k(\cdot)$ does not have any implicit dependence on the $p_1, \ldots, p_L$, we see that the right-hand side of (9) cannot depend on p_l for $l \neq k$, unless $L = 2$ in which case $p_2 = 1 - p_1$. Since this is true for each $k \in \{1, \ldots, L\}$, we conclude that for $L \geq 3$, it must be that $\lambda(p_1, \ldots, p_L)$ is simply a constant λ that does not depend on $p_1, \ldots, p_L$. Hence, specializing (9), we obtain that $\phi_k(\cdot)$ is a solution to the differential equation

$$\dot{\phi}_k(q) = -\frac{\lambda}{q}, \tag{10}$$

i.e.,

$$\phi_k(q) = -\lambda \ln q + c_k, \quad k = 1, \ldots, L \tag{11}$$

where $\lambda > 0$ and c_k are arbitrary constants. Returning to our original notation, this means that

$$\phi(x, q(x)) = -A \log q(x) + B(x)$$

for $A = \lambda / \log(e)$ and $B(x_k) = c_k$.[1] $\square$

Henceforth, we will, without loss of generality, restrict our attention to the version of the log-loss cost function with $A = 1$ and $B(\cdot) = 0$. Moreover, the associated log will be either base 2 (i.e., $\log_2$) or base e (i.e., ln), depending on whether we wish to measure the cost in bits or nats, as we will later interpret.

13.2 Self-Information Loss and Entropy

Let us now study the resulting *minimum* expected cost associated with the log-loss criterion.

We begin with the cost in absence of observations, so $p_{\mathsf{x}|\mathsf{y}} = p_{\mathsf{x}}$. Since the log-loss cost criterion is proper, we know that the prior p_{x} achieves the minimum. The "prior" cost is

$$\begin{aligned}
\min_{\{q(\cdot)\geq 0:\ \sum_a q(a)=1\}} \mathbb{E}\left[C(\mathsf{x}, q)\right] &= \mathbb{E}\left[C(\mathsf{x}, p_{\mathsf{x}})\right] && (12)\\
&= -\,\mathbb{E}\left[\log p_{\mathsf{x}}(\mathsf{x})\right] && (13)\\
&= -\sum_a p_{\mathsf{x}}(a) \log p_{\mathsf{x}}(a) \triangleq H(\mathsf{x}). && (14)
\end{aligned}$$

Eq. (14) defines a very special information-theoretic quantity; it is referred to as the *entropy*, or "self-information", of x. The entropy is a measure of the (average) uncertainty or randomness in x. It is measured in bits (when base-2 logarithms are used). In an appropriate sense, $H(\mathsf{x})$ reflects the number of bits needed to describe x.

It is also worth emphasizing that entropy is a measure of uncertainty that does not depend on how the symbols in the alphabet are labeled, just their probabilities (for example, the elements could be $\{\heartsuit, \spadesuit, \diamondsuit, \clubsuit, \dots\}$ or $\{0, 1, 2, 3, \dots\}$ or ...). In contrast, for example, variance as a measure of uncertainty is only defined for *numeric* random variables.

The entropy has many interesting properties, some of which we show here.

Claim 3. *Let x be a discrete random variable defined over an alphabet $\mathcal{X}$. Then*

$$0 \leq H(\mathsf{x}) \leq \log |\mathcal{X}|, \tag{15}$$

where the lower bound is tight when x is deterministic, and the upper bound is tight when x is uniformly distributed over $\mathcal{X}$.

We leave the proof as an exercise.

[1] Note that we have used the easily verified log-conversion identity: $\log(u) = \log(e) \ln(u)$ valid for all $u > 0$.

Finally, sometimes it will be convenient to use the alternative notation $H(p)$ to denote the entropy of a random variable distributed according to distribution $p(\cdot)$.

Whenever dealing with information measures, we adopt the convention that $0 \log 0 \equiv 0$. Note that with this convention, $H(p)$ is a continuous function of p. This follows from a simple application of L'Hôpital's Rule:

$$\lim_{t \to 0} t \log t = \lim_{t \to 0} \frac{\log t}{1/t} = \lim_{t \to 0} \frac{\log(e)/t}{-1/t^2} = 0.$$

Example 1. Let v be a Bernoulli random variable, i.e., $\mathsf{v} \sim \mathsf{B}(p)$ for some parameter p, which is the probability of one of the two possible symbols. Then

$$H(\mathsf{v}) = H_{\mathrm{B}}(p) \triangleq -p \log p - (1-p) \log(1-p), \tag{16}$$

where $H_{\mathrm{B}}(\cdot)$ is referred to as the binary entropy function. It satisfies $H_{\mathrm{B}}(0) = H_{\mathrm{B}}(1) = 0$ and $H_{\mathrm{B}}(1/2) = 1$ bit.

Once we observe $\mathsf{y} = y$, the log-loss is minimized for the posterior probability distribution $p_{\mathsf{x}|\mathsf{y}}(\cdot|y)$. The resulting "posterior" cost is

$$\begin{aligned}
\min_{\{q(\cdot) \geq 0:\ \sum_a q(a) = 1\}} \mathbb{E}\left[C(\mathsf{x}, q) | \mathsf{y} = y\right] &= \mathbb{E}\left[C(\mathsf{x}, p_{\mathsf{x}|\mathsf{y}}) | \mathsf{y} = y\right] && (17)\\
&= -\,\mathbb{E}\left[\log p_{\mathsf{x}|\mathsf{y}}(\mathsf{x}|\mathsf{y}) | \mathsf{y} = y\right] && (18)\\
&= -\sum_a p_{\mathsf{x}|\mathsf{y}}(a|y) \log p_{\mathsf{x}|\mathsf{y}}(a|y) \triangleq H(\mathsf{x}|\mathsf{y} = y), && (19)
\end{aligned}$$

which is referred to as the *conditional* entropy of x given a particular observation $\mathsf{y} = y$. Taking the expectation of (19) over all possible values of the observation yields the average posterior cost

$$\begin{aligned}
\mathbb{E}\left[C(\mathsf{x}, p_{\mathsf{x}|\mathsf{y}})\right] &= \mathbb{E}\left[\mathbb{E}\left[C(\mathsf{x}, p_{\mathsf{x}|\mathsf{y}}(\mathsf{x}|\mathsf{y})) | \mathsf{y} = y\right]\right] && (20)\\
&= -\sum_y p_{\mathsf{y}}(y) H(\mathsf{x}|\mathsf{y} = y) && (21)\\
&= -\sum_{a,b} p_{\mathsf{x},\mathsf{y}}(a, b) \log p_{\mathsf{x}|\mathsf{y}}(a|b) \triangleq H(\mathsf{x}|\mathsf{y}), && (22)
\end{aligned}$$

which is referred to as simply the conditional entropy of x given y.

We see that $H(\mathsf{x}|\mathsf{y})$ is a measure of the uncertainty in x having observed y. It is straightforward to verify that

$$0 \leq H(\mathsf{x}|\mathsf{y}) \leq H(\mathsf{x}), \tag{23}$$

where the left inequality is tight if x is a deterministic function of y, and where the right inequality is tight if x is independent of y, which we frequently express using the shorthand notation $\mathsf{x} \perp\!\!\!\perp \mathsf{y}$. Note that the right inequality also implies that conditioning never increases uncertainty as measured by entropy, which is intuitively pleasing. Finally, we note that since $z = (x, y)$ is just another random variable, it is a simple exercise to verify that

$$H(z) = H(x, y) = \sum_{a,b} p_{x,y}(a, b) \log p_{x,y}(a, b) = H(x|y) + H(y), \tag{24}$$

which can be referred to as the *joint* uncertainty in x and y.

13.3 Cost Reduction and Mutual Information

We can now quantify the (average) cost reduction achieved by making an observation in the case of the log-loss cost function. In particular, the difference between the prior and posterior expected costs

$$\Delta \, \mathbb{E}\left[C(x, q)\right] = H(x) - H(x|y) \triangleq I(x; y) \tag{25}$$

is the *mutual information* between x and y. It can be equivalently expressed in the form

$$I(x; y) = \sum_{a,b} p_{x,y}(a, b) \log \frac{p_{x,y}(a, b)}{p_x(a) p_y(b)}. \tag{26}$$

The mutual information between two random variables x and y is a fundamental quantity. It characterizes how much y tells us (on average) about x for the purposes of inference. More specifically, it reflects how many fewer bits it takes to describe x once y is observed.

The following properties follow readily:

$$\begin{align} I(x; y) &\geq 0 \quad \text{(nonnegativity)} \tag{27} \\ I(x; y) &= 0 \quad \text{if and only if } x \perp\!\!\!\perp y \tag{28} \\ I(x; y) &= I(y; x) \quad \text{(symmetry)} \tag{29} \\ I(x; y, z) &= I(x; z) + I(x; y|z) \quad \text{(chain rule)}, \tag{30} \end{align}$$

where in the chain rule, the conditional mutual information is defined as

$$I(x; y|z) \triangleq H(x|z) - H(x|y, z). \tag{31}$$

A variety of useful identities can be verified; the Venn diagram of Fig. 1 serves as a simple—albeit imperfect—mnemonic device. For example, we have

$$H(x, y) = H(x) + H(y) - I(x; y). \tag{32}$$

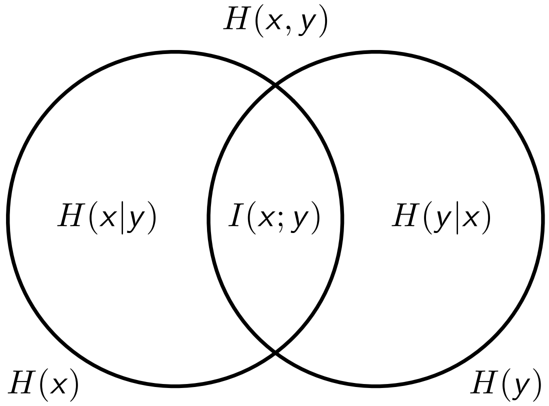

Figure 1: Venn diagram of relationships between mutual information and entropies.

13.4 Sufficient Statistics and the Data Processing Inequality

The notion of a sufficient statistic, as developed earlier, also has a useful information-theoretic interpretation, as the following theorem establishes.

Theorem 2 (Data processing inequality)**.** *If $x \leftrightarrow y \leftrightarrow t$ is a Markov chain, i.e., if t is any statistic, then*

$$I(x;y) \geq I(x;t), \tag{33}$$

with equality if and only if $x \leftrightarrow t \leftrightarrow y$ is a Markov chain.

Proof. We have

$$\begin{aligned} I(x;t) &= I(x;y,t) - I(x;y|t) && (34)\\ &= I(x;y) + I(x;t|y) - I(x;y|t) && (35)\\ &= I(x;y) - I(x;y|t) && (36)\\ &\leq I(x;y) && (37) \end{aligned}$$

where to obtain (34) we have used the chain rule of mutual information, to obtain (35) we have used the other possible chain rule expansion, to obtain (36) we have used that x and t are conditionally independent (given y) due to the Markov structure, and to obtain (37) we have used that mutual information is nonnegative. Moreover, the application of (28) in the world conditioned on t implies that the equality in (37) holds if and only if x and y are conditionally independent given t, which corresponds to $x \leftrightarrow t \leftrightarrow y$ forming a Markov chain. □

The following corollary immediately follows from Theorem 2.

Lemma 1. *A statistic $t = t(y)$ is* sufficient *if and only if $I(x;t) = I(x;y)$.*

In essence, if t provides the same cost reduction as does y directly, then from the point of view of inference, t summarizes everything we need to know about y. Thus after constructing t from y, we may discard y.

Theorem 2 says that (pre)processing the data can never give greater cost reduction, and also reminds us of our earlier Markov chain interpretation of sufficiency. Note the following special case of Theorem 2:

Corollary 1. *For any deterministic function $g(\cdot)$, $I(\mathsf{x};\mathsf{y}) \geq I(\mathsf{x};g(\mathsf{y}))$.*

13.5 Imperfect Inference and Information Divergence

Sometimes we cannot implement true beliefs, but must approximate them. In such cases, we would like to quantify the approximation loss. For example, if $p(\cdot)$ represents the true (prior or posterior) belief, but we must approximate it with some distribution $q(\cdot)$, then the approximation loss can be expressed as

$$\begin{aligned} \Delta\,\mathbb{E}\left[C(\mathsf{x},q)\right] &= -\,\mathbb{E}_p\left[\log q(\mathsf{x})\right] + \mathbb{E}_p\left[\log p(\mathsf{x})\right] && (38)\\ &= \mathbb{E}_p\left[\log\frac{p(\mathsf{x})}{q(\mathsf{x})}\right] \triangleq D(p\|q), && (39)\end{aligned}$$

where

$$D(p\|q) = \sum_a p(a)\log\frac{p(a)}{q(a)} \tag{40}$$

is referred to as the (information) *divergence* of $q(\cdot)$ from $p_{\mathsf{x}}(\cdot)$, or, alternatively, the *relative entropy* of $q(\cdot)$ with respect to $p_{\mathsf{x}}(\cdot)$. It is also sometimes referred to as the Kullback-Leibler (KL) distance of $q(\cdot)$ from $p_{\mathsf{x}}(\cdot)$, though it is not a true distance—indeed, it is not even symmetric: $D(p\|q) \neq D(q\|p)$.

The divergence can be interpreted as the number of additional bits required to represent x by virtue of not using its correct distribution. The divergence is perhaps the most fundamental information theoretic quantity, and will play a central role in the rich information geometry we are to develop.

Gibbs' inequality implies

$$D(p\|q) \geq 0 \tag{41}$$

for any distributions p and q.

Other useful identities include

$$D(p\|\mathtt{U}) = \log|\mathcal{X}| - H(p), \tag{42}$$

where $\mathtt{U}$ denotes the uniform distribution over the alphabet $\mathcal{X}$, and

$$D(p_{\mathsf{x},\mathsf{y}}\|p_{\mathsf{x}}p_{\mathsf{y}}) = I(\mathsf{x};\mathsf{y}). \tag{43}$$

Note, too, that with our convention $0\log(0) \equiv 0$, information divergence $D(p\|q)$ is a continuous function of p and q.

Example 2. Let $p = \mathtt{B}(\epsilon)$ and $q = \mathtt{B}(\delta)$. Then

$$D(p\|q) = \epsilon \log \frac{\epsilon}{\delta} + (1-\epsilon) \log \frac{1-\epsilon}{1-\delta} \triangleq D_{\mathrm{B}}(\epsilon\|\delta). \tag{44}$$

You may find it insightful to sketch $D_{\mathrm{B}}(\epsilon\|\delta)$ as a function of δ for various fixed values of ϵ.

As a final comment, note that expressing the approximation loss for posterior beliefs requires a little additional care. In particular, divergence extends naturally to conditional distributions via

$$D(p_{\mathsf{x}|\mathsf{y}}(\cdot|y) \parallel q_{\mathsf{x}|\mathsf{y}}(\cdot|y)) = \mathbb{E}_{p_{\mathsf{x}|\mathsf{y}}(\cdot|y)} \left[\log \frac{p_{\mathsf{x}|\mathsf{y}}(\cdot|y)}{q_{\mathsf{x}|\mathsf{y}}(\cdot|y)} \right] = \sum_a p_{\mathsf{x}|\mathsf{y}}(a|y) \log \frac{p_{\mathsf{x}|\mathsf{y}}(a|y)}{q_{\mathsf{x}|\mathsf{y}}(a|y)}. \tag{45}$$

However, to obtain the loss in approximating a true posterior $p_{\mathsf{x}|\mathsf{y}}$ by $q_{\mathsf{x}|\mathsf{y}}$, we must average (45) over all y, i.e.,

$$\begin{aligned} \Delta\, \mathbb{E}\left[C(\mathsf{x}, q)\right] &= \mathbb{E}_{p_{\mathsf{y}}} \left[D(p_{\mathsf{x}|\mathsf{y}}(\cdot|\mathsf{y}) \parallel q_{\mathsf{x}|\mathsf{y}}(\cdot|\mathsf{y})) \right] \\ &= \sum_b p_{\mathsf{y}}(b)\, D(p_{\mathsf{x}|\mathsf{y}}(\cdot|b) \parallel q_{\mathsf{x}|\mathsf{y}}(\cdot|b)) \\ &= \sum_{a,b} p_{\mathsf{x},\mathsf{y}}(a,b) \log \frac{p_{\mathsf{x}|\mathsf{y}}(a|b)}{q_{\mathsf{x}|\mathsf{y}}(a|b)} \\ &\triangleq \bar{D}(p_{\mathsf{x}|\mathsf{y}} \| q_{\mathsf{x}|\mathsf{y}}). \end{aligned}$$

Sometimes it is convenient to recognize that

$$\bar{D}(p_{\mathsf{x}|\mathsf{y}} \| q_{\mathsf{x}|\mathsf{y}}) = D(p_{\mathsf{x}|\mathsf{y}} p_{\mathsf{y}} \| q_{\mathsf{x}|\mathsf{y}} p_{\mathsf{y}}), \tag{46}$$

and, in addition, the alternative notation $D(p_{\mathsf{x}|\mathsf{y}} \| q_{\mathsf{x}|\mathsf{y}} | p_{\mathsf{y}})$ is sometimes used for (46) to make explicit the relevant distribution for the data, though we will avoid it.

13.6 Further reading

For additional detail and insights, see, e.g., the books by Bernardo and Smith, and by Cover and Thomas, as well as the paper by Feder and Merhav, although the latter will be more accessible later.