1 Overview of course

2 Some definitions

An ensemble 'X' is a random variable x with a set of possible *outcomes*, $\mathcal{A}_X=\{a_1, a_2, \ldots a_i, \ldots a_I\}$, having probabilities $\{p_1, p_2, \ldots p_I\}$, with $P(x{=}a_i) = p_i$, $p_i \geq 0$ and $\sum_{x\in\mathcal{A}_X} P(x) = 1$.

A joint ensemble 'XY' is an ensemble in which each outcome is an ordered pair x, y with $x \in \mathcal{A}_X=\{a_1, \ldots, a_I\}$ and $y \in \mathcal{A}_Y=\{b_1, \ldots, b_J\}$. From the joint probability $P(x, y)$ we can obtain the following:

Marginal probability

$$P(x{=}a_i) \equiv \sum_{y\in\mathcal{A}_Y} P(x{=}a_i, y);\ P(y) \equiv \sum_{x\in\mathcal{A}_X} P(x, y)$$

Conditional probability

$$P(x{=}a_i|y{=}b_j) \equiv \frac{P(x{=}a_i, y{=}b_j)}{P(y{=}b_j)} \text{ if } P(y{=}b_j) \neq 0$$

Rather than writing down the joint probability directly, we will often define an ensemble in terms of a collection of conditional probabilities. The following rules of probability theory will be useful. ($\mathcal{H}$ denotes background assumptions.)

Product rule

$$P(x, y|\mathcal{H}) = P(x|y, \mathcal{H})P(y|\mathcal{H})$$

Sum rule

$$\begin{aligned} P(x|\mathcal{H}) &= \sum_y P(x, y|\mathcal{H}) \\ &= \sum_y P(x|y, \mathcal{H})P(y|\mathcal{H}) \end{aligned}$$

Bayes' theorem —obtained from the product rule:

$$\begin{aligned} P(y|x, \mathcal{H}) &= \frac{P(x|y, \mathcal{H})P(y|\mathcal{H})}{P(x|\mathcal{H})} \\ &= \frac{P(x|y, \mathcal{H})P(y|\mathcal{H})}{\sum_{y'} P(x|y', \mathcal{H})P(y'|\mathcal{H})} \end{aligned}$$

Definition of Entropy and related functions

The entropy of X is defined by:

$$H(X) \equiv \sum_{x\in\mathcal{A}_X} P(x) \log \frac{1}{P(x)},$$

with the convention for $P(x){=}0$ that $0 \times \log 1/0 \equiv 0$, since $\lim_{\theta\to 0^+} \theta \log 1/\theta{=}0$. The entropy measures the information content or 'uncertainty' of x.

$H(X) \geq 0$ with equality iff $p_i{=}1$ for one i.

$H(X) \leq \log(|X|)$ with equality iff $p_i{=}1/|X|$ for all i. ($|X|$ denotes the number of elements in the set $\mathcal{A}_X$.)

The joint entropy of X, Y is then:

$$H(X, Y) = \sum_{x,y\in\mathcal{A}_X,\mathcal{A}_Y} P(x, y) \log \frac{1}{P(x, y)}.$$

Entropy is additive for independent random variables:

$$H(X, Y) = H(X) + H(Y) \text{ iff } P(x, y) = P(x)P(y).$$

The conditional entropy of X **given** $y = b_k$ is the entropy of the probability distribution $P(x|y{=}b_k)$.

$$H(X|y{=}b_k) \equiv \sum_{x\in\mathcal{A}_X} P(x|y{=}b_k) \log \frac{1}{P(x|y{=}b_k)}.$$

The conditional entropy of X **given** Y is the average of the above, summing over all values of y, weighted by their probability.

$$\begin{aligned} H(X|Y) &\equiv \sum_{y\in\mathcal{A}_Y} P(y) \left[\sum_{x\in\mathcal{A}_X} P(x|y) \log \frac{1}{P(x|y)}\right] \\ &= \sum_{x,y\in\mathcal{A}_X,\mathcal{A}_Y} P(x, y) \log \frac{1}{P(x|y)}. \end{aligned}$$

This measures the average uncertainty that remains about x when y is known.[1] So data is helpful—it doesn't increase uncertainty, on average!

Chain rule for Entropy The joint entropy, conditional entropy and marginal entropy are related by:[2]

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

The mutual information between X **and** Y is

$$H(X; Y) \equiv H(X) - H(X|Y),$$

and satisfies $H(X; Y) = H(Y; X)$,[3] and $H(X; Y) \geq 0$. It measures the average reduction in uncertainty about x that results from learning the value of y, **or vice versa**. Equivalently, it measures the amount of information that x conveys about y.

The distance between two random variables is the difference between their joint entropy and their mutual information:

$$D(X, Y) \equiv H(X, Y) - H(X; Y).$$

This quantity satisfies the axioms for a distance—$D(X, Y) \geq 0$, $D(X, X) = 0$, $D(X, Y){=}D(Y, X)$, and $D(X, Z) \leq D(X, Y) + D(Y, Z)$.[4]

[1] Exercise: confirm that it is possible for $H(X|y = b_k)$ to exceed $H(X)$, but that the average, $H(X|Y)$ is less than $H(X)$.

[2] Exercise: confirm this.

[3] Exercise: confirm this.

[4] Optional exercise: prove these relations. Incidentally, we are unlikely to see $D(X, Y)$ again in these lectures.

OTHER USEFUL DEFINITIONS

The relative entropy *or* **Kullback-Leibler distance** between two probability distributions $p(x)$ and $q(x)$ that are defined over the same $x \in \mathcal{A}_X$ is

$$D_{\mathrm{KL}}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}.$$

The relative entropy satisfies $D_{\mathrm{KL}}(p||q) \geq 0$ (Gibbs' inequality) with equality only if $p{=}q$.[5] Note that in general $D_{\mathrm{KL}}(p||q) \neq D_{\mathrm{KL}}(q||p)$, so this is not strictly a 'distance'. It is also known as the 'discrimination'.

Convex function. A function f(x) is *convex* over (a, b) if for all $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2).$$

A function f is strictly convex if the equality holds only for $\lambda{=}0$ and $\lambda{=}1$. Some strictly convex functions are x^2, e^x, $\log(1/x)$ for $x > 0$, and $x \log x$ for $x > 0$.

Jensen's inequality. If f is a convex function and x is a random variable then:

$$E\left[f(x)\right] \geq f\left(E[x]\right),$$

where E denotes expectation. If f is strictly convex and $E\left[f(x)\right] {=} f\left(E[x]\right)$, then x is a constant (with probability 1).

References

BLAHUT, R. E. (1987) *Principles and Practice of Information Theory*. New York: Addison-Wesley.

COVER, T. M., and THOMAS, J. A. (1991) *Elements of Information Theory*. New York: Wiley.

MCELIECE, R. J. (1977) *The theory of information and coding: a mathematical framework for communication*. Reading, Mass.: Addison-Wesley. Reprinted (1984) by C.U.P. in *Encyclopedia of Mathematics.*

Coding theory only:

WITTEN, I. H., NEAL, R. M., and CLEARY, J. G. (1987) Arithmetic coding for data compression. *Communications of the ACM* **30** (6): 520–540.

LIN, S., and COSTELLO, JR., D. J. (1983) *Error control coding: fundamentals and applications*. Englewood Cliffs, N.J.: Prentice-Hall.

BERLEKAMP, E. R. (1968) *Algebraic Coding Theory*. New York: McGraw-Hill.

MACWILLIAMS, F. J., and SLOANE, N. J. A. (1977) *The theory of error-correcting codes*. Amsterdam: North-Holland.

PETERSON, W. W., and WELDON, JR., E. J. (1972) *Error-Correcting Codes*. Cambridge, Massachusetts: MIT Press, 2nd edition.

RAO, T. R. N., and FUJIWARA, E. (1989) *Error-control Coding for Computer Systems*. Prentice-Hall.

[5] Exercise: prove this. You may find this helps with the previous two exercises. Note $H(X;Y) = D_{\mathrm{KL}}(P(x,y)||P(x)P(y))$.