EE 229A

Venkat Anantharam

Davis Foote[*]

University of California, Berkeley

August 27th, 2015 – December 10th, 2015

Contents

1 Defining Information

Lecture 1
September 1st, 2015

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

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

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

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 $(p_1, \ldots, p_m)$.

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

[*] djfoote@berkeley.edu

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.

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

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 $\eta \in [0, 1]$

1.7 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.

1.8 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.

1.9 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 $H(p_1, \ldots, p_m)$ as a function from probability distributions on $\{1, \ldots, m\}$ to real numbers.

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

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

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

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

$H(p_1, \ldots, p_m)$ 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 $\mathbf{p}^{(\mathbf{0})} = (p_1^{(0)}, \ldots, p_m^{(0)})$ and $\mathbf{p}^{(\mathbf{1})} = (p_1^{(1)}, \ldots, p_m^{(1)})$ and $\eta \in [0, 1]$,

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

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

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

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 $p(x, y)$.

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

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.

1.18 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 $p(x|y)$).

1.19 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 $H(X|Y=y)$.

1.20 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.

2 Toward a Calculus of Entropy

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.

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.

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.

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. $\square$

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

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

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