EE 229A

Venkat Anantharam

Davis Foote[*]

University of California, Berkeley

August 27^{th}, 2015 – December 10^{th}, 2015

Contents

1 Toward a Calculus of Information

1.1 *Defining Information*

Lecture 1
August 27^{th}, 2015

1.1.1 NOTATION. For this class, log refers to the logarithm in base 2.

Lecture 2
September 1^{st}, 2015

1.1.2 DEFINITION. Let $(p_1, \ldots, p_m)$ be a probability distribution on $\{1, \ldots, m\}$. The **entropy** of D is defined as

$$H(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i = \sum_{i=1}^{m} p_i \log \frac{1}{p_i}$$

[*]djfoote@berkeley.edu

This can be intuitively decomposed into two components: the coefficient p_i is the probability with which symbol i occurs, and $\log \frac{1}{p_i}$ is the information content inherent to symbol i as a member of the distribution $(p_1, \ldots, p_m)$.

1.1.3 NOTATION. If X is a random variable taking values in a finite set $\mathcal{X}$, we write $H(X)$ to mean $H((p_X(x) : x \in \mathcal{X}))$.

1.1.4 NOTATION. A somewhat dangerous but common convention is to drop the subscript in $p_X(x)$ to just be $p(x)$.

The first and second derivatives of log are:

1. $\frac{\mathrm{d}}{\mathrm{d}u} \log u = (\log u)' = (\log e)\frac{1}{u} > 0$ for $u > 0$
2. $\frac{\mathrm{d}^2}{\mathrm{d}u^2} \log u = -\frac{\log e}{u^2} < 0$ for $u > 0$

Since the derivative is nonincreasing this is a "concave" function.

1.1.5 DEFINITION. A **concave function** is the negative of a convex function.

1.1.6 DEFINITION. A **convex function** is a real-valued function on a convex set, such that

$$f(\eta x_1 + (1-\eta)x_0) \leq \eta f(x_1) + (1-\eta) f(x_0)$$

for all x_0, x_1 in the domain and all $\eta \in [0,1]$

1.1.7 DEFINITION. A **convex set** is a subset C of $\mathbb{R}^d$ (for some $d \geq 1$) such that for all $x_0, x_1 \in C$ and all $\eta \in [0,1]$, $\eta x_1 + (1-\eta)x_0 \in C$.

Note that since the domain of a convex function must be a convex set, the definition given for convex function makes sense.

1.1.8 EXAMPLE. Consider a convex subset C of the real line. Given any two points in C, every point on the real line betwee these two points must also be in C.

1.1.9 THEOREM. *For a real-valued function f on a convex subset C of the real line, if f has a nonnegative second derivative it is convex. Hence if f has a nonpositive second derivative it is concave.*

TODO: include drawing of $u \mapsto u \log u$

Need to first understand the limit as u goes to 0 from the right, $\lim_{u \to 0^+} u \log u$. In fact, this limit is 0 because

$$u \log u = -u \log \frac{1}{u} = \frac{\log \frac{1}{u}}{2^{\log \frac{1}{u}}}$$

Since the numerator is approaching infinity linearly in $\log \frac{1}{u}$ and the denominator is approaching infinity exponentially in $\log \frac{1}{u}$, the limit must be 0.

The first and second derivatives of $u \log u$ are

- $\frac{\mathrm{d}}{\mathrm{d}u} u \log u = (u \log u)' = \log u + \log e$, which is negative if $u < 1/e$ and positive when $u > 1/e$
- $\frac{\mathrm{d}^2}{\mathrm{d}u^2} u \log u (u \log u)'' = (\log e) \frac{1}{u} > 0$ if $u > 0$, so the function is convex.

The purpose of this is to get a feeling for $H(p_1, \dots, p_m)$ as a function from probability distributions on $\{1, \dots, m\}$ to real numbers.

1.1.10 DEFINITION. When $m = 2$,

$$H(p, 1-p) = -p \log p - (1-p) \log(1-p)$$

This function is called the **binary entropy function**. The function is nonnegative, its maximum occurs at $p = \frac{1}{2}$, the derivative at 0 is $+\infty$, and the derivative at 1 is $-\infty$. The function is concave.

TODO: include drawing of binary entropy function

1.1.11 NOTATION. For binary distributions, we often just write $H(p)$.

1.1.12 DEFINITION. The set of probability distributions on $\{1, \dots, m\}$ can be visualized as a convex subset of $\mathbb{R}^m$, the **convex simplex** in $\mathbb{R}^m$.

1.1.13 EXAMPLE. For $m = 3$, this is the (filled in) triangle connecting the points $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$.

$H(p_1, \dots, p_m)$ viewed as a real-valued function on the unit simplex in $\mathbb{R}^m$ is nonnegative (because each $p_i \in [0,1]$). It is also concave, because given $\mathbf{p}^{(\mathbf{0})} = (p_1^{(0)}, \dots, p_m^{(0)})$ and $\mathbf{p}^{(\mathbf{1})} = (p_1^{(1)}, \dots, p_m^{(1)})$ and $\eta \in [0,1]$,

$$\eta \mathbf{p}^{(\mathbf{1})} + (1-\eta) \mathbf{p}^{(\mathbf{0})}$$

TODO: finish derivation using concavity of $-u \log u$

1.1.14 FACT. $H(p_1, \dots, p_m)$ is invariant to permutations of the coordinates. This is $m!$-fold symmetry. Its maximum occurs at $p_i = 1/m$ for $1 \leq i \leq m$. This should match intuition because the uniform distribution is the "most uncertain."

1.1.15 NOTATION. Consider two random variables X and Y; X taking values in $\mathcal{X}$ and Y taking values in $\mathcal{Y}$. Both $\mathcal{X}$ and $\mathcal{Y}$ are finite sets. They have joint probability distribution

$$p_{XY}(x, y) = P(X = x, Y = y)$$

which we'll abbreviate as $p(x,y)$.

1.1.16 DEFINITION. The **joint entropy** $H(x,y) = -\sum_{x,y} p(x,y) \log p(x,y)$.

1.1.17 EXAMPLE. Suppose $\mathcal{X} = \{1,2,3,4\}$, $\mathcal{Y} = \{1,2,3\}$, and $p(x,y)$ is uniform on $\{(x,y) \in \mathcal{X} \times \mathcal{Y} : x \geq y\}$.

$H(x,y) = \log 9$
$H(x) = \frac{1}{9}\log 9 + \frac{2}{9}\log\frac{9}{2} + 2 \cdot \frac{3}{9}\log\frac{9}{3} = \log 9 - \frac{2}{9}\log 2 - \frac{2}{3}\log 3$
$H(y)$ is computed similarly.

Suppose that X and Y are not independent; that is, knowing X gives you some information about Y. Intuitively one should expect, then, that the total uncertainty between X and Y is less than the sum of their individual uncertainties.

1.1.18 NOTATION. Let us condition on the event $\{Y = y\}$ for some $y \in \mathcal{Y}$. The conditional distribution of X fiven that $Y = y$ is

$$p_{X|Y}(x|y), x \in \mathcal{X}$$

(which we'll abbreviate as $p(x|y)$).

1.1.19 NOTATION. This conditional probability distribution has an entropy. According to the formula, this entropy is

$$-\sum_{x\in\mathcal{X}} p(x|y)\log p(x|y)$$

which we denote as $H(X|Y=y)$.

1.1.20 DEFINITION. Let $p_Y(y) = P(Y=y) = \sum_{x'} p(x',y)$ be denoted $p(y)$. Then

$$\sum_{y\in\mathcal{Y}} p(y) H(X|Y=y)$$

is denoted $H(X|Y)$ and is called the **conditional entropy** of X given Y.

1.2 *Some Mechanics*

Lecture 3
September 3rd, 2015

1.2.1 THEOREM. $H(X,Y) = H(Y) + H(X|Y)$

Proof.

$$
\begin{aligned}
H(X,Y) &= -\sum_{x,y} p(x,y)\log p(x,y) \\
&= -\sum_{x,y} p(x,y)\log p(y) - \sum_{x,y} p(x,y)\log p(x|y) \\
&= -\sum_{y} p(y)\log p(y) - \sum_{y} p(y)\sum_{x} p(x|y)\log p(x|y) \\
&= H(Y) + \sum_{y} p(y) H(X|Y=y) \\
&= H(Y) + H(X|Y)
\end{aligned}
$$

□

Note that $H(X|Y) \neq H(Y|X)$ in general.

1.2.2 theorem (Chain rule for entropy). *Suppose you are given n random variables $X_1, \ldots, X_n$, each discrete and finite-valued. Then we have*

$$H(X_1,\ldots,X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_1,X_2) + \ldots + H(X_n|X_1,X_2,\ldots,X_n)$$

There are $n!$ valid such formulae.

This can be proved simply by induction using a similar derivation as that of the formula in the previous theorem.

Suppose we are given two discrete, finite-valued random variables X and Y. The expression $H(X) - H(X|Y)$ seems to capture the amount by which the uncertainty about X is reduced (on average) when learning Y.

1.2.3 definition. This quantity $H(X) - H(X|Y)$ is denoted $I(X;Y)$. It is also often written as $H(X\ Y)$. It is called the **mutual information** between X and Y.

1.2.4 theorem. $I(X;Y) = I(Y;X)$.

Proof.

$$
\begin{aligned}
I(X;Y) &\triangleq H(X) - H(X|Y) \\
&= -\sum_{x,y} p(x,y)\log p(x) + \sum_{x,y} p(y)p(x|y)\log p(x|y) \\
&= \sum_{x,y} \log p(x,y)\frac{p(x,y)}{p(x)p(y)} \\
&= I(Y;X)
\end{aligned}
$$

where the last equality holds because the expression in the second-to-last line is symmetric with respect to x and y. $\square$

1.2.5 DEFINITION. Given three random variables X, Y, and Z (all discrete, finite), the **conditional mutual information** between X and Y conditioned on $Z = z$ is

$$\begin{aligned} I(X;Y|Z=z) &\triangleq H(X|Z=z) - H(X|Y=y, Z=z) \\ &= \sum_{x,y} p(x,y|z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \end{aligned}$$

We then define the **conditional mutual information** between X and Y given Z as

$$I(X;Y|Z) \triangleq \sum_z p(z) I(X;Y|Z=z)$$

Note that

$$\begin{aligned} I(X;Y|Z=z) &= \sum_x p(x|z) \log \frac{1}{p(x|z)} + \sum_y p(y|z) \sum_x p(x|y,z) \log p(x|y,z) \\ &= H(X|Z=z) - H(X|Y, Z=z) \end{aligned}$$

where $H(X|Y,Z=z) \triangleq \sum_y P(Y=y|Z=z) H(X|Y=y, Z=z)$.

1.2.6 THEOREM (Chain rule for information). $I(X, Y_1, Y_2, \ldots, Y_n) = I(X;Y_1) + I(X;Y_2|Y_1) + \cdots + I(X;Y_n|Y_1, Y_2, \ldots, Y_{n-1})$.

Lecture 4
September 8^{th}, 2015

1.2.7 NOTATION.

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = \mathbb{E}\left[\log \frac{1}{p_X(X)}\right]$$

This is also written by abuse of notation as

$$H(X) = \mathbb{E}\left[\log \frac{1}{p(X)}\right]$$

Similarly,

$$H(X|Y) = \mathbb{E}\left[\log \frac{1}{p(X|Y)}\right]$$

1.2.8 THEOREM (Chain rule for mutual information).

$$I(X;Y_1,\ldots,Y_n) = I(X;Y_1) + I(X;Y_2 \mid Y_1) + I(X;Y_3 \mid Y_1, Y_2) + \cdots + I(X;Y_n \mid Y_1, \ldots, Y_{n-1})$$

There are $n!$ such formulae.

The proof of this theorem simply involves splitting each I into a difference of entropies and following algebraically.

1.3 *KL-Divergence*

1.3.1 DEFINITION.

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

where $(p(x), x \in \mathcal{X})$ and $(q(x), x \in \mathcal{X})$ are two probability distributions on the same finite set $\mathcal{X}$. This quantity has several names:

- The **relative entropy** of p with respect to q.
- The **information divergence** (or **divergence**) of p from q.
- The **Kullback-Liebler distance** (or **information distance**) of p from q.

This measure should be interpreted as a sort of distance measure between p and q (see below).

1.3.2 FACT.

$$I(X;Y) = D(p(x,y)\|p(x)p(y))$$

This is true because

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

Some facts about $D(p\|q)$:

- $D(p\|q)$ can equal $+\infty$. This happens if and only if there is some $x \in \mathcal{X}$ for which $q(x) = 0$ but $p(x) > 0$. The intuition here is that if something can show up for p and not q then we have an infinite ability to determine that we are dealing with p and not q if that value appears. Note that this cannot happen with $I(X;Y)$.
- $D(p\|q) \neq D(q\|p)$ in general. For example, one can be $+\infty$ and the other can be 0.

1.3.3 THEOREM (Jensen's Inequality). *If $f(u)$ is a convex function defined for u in a convex subset S of $\mathbb{R}^n$ for some $n \geq 1$ and U is a random variable taking values in the domain S of f, then*

$$\mathbb{E}[f(U)] \geq f(\mathbb{E}[U])$$

The proof comes from noting that a convex function at the centroid of a convex set should be lower than the polytope connecting the points $\{(u, f(u) : u \in S\}$.

1.3.4 DEFINITION. A convex function f defined on a convex set $C \subset \mathbb{R}^n$ is called **strictly convex** if $(1-\eta)f(u_0) + \eta f(u_1) > f((1-\eta)u_0 + \eta u_1)$ unless either $u_0 = u_1$ or $\eta = 0$ or $\eta = 1$.

A test for strict convexity is that the second derivative be strictly positive.

1.3.5 THEOREM. *It is always the case that $D(p\|q) \geq 0$. We have equality if and only if $p = q$.*

Proof. Write

$$D(p\|q) = \sum_x q(x) \left(\frac{p(x)}{q(x)} \log \frac{p(x)}{q(x)} \right)$$

and consider U taking values in $[0, \infty)$ taking the value $\frac{p(x)}{q(x)}$ with probability $q(x)$. Use Jensen's Inequality for the strictly convex function $f(u) = u \log u$. This gives

$$\begin{aligned} D(p\|q) &= \mathbb{E}[f(U)] \\ &\geq f(\mathbb{E}[U]) \\ &= \left(\sum_x q(x) \frac{p(x)}{q(x)} \right) \log \left(\sum_x q(x) \frac{p(x)}{q(x)} \right) \\ &= \left(\sum_x q(x) \frac{p(x)}{q(x)} \right) \log \left(\sum_x p(x) \right) \\ &= 0 \end{aligned}$$

Since f is strictly convex, we have inequality if and only if q is a one-point distribution (in which case $D(p\|q) = \infty$ unless p is also a one-point distribution on the same point) or if all $\frac{p(x)}{q(x)}$ are the same, in which case they must all equal 1. □

1.3.6 THEOREM. *It is always true that $I(X;Y) \geq 0$. We have equality if and only if X and Y are independent.*

Proof. $I(X;Y) = D(p(x,y)\|p(x)p(y))$, and $D(p(x,y)\|p(x)p(y)) = 0$ if and only if $p(x,y)$ and $p(x)p(y)$ are the same distribution, i.e. if X and Y are independent. □

1.4 *Entropy Rate*

1.4.1 definition. A **stochastic process** is simply a sequence of random variables.

1.4.2 definition. A **discrete-time stationary stochastic process** is a sequence of random variables $(X_1, X_2, X_3, \ldots)$ with the property that for every $m \geq 1$, $p(x_1, \ldots, x_m) = p(x_{n+1}, x_{n+2}, \ldots, x_{n+m})$ for all $n \geq 0$.

1.4.3 theorem (Entropy rate). *For any discrete-time stationary stochastic process,*

$$\lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n) = \lim_{n\to\infty} H(X_{n+1} \mid X_1, \ldots, X_n)$$

*and this limit exists. This quantity is called the **entropy rate** of the process.*

Lecture 5
September 10th, 2015

Proof. In fact,

(1) $\quad H(X_1) \geq H(X_2|X_1) \geq H(X_3|X_1, X_2) \geq \cdots$

This is because

$$H(X_{n+2}|X_1, \ldots, X_{n+1}) \leq H(X_{n+1}|X_1, \ldots, X_n) = H(X_{n+2}|X_2, \ldots, X_{n+1})$$

The equality holds because of stationarity, and the inequality holds because we are conditioning by more variables on the left. In fact,

$$H(X_{n+2}|X_1, \ldots, X_{n+1}) - H(X_{n+2}|X_2, \ldots, X_{n+1}) = I(X_{n+2}; X_1|X_2, \ldots, x_{n+1}) \geq 0$$

It follows from (1) that $\lim_{n\to\infty} H(X_{n+1}|X_1, \ldots, X_n)$ exists because it is a non-increasing sequence of nonnegative numbers. Call this limit A. For every $\varepsilon > 0$, there is some finite N such that for all $n \geq N$, $\frac{1}{n} H(X_1, \ldots, X_n) \leq A + \varepsilon$. This is because $H(X_1, \ldots, X_n) = H(X_1) + \cdots + H(X_n|x_1, \ldots, X_{n-1})$. Hence the proof is done. □

1.4.4 definition. Given a sequence of numbers $u_1, u_2, \ldots$, the sequence $w_1, w_2, \ldots$ where $w_n = \frac{1}{n}(u_1 + \cdots + u_n)$ is called the sequence of Cesaro means associated to the original sequence.

1.4.5 example. Suppose we have a sequence of i.i.d. random variables $\ldots, X_{-1}, X_0, X_1, X_2, \ldots$ where $p(x_1, \ldots, x_k) = p(x_1)p(x_2)\cdots p(x_k)$ for some $p(x)$ defined over $x \in \mathcal{X}$. Here $H(X_1, \ldots, X_n) = nH(X_1)$, so the entropy rate is $H(X_1)$.

1.4.6 example (Stationary Markov chains). Here we start with some transition probability matrix $[P_{ij}]$, $i, j \in \mathcal{X}$ where $P_{ij} \geq 0$ for all $i, j \in \mathcal{X}$ and $\sum_j P_{ij} = 1$ for all $i \in \mathcal{X}$ and given some probability distribution $(\pi_i, i \in \mathcal{X})$ satisfying $\sum_i \pi_i P_{ij} = \pi_j$ for all $j \in \mathcal{X}$, we can define a stationary stochastic process

$\ldots, X_{-1}, X_0, X_1, \ldots$ where

$$\begin{aligned} p(x_1, \ldots, x_n) &= \pi_{x_1} P_{x_1 x_2} P_{x_2 x_3} \cdots P_{x_{n-1} x_n} \\ &= p(x_1) p(x_2|x_1) p(x_3|x_2) \cdots p(x_n|x_{n-1}) \end{aligned}$$

More generally, a sequence of discrete random variables $\ldots, Y_{-1}, Y_0, Y_1, \ldots$ is called Markov if $p(y_{k+1}|y_m, \ldots, y_k) = p(y_{k+1}|p_k)$ holds for all k and all $m \leq k$.

What is $\lim_{n\to\infty} H(X_{n+1}|X_1, \ldots, X_n)$ for a stationary Markov process?

The key observation is that $H(X_{n+1}|X_1, \ldots, X_n) = H(X_{n+1}|X_n)$. This gives $H(X_3|X_1, X_2) = H(X_3|X_2) = H(X_2|X_1)$ and, more generally, $H(X_{n+1}|X_1, \ldots, X_n) = H(X_{n+1}|X_n) = H(X_2|X_1)$. Thus the above limit is $H(X_2|X_1)$, and this is the entropy rate of a stationary Markov process.

$$\begin{aligned} H(X_2|X_1) &= -\sum_{x_1, x_2} p(x_1, x_2) \log p(x_2|x_1) \\ &= -\sum_{i,j} \pi_i P_{ij} \log P_{ij} \end{aligned}$$

1.4.7 example (Memory-k Markov processes). This can be generalized to memory-k Markov processes for some $k \geq 2$ ($k = 1$ is the original Markov case). This means that you are given $[P_{j|i_1,\ldots,i_k}]$ where $P_{j|i_1,\ldots,i_k} \geq 0$ and $\sum_j P_{j|i_1,\ldots,i_k} = 1$ for all $(i_1, \ldots, i_k)$ and given $\pi_{i_1,\ldots,i_k} \geq 0$ with $\sum_{i_1,\ldots,i_k} \pi_{i_1,\ldots,i_k} = 1$ with $\sum_{i_1,\ldots,i_k} \pi_{i_1,\ldots,i_k} P_{i_{k+1}|i_1,\ldots,i_k} = \pi_{i_1,\ldots,i_k}$ for all $i_1, \ldots, i_k$ and $\ldots, X_{-1}, X_0, X_1, \ldots$ has $p(x_1, \ldots, x_l)$ derived from this. That is, for $l \geq k$,

$$p(x_1, \ldots, x_l) = \pi_{i_1,\ldots,x_k} P_{x_n|x_1,\ldots,x_k}$$

For $l < k$, it is simply derived from π.

Here, the entropy rate is $H(X_{k+1}|X_1, \ldots, X_k)$.

One can also see this by constructing an $\mathcal{X}^k$-valued stationary stochastic process $\ldots, Z_{-1}, Z_0, Z_1, \ldots$ from the original memory-k Markov process where $Z_n = (X_{n-k+1}, \ldots, X_n)$. This will be a Markov process (i.e. $k = 1$) and its entropy rate is the same as that of the original process. This is because it is a deterministic function of the original process and the original process is a deterministic function of this new process, so they must have the same entropy.

1.4.8 example (Renewal processes). Start with a probability distribution $(q_1, q_2, \ldots)$ on the positive integers, i.e. $q \geq 0$ for all $i = 1, 2, \ldots$ and $\sum_i q_i = 1$. Also assume $\sum_i iq_i = a$ is finite. Associated to this define its renewal distribution $r_l = \frac{1}{a} \sum_{j=l+1}^{\infty} q_j$ for $l = 0, 1, 2, \ldots$. Note that $\sum_{l=0}^{\infty} r_l = 1$.

We can define a $\{0, 1\}$-valued stationary stochastic process $\ldots, X_{-1}, X_0, X_1, \ldots$

with

$$P(\text{Most recent of the } X_i \text{ that is 1 being } X_{-l}) = r_l$$

and with

$$P(X_1 = 1 | \text{most recent of } X_i \text{ that is 1 is } X_{-l}) = \frac{q_{l+1}}{\sum_{j=l+1}^{\infty} q_j}$$