Lesson 11: Entropy

Theme: Review of basic probability theory and entropy.

1 A review on basic probability theory

A *probability space* is a system $(\Omega, \mathbf{Pr})$, where Ω is a set called *sample space*, and $\mathbf{Pr} : 2^\Omega \to \mathbb{R}$ is a probability function satisfying the following conditions.

- $\mathbf{Pr}(\Omega) = 1$,
- $0 \leqslant \mathbf{Pr}(E) \leqslant 1$, for every $E \in 2^\Omega$,
- for any countably infinite sequence of pairwise disjoint sets $E_1, E_2, \ldots$, $\mathbf{Pr}(\bigcup_{i \geqslant 1} E_i) = \sum_{i \geqslant 1} \mathbf{Pr}(E_i)$.

The sets in 2^Ω are usually called *events*, and the singletons $\{e\}$ *elementary events*. To avoid too many bracketing, we will write $\mathbf{Pr}[E]$, instead of $\mathbf{Pr}(E)$, and $\mathbf{Pr}[e]$, instead of $\mathbf{Pr}[\{e\}]$. We will only deal with discrete probability space, i.e., when Ω is a countable set. Without loss of generality, we can assume that $\mathbf{Pr}[e] > 0$, for every $e \in \Omega$.

We say that *two events E and F are independent*, if $\mathbf{Pr}[X \cap Y] = \mathbf{Pr}[X] \cdot \mathbf{Pr}[Y]$. Likewise, *a collection of events $E_1, \ldots, E_k$ are independent*, if for every $I \subseteq \{1, \ldots, k\}$,

$$\mathbf{Pr}\Big[\bigcap_{i \in I} E_i\Big] \;=\; \prod_{i \in I} \mathbf{Pr}[E_i]$$

The *conditional probability that event E occurs given that event F occurs* is defined as:

$$\mathbf{Pr}[E \mid F] \;:=\; \frac{\mathbf{Pr}[E \cap F]}{\mathbf{Pr}[F]}$$

A (discrete) *random variable* is a function $X : \Omega \to \mathbb{R}$. The *probability of the event "$X = a$"* is defined as the probability of the event $\{e \in \Omega \mid X(e) = a\}$, i.e.,

$$\mathbf{Pr}[X = a] \;:=\; \sum_{e \in \Omega \text{ such that } X(e) = a} \mathbf{Pr}[e]$$

The probabilities $\mathbf{Pr}[X \circledast a]$, where $\circledast \in \{\leqslant, \geqslant, <, >, \neq\}$ can be defined in a similar manner. We say that *a random variable X is uniformly distributed on $\mathsf{range}(X)$*, if $\mathbf{Pr}[X = a] = \mathbf{Pr}[X = b]$, for every $a, b \in \mathsf{range}(X)$.

We say that *two random variables X, Y are independent*, if $\mathbf{Pr}[X = x \cap Y = y] = \mathbf{Pr}[X = x] \cdot \mathbf{Pr}[Y = y]$, for every possible values x and y. Likewise, *a collection of random variables $X_1, \ldots, X_k$ are independent*, if for every $I \subseteq \{1, \ldots, k\}$, for every $i \in I$, for every value x_i,

$$\mathbf{Pr}\Big[\bigcap_{i \in I} X_i = x_i\Big] \;=\; \prod_{i \in I} \mathbf{Pr}[X_i = x_i]$$

The *expectation* of a random variable X is defined as $\mathbf{E}[X] := \sum_i \; i \cdot \mathbf{Pr}[X = i]$. It is known that for every two random variables X and Y, and for every constant c,

- $\mathbf{E}[X + Y] = \mathbf{E}[X] + \mathbf{E}[Y]$,
- $\mathbf{E}[cX] = c\mathbf{E}[X]$.

A pair (X, Y) of random variables can be viewed as a random variable Z with an appropriate "pairing function" $\langle \cdot, \cdot \rangle$, where $\mathbf{Pr}[Z = c] = \mathbf{Pr}[(X = a, Y = b)]$, where $\langle a, b \rangle = c$. For convenience, we will simply write (X, Y) to denote a random variable Z obtained in this way.

2 Entropy

Here the logarithm is always of base 2, and the sample space is a finite set. The *entropy* $H(X)$ *of a random variable* X is defined as:

$$H(X) \;:=\; -\sum_{x \,\in\, \mathsf{range}(X)} \mathbf{Pr}[X = x] \cdot \log(\mathbf{Pr}[X = x])$$

Proposition 11.1 *For every random variable* X, $H(X) \leqslant \log|\mathsf{range}(X)|$. *Equality holds when* X *is uniformly distributed on* $\mathsf{range}(X)$.

Proof. We will show that $\sum_{i=1}^{m} p_i \log(1/p_i) \leqslant \log(m)$, whenever $\sum_{i=1}^{m} p_i = 1$. The proof is by induction on m. The base case $m = 2$ is straightforward. The entropy $H(X)$ can be viewed as a function on p:

$$H(X) \;=\; p\log(1/p) \;+\; (1-p)\log(1/(1-p))$$

Taking the derivative of $H(X)$ on p, we obtain that $H(X)$ is maximal when $p = 1/2$. So, $H(X) \leqslant \log 2$.

The induction step is as follows. Without loss of generality, we assume m is even. If m is odd, we can "split" one term $p\log(1/p)$ into two $p/2\log(2/p) + p/2\log(2/p)$, and the inequality is not effected since $p\log(1/p) \leqslant p\log(2/p)$. Let $\lambda = \sum_{i=1}^{m/2} p_i$. Then,

$$\begin{aligned}
\sum_{i=1}^{m} p_i \log(1/p_i) \;&=\; \sum_{i=1}^{m/2} p_i \log(1/p_i) \;+\; \sum_{i=m/2+1}^{m} p_i \log(1/p_i) \\
&=\; \lambda \sum_{i=1}^{m/2} \frac{p_i}{\lambda} \log(\frac{\lambda}{p_i} \cdot \frac{1}{\lambda}) \;+\; (1-\lambda) \sum_{i=m/2+1}^{m/2} \frac{p_i}{1-\lambda} \log(\frac{1-\lambda}{p_i} \cdot \frac{1}{1-\lambda}) \\
&\leqslant\; \lambda \sum_{i=1}^{m/2} \frac{p_i}{\lambda} \log(\frac{\lambda}{p_i}) \;+\; \lambda \sum_{i=1}^{m/2} \frac{p_i}{\lambda} \log(\frac{1}{\lambda}) \\
&\quad +\; (1-\lambda) \sum_{i=m/2+1}^{m} \frac{p_i}{1-\lambda} \log(\frac{1-\lambda}{p_i}) \;+\; (1-\lambda) \sum_{i=m/2+1}^{m} \frac{p_i}{1-\lambda} \log(\frac{1}{1-\lambda}) \\
&=\; \lambda \sum_{i=1}^{m/2} \frac{p_i}{\lambda} \log(\frac{\lambda}{p_i}) \;+\; \lambda \log(\frac{1}{\lambda}) \\
&\quad +\; (1-\lambda) \sum_{i=m/2+1}^{m} \frac{p_i}{1-\lambda} \log(\frac{1-\lambda}{p_i}) \;+\; (1-\lambda)\log(\frac{1}{1-\lambda})
\end{aligned}$$

Applying induction hypothesis on the first and third terms,

$$\begin{aligned}
\sum_{i=1}^{m} p_i \log(1/p_i) \;&\leqslant\; \lambda\log(m/2) \;+\; \lambda\log(\frac{1}{\lambda}) \;+\; (1-\lambda)\log(m/2) \;+\; (1-\lambda)\log(\frac{1}{1-\lambda}) \\
&\leqslant\; \log(m/2) \;+\; \lambda\log(\frac{1}{\lambda}) \;+\; (1-\lambda)\log(\frac{1}{1-\lambda})
\end{aligned}$$

Applying the base case on the second and third terms,

$$\sum_{i=1}^{m} p_i \log(1/p_i) \;\leqslant\; \log(m/2) + \log 2 \;= \log m$$

■

For two random variables X and Y, the *conditional entropy* $H(Y \mid X)$ is defined as:

$$H(Y \mid X) \quad := \quad -\sum_{x,y} \mathbf{Pr}[Y = y \cap X = x] \cdot \log(\mathbf{Pr}[Y = y \mid X = x])$$

Proposition 11.2 *If X and Y are independent random variables, then $H(X \mid Y) = H(X)$.*

Proposition 11.3 *For every random variables X, Y, Z, the following holds.*[1]

- $H(X, Y) = H(X) + H(Y \mid X)$.
- $H(X, Y \mid Z) = H(X \mid Z) + H(Y \mid Z, X)$.

Corollary 11.4 *In general, for random variables $X_1, \ldots, X_k$,*

$$H(X_1, \ldots, X_k) \quad = \quad H(X_1) \; + \; H(X_2 \mid X_1) \; + \; \cdots \; + \; H(X_k \mid X_1, \ldots, X_{k-1}).$$

Moreover, if $X_1, \ldots, X_k$ are independent, then $H(X_1, \ldots, X_k) = H(X_1) + \cdots + H(X_k)$.

Proposition 11.5 *Let $X_1, \ldots, X_k, Y$ be random variables. Then, $H(Y \mid X_1, \ldots, X_k) \leqslant H(Y \mid X_1, \ldots, X_{k-1})$. More generally, for every $I \subseteq J \subseteq \{1, \ldots, k\}$, $H(Y \mid X_J) \leqslant H(Y \mid X_I)$, where $X_{\{i_1,\ldots,i_m\}}$ denotes $(X_{i_1}, \ldots, X_{i_m})$.*

Proof. We will show that $H(X, Y) \leqslant H(X) + H(Y)$, which combines with the fact that $H(X, Y) = H(X) + H(Y \mid X)$, will imply $H(Y \mid X) \leqslant H(Y)$. Proposition 11.5 follows immediately by straightforward induction.

We need to use following claim.

Claim 1 log *is concave, i.e., for every real numbers x, y, for every $0 \leqslant \lambda \leqslant 1$,*

$$\lambda \log x \; + \; (1 - \lambda) \log y \quad \leqslant \quad \log \Big(\lambda x \; + \; (1 - \lambda) y \Big)$$

In general, for every $x_1, \ldots, x_m$ and every $\lambda_1, \ldots, \lambda_m$, where $\lambda_1 + \cdots + \lambda_m = 1$,

$$\lambda_1 \log x_1 \; + \; \cdots \; + \; \lambda_m \log x_m \quad \leqslant \quad \log \Big(\lambda_1 x_1 \; + \; \cdots \; + \; \lambda_m x_m \Big)$$

What this claim states is that if we draw a line segment between two points $(x, \log x)$ and $(y, \log y)$ in 2D, the line segment is always "below" the graph of the log function.

[1] Note that if we want to be technically correct, we should have written $H((X, Y))$, $H((X, Y) \mid Z)$, and $H(Y \mid (Z, X))$ in this proposition. However, to avoid cumbersome brackets, we omit them.

Now, back to calculating $H(Y, X)$,

$$\begin{aligned}
H(Y,X) &= \sum_{x,y} \mathbf{Pr}[Y=y \cap X=x] \cdot \log \frac{1}{\mathbf{Pr}[Y=y \cap X=x]} \\
&= \sum_{x,y} \mathbf{Pr}[Y=y \cap X=x] \cdot \log \frac{\mathbf{Pr}[X=x]\mathbf{Pr}[Y=y]}{\mathbf{Pr}[Y=y \cap X=x]\mathbf{Pr}[X=x]\mathbf{Pr}[Y=y]} \\
&= \sum_{x,y} \mathbf{Pr}[Y=y \cap X=x] \cdot \log \frac{\mathbf{Pr}[X=x]\mathbf{Pr}[Y=y]}{\mathbf{Pr}[Y=y \cap X=x]} \\
&\quad + \sum_{x,y} \mathbf{Pr}[Y=y \cap X=x] \cdot \log \frac{1}{\mathbf{Pr}[X=x]} \\
&\quad + \sum_{x,y} \mathbf{Pr}[Y=y \cap X=x] \cdot \log \frac{1}{\mathbf{Pr}[Y=y]} \\
&= \sum_{x,y} \mathbf{Pr}[Y=y \cap X=x] \cdot \log \frac{\mathbf{Pr}[X=x]\mathbf{Pr}[Y=y]}{\mathbf{Pr}[Y=y \cap X=x]} \quad + \quad H(X) \quad + \quad H(Y)
\end{aligned}$$

Using the fact that log is concave,

$$\begin{aligned}
H(Y,X) &\leqslant \log \sum_{x,y} \mathbf{Pr}[Y=y \cap X=x] \frac{\mathbf{Pr}[X=x]\mathbf{Pr}[Y=y]}{\mathbf{Pr}[Y=y \cap X=x]} \quad + \quad H(X) \quad + \quad H(Y) \\
&= \log \sum_{x,y} \mathbf{Pr}[Y=y]\mathbf{Pr}[X=x] \quad + \quad H(X) \quad + \quad H(Y) \\
&= \log 1 \quad + \quad H(X) \quad + \quad H(Y) \\
H(Y,X) &\leqslant H(X) + H(Y)
\end{aligned}$$

This completes the proof of Proposition 11.5. ■

3 Counting with entropy

The material in this section is taken from [2].

Proposition 11.6 *Let $\{p_1, \ldots, p_n\}$ be a set of points in 3D with n_1 distinct projections on the (x, y) coordinates; n_2 distinct projections on the (y, z) coordinates; n_3 distinct projections on the (x, z) coordinates. Then, $n^2 \leqslant n_1 n_2 n_3$.*

Proof. We pick randomly one of the n points with uniform distribution. Let $P = (X, Y, Z)$ be its random variable. Define the following random variables.

$$P_1 := (X, Y) \qquad P_2 := (X, Z) \qquad P_3 := (Y, Z)$$

By Corollary 11.4, we have:

$$\begin{aligned}
H(P) &= H(X) + H(Y \mid X) + H(Z \mid X, Y) \\
H(P_1) &= H(X) + H(Y \mid X) \\
H(P_2) &= H(X) + H(Z \mid X) \\
H(P_3) &= H(Y) + H(Z \mid Y)
\end{aligned}$$

Then,

$$2H(P) \leqslant 2H(X) + 2H(Y \mid X) + 2H(Z \mid X, Y)$$

By Proposition 11.5, $H(Y \mid X) \leqslant H(Y)$ and $H(Z \mid X, Y) \leqslant H(Z \mid X),\ H(Z \mid Y)$. So,

$$2H(P) \leqslant H(P_1) + H(P_2) + H(P_3).$$

By Proposition 11.1, $H(P) = \log n$, and $H(P_1) \leqslant \log n_1$, $H(P_2) \leqslant \log n_2$ and $H(P_3) \leqslant \log n_3$. This completes the proof of proposition. ■

Theorem 11.7 (Shearer's entropy lemma [1]) *Let $X_1, \ldots, X_n$ be random variables and $X = (X_1, \ldots, X_n)$. Let $A_1, \ldots, A_m \subseteq \{1, \ldots, n\}$ such that every element $i \in \{1, \ldots, n\}$ appears in at least k number of A_i's. Then,*

$$\sum_{i=1}^{m} H(X_{A_i}) \geqslant k \cdot H(X)$$

Here, for a set $B = \{i_1, \ldots, i_l\}$, the random variable X_B denotes $(X_{i_1}, \ldots, X_{i_l})$.

Proof. By Corollary 11.4,

$$\begin{aligned} H(X) &= H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_k \mid X_1, \ldots, X_{k-1}). \\ H(X_{A_j}) &= H(X_{i_1}) + H(X_{i_2} \mid X_{i_1}) + \cdots + H(X_{i_{l_j}} \mid X_{i_1}, \ldots, X_{i_{l_j-1}}). \end{aligned}$$

where $A_j = \{i_1, \ldots, i_{l_j}\}$ and $i_1 < \cdots < i_{l_j}$.

Consider the sum:

$$\sum_{i=1}^{m} H(X_{A_j}) = \sum_{i=1}^{m} H(X_{i_1}) + H(X_{i_2} \mid X_{i_1}) + \cdots + H(X_{i_{l_j}} \mid X_{i_1}, \ldots, X_{i_{l_j-1}}).$$

Since every element $i \in \{1, \ldots, n\}$ appears in at least k number of A_j's, there are at least k terms of the form $H(X_i \mid X_{A_j \cap \{1,\ldots,i-1\}})$ on the right side.

By Proposition 11.5,

$$H(X_i \mid X_{A_j \cap \{1,\ldots,i-1\}}) \geqslant H(X_i \mid X_1, \ldots, X_{i-1})$$

Thus,

$$\sum_{i=1}^{m} H(X_{A_j}) \geqslant kH(X)$$

This completes our proof. ■

Corollary 11.8 *Let $\mathcal{H} = (V, \mathcal{E})$ be a hypergraph, and let $\{A_1, \ldots, A_m\}$ be a collection of subsets of V. Suppose every element in V appears in at least k number of A_i's. Let $\mathcal{E}_i = \{e \cap A_i \mid e \in \mathcal{E}\}$. Then,*

$$|\mathcal{E}|^k \leqslant \prod_{i=1}^{m} |\mathcal{E}_i|$$

Proof. Assume that $V = \{1, \ldots, n\}$. We pick randomly a set $e \in \mathcal{E}$ with uniform distribution. Let $Z = (X_1, \ldots, X_n)$ be the characteristic random variable for e, i.e., each X_i is the indicator random variable for $i \in e$. Thus, $H(Z) = \log |\mathcal{E}|$.

For each A_j, let Z_j be the projection of Z to components in A_j. Then, $H(Z_j) \leqslant \log |\mathcal{E}_j|$. By Theorem 11.7, $kH(Z) \leqslant \sum_{j=1}^{m} H(Z_j)$. Hence,

$$k \log |\mathcal{E}| \leqslant \sum_{j=1}^{m} \log |\mathcal{E}_j|$$

This completes the proof of corollary. ■

References

[1] F. Chung, R. Graham, P. Frankl, and J. Shearer. Some intersection theorems for ordered sets and graphs. *Journal of Cominatorial Theory, Ser. A*, 43(1):23–37, 1986.

[2] J. Radhakrishnan. Entropy and counting. In J. C. Mishra, editor, *Computational mathematics, modelling and algorithms*, 2001. URL: `http://www.tcs.tifr.res.in/~jaikumar/Papers/EntropyAndCounting.pdf`.