Fundamentals of Mathematical Informatics

Random Variables, Entropy, and Information

Francesco Buscemi

Lecture One

The Information Theory

Great scientific theories, like great symphonies and great novels, are among man's proudest—and rarest—creations. What sets the scientific theory apart from and, in a sense, above the other creations is that it may profoundly and rapidly alter man's view of his world. Within the last five years a new theory has appeared that seems to bear some of the same hallmarks of greatness. It may be no exaggeration to say that man's progress in peace, and security in war, depend more on fruitful applications of information theory than on physical demonstrations, either in bombs or power plants, that Einstein's famous equation works."

from "The Information Theory," Fortune, pp. 136-158, Dec. 1953,
five years after the publication of Shannon's seminal paper (1948).

Example: predicting a roll of a die

- Let us consider a six-face die $\{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$.
- The die is called *fair* if $\Pr\{$⚀$\} = \Pr\{$⚁$\} = \cdots = \Pr\{$⚅$\} = \frac{1}{6}$.
- Imagine now a *biased* die, made so that $\Pr\{$⚀$\} = 0.95$ and $\Pr\{$⚁$\} = \cdots = \Pr\{$⚅$\} = 0.01$.
- **Question.** Which die is 'more uncertain,' $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6},)$ or $(0.95, 0.01, 0.01, 0.01, 0.01, 0.01)$?
- Imagine now two other biased dice with different biases, for example, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12})$ and $(\frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{12}, \frac{1}{12}, \frac{1}{20})$.
- **Question.** Which die is 'more uncertain' now?

Random variables and their entropy

- A **random variable** (RV) X is like a 'device' that outputs an element of a set $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ (called the **range** of X) with probability $\Pr\{X = x_i\} \overset{\text{def}}{=} p_i$.
- The **entropy** of a RV X with probability distribution $(p_1, \cdots, p_n)$ is defined as

$$H(X) \overset{\text{def}}{=} H(p_1, \cdots, p_n) \overset{\text{def}}{=} -\sum_{i=1}^{n} p_i \log_2 p_i .$$

- The unit of entropy is called 'bit.'
- **Remark:** the entropy depends only on the probability distribution, not on the range.
- When computing the entropy, we use the convention $0 \log_2 0 = 0$: events that never happen do not contribute to the entropy, namely, $H(p_1, \cdots, p_n, 0) = H(p_1, \cdots, p_n)$.
- For the dice of the **previous example**, $H(\frac{1}{3}, \frac{1}{3}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}) \approx 2.25 < H(\frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{12}, \frac{1}{12}, \frac{1}{20}) \approx 2.31$.
- Next week, we will see why (and in which sense) this fact tells us that the die $(\frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{12}, \frac{1}{12}, \frac{1}{20})$ is 'more uncertain' than $(\frac{1}{3}, \frac{1}{3}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12})$.

Example: a lottery

- Consider an **urn with four balls** $\left\{\textcircled{A}, \textcircled{A}, \textcircled{B}, \textcircled{C}\right\}$. A part from the labels, the balls are identical.
- Imagine to draw one ball at random and read the label: how to model the associated random variable?
- There are four identical balls: each one is picked with probability $1/4$.
- Two balls are marked with the letter A: the probability of getting the letter A is therefore $1/4 + 1/4 = 1/2$.
- The probability of getting B or C is $1/4$ in both cases.
- Therefore, this situation is modeled by a random variable X with **three possible outcomes**, $\mathcal{X} = \{x_1 \equiv \text{'A'}, x_2 \equiv \text{'B'}, x_3 \equiv \text{'C'}\}$, and $p_1 = 1/2$, $p_2 = p_3 = 1/4$.
- The **associated entropy** is equal to: $H(X) = H(\frac{1}{2}, \frac{1}{4}, \frac{1}{4}) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{4} = 1.5$ bits.

Example: a horse race

- Imagine a horse race with **eight horses**: Star, Dakota, Cheyenne, Spirit, Misty, Cowboy, Blaze, and Lucky.
- Imagine that each one can win the race with probabilities:
  $\Pr\{\text{'Star wins'}\} = \frac{1}{2}$,
  $\Pr\{\text{'Dakota wins'}\} = \frac{1}{4}$,
  $\Pr\{\text{'Cheyenne wins'}\} = \frac{1}{8}$,
  $\Pr\{\text{'Spirit wins'}\} = \frac{1}{16}$,
  $\Pr\{\text{'Misty wins'}\} = \Pr\{\text{'Cowboy wins'}\} = \Pr\{\text{'Blaze wins'}\} = \Pr\{\text{'Lucky wins'}\} = \frac{1}{64}$.
- The result of this race is modeled by a RV X with **eight possible outcomes** $\{x_1 = \text{'Star'}, x_2 = \text{'Dakota'}, x_3 = \text{'Cheyenne'}, x_4 = \text{'Spirit'}, x_5 = \text{'Misty'}, x_6 = \text{'Cowboy'}, x_7 = \text{'Blaze'}, x_8 = \text{'Lucky'}\}$ and probability distribution $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{4}$, $p_3 = \frac{1}{8}$, $p_4 = \frac{1}{16}$, $p_5 = p_6 = p_7 = p_8 = \frac{1}{64}$.
- The **associate entropy** is equal to $H(X) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{4}\log_2\frac{1}{4} - \frac{1}{8}\log_2\frac{1}{8} - \frac{1}{16}\log_2\frac{1}{16} - \frac{1}{16}\log_2\frac{1}{64} = 2$ bits.

Properties of the entropy function

- **Positivity.** $H(p_1, \cdots, p_n) \geq 0$, and $H(p_1, \cdots, p_n) = 0$ iff $\exists \bar{k} : p_i = \begin{cases} 1, & \text{if } i = \bar{k} \\ 0, & \text{if } i \neq \bar{k}. \end{cases}$ (In Kronecker's notation, $p_i = \delta_{i,\bar{k}}$).
- **Symmetry.** For any permutation π of $\{1, 2, \cdots, n\}$, $H(p_1, \cdots, p_n) = H(p_{\pi(1)}, \cdots, p_{\pi(n)})$.
- **Key Lemma.** Given a probability distribution $(p_1, \cdots, p_n)$, $H(p_1, \cdots, p_n) \leq -\sum_i p_i \log_2 q_i$, for all probability distributions $(q_1, \cdots, q_n)$, with equality iff $q_1 = p_1, \cdots, q_n = p_n$.

  Proof. Since $\log_e x \leq x - 1$, we have $\log_e(q_k/p_k) \leq (q_k/p_k) - 1$, so that $\sum_i p_i \log_e(q_i/p_i) \leq \sum_i q_i - \sum_i p_i = 0$. But then, the identity $\log_2 x = (\log_e x)/(\log_e 2)$ implies $-\sum_i p_i \log_2 p_i \overset{\text{def}}{=} H(p_1, \cdots, p_n) \leq -\sum_i p_i \log_2 q_i$. □
- **Theorem.** $H(p_1, \cdots, p_n) \leq \log_2 n$, with equality iff $p_1 = p_2 = \cdots = p_n = 1/n$.

  Proof. Apply the Key Lemma to the case $q_1 = q_2 = \cdots = q_n = 1/n$. □
- The probability distribution $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$ is called **uniform distribution**.

Dependent RVs

- Consider a pair of RVs (X, Y) with **joint probability distribution** $\Pr\{X = x_i \text{ and } Y = y_j\} \overset{\text{def}}{=} t_{ij}$.
- The **marginal distributions** are defined as:
  $r_i \overset{\text{def}}{=} \Pr\{X = x_i \text{ independently of the value of } Y\} = \sum_j t_{ij}$ and
  $s_j \overset{\text{def}}{=} \Pr\{Y = y_j \text{ independently of the value of } X\} = \sum_i t_{ij}$.
- X and Y are called **independent** iff $t_{ij} = r_i s_j$ for all i, j.
- **Theorem.** $H(X, Y) \leq H(X) + H(Y)$, with equality iff X and Y are independent.

  Proof. First, $H(X) + H(Y) = -\sum_i r_i \log_2 r_i - \sum_j s_j \log_2 s_j = -\sum_i (\sum_j t_{ij}) \log_2 r_i - \sum_j (\sum_i t_{ij}) \log_2 s_j = -\sum_{ij} t_{ij} (\log_2 r_i + \log_2 s_j) = -\sum_{ij} t_{ij} \log_2(r_i s_j)$. By the Key Lemma, $-\sum_{ij} t_{ij} \log_2(r_i s_j) \geq -\sum_{ij} t_{ij} \log_2 t_{ij}$, namely, $H(X) + H(Y) \geq H(X, Y)$. Moreover, the Key Lemma states that equality holds iff $t_{ij} = r_i s_j$ for all i, j, namely, iff X and Y are independent. □
- The difference $H(X) + H(Y) - H(X, Y) \geq 0$ can hence be used to measure 'how dependent' two RV are. (See also slide 14, 'Information.')

Example: shapes and colors

- Take the set $\{{\color{red}\square}, {\color{red}\bigcirc}, {\color{blue}\square}, {\color{blue}\bigcirc}\}$ which is: $\{\text{red}, \text{blue}\} \times \{\square, \bigcirc\}$.
- Two RVs, X for the color and Y for the shape.
- For example, $\Pr\{\text{red circle}\} = \Pr\{\text{‘}X = \text{red’ and ‘}Y = \bigcirc\text{’}\}$.
- **Question:** what is the marginal probability for X (the color)?
- For example,

$$\begin{aligned}\Pr\{X = \text{red}\} &= \Pr\{\text{red square}\} + \Pr\{\text{red circle}\}\\ &= \Pr\{\text{‘}X = \text{red’ and ‘}Y = \square\text{’}\} + \Pr\{\text{‘}X = \text{red’ and ‘}Y = \bigcirc\text{’}\}\\ &= \sum_{y\in\{\square,\bigcirc\}} \Pr\{\text{‘}X = \text{red’ and ‘}Y = y\text{’}\}.\end{aligned}$$

- **Question.** Let the probability distribution for $\{{\color{red}\square}, {\color{red}\bigcirc}, {\color{blue}\square}, {\color{blue}\bigcirc}\}$ be $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Are shape and color independent? (Yes.)
- **Question.** What about $(0.49, 0.01, 0.01, 0.49)$? (No.)
- **Question.** And what about $(\frac{3}{15}, \frac{2}{15}, \frac{6}{15}, \frac{4}{15})$? (Yes.)

Example: red shirts

- Consider a population of 100 people: 60 women and 40 men. Suppose that, in such a population, 10 men and 30 women are wearing a red shirt. Imagine to pick one person at random.
- The probability of picking one woman is $\Pr\{\text{‘woman’}\} = 60/100 = 0.6$. (Of course, $\Pr\{\text{‘man’}\} = 0.4$.)
- The probability of picking one person wearing a red shirt is $\Pr\{\text{‘red’}\} = (10 + 30)/100 = 0.4$. (Of course, $\Pr\{\text{‘other color’}\} = 0.6$.)
- If the chosen person is a woman, the probability that she is wearing a red shirt is $\Pr\{\text{‘red’}|\text{‘woman’}\} = 30/60 = 0.5$.
- If the chosen person is a man, the probability that he is wearing a red shirt is $\Pr\{\text{‘red’}|\text{‘man’}\} = 10/40 = 0.25$.
- **Question.** What is the probability that the chosen person is a woman wearing a red shirt?
- **Question.** If the chosen person is not wearing a red shirt, what is the probability that such person is a man?

Conditional probabilities

- Consider a pair of RVs (X, Y) with joint probability distribution $\Pr\{X = x_i \text{ and } Y = y_j\} = t_{ij}$ and marginal distributions $\Pr\{X = x_i\} = r_i$ and $\Pr\{Y = y_j\} = s_j$.
- The **conditional probabilities** are given by:
  $\Pr\{X = x_i | Y = y_j\} \overset{\text{def}}{=} \Pr\{X = x_i \text{ given that } Y = y_j\} = \frac{t_{ij}}{s_j}$ and
  $\Pr\{Y = y_j | X = x_i\} \overset{\text{def}}{=} \Pr\{Y = y_j \text{ given that } X = x_i\} = \frac{t_{ij}}{r_i}$.
- **Sum rule**: $\sum_i \Pr\{X = x_i | Y = y_j\} = 1$ for all j and $\sum_j \Pr\{Y = y_j | X = x_i\} = 1$ for all i.
- In the previous example:
  $\Pr\{\text{'woman in red'}\} = \Pr\{\text{'woman' and 'red'}\} = \Pr\{\text{'red'}|\text{'woman'}\} \times \Pr\{\text{'woman'}\} = 0.5 \times 0.6 = 0.3$.
- In the previous example:
  $\Pr\{\text{'man'}|\text{'other color'}\} = \frac{\Pr\{\text{'man' and 'other color'}\}}{\Pr\{\text{'other color'}\}} = \frac{\Pr\{\text{'other color'}|\text{'man'}\} \times \Pr\{\text{'man'}\}}{1 - \Pr\{\text{'red'}\}} = \frac{(1 - \Pr\{\text{'red'}|\text{'man'}\}) \times \Pr\{\text{'man'}\}}{1 - \Pr\{\text{'red'}\}} = \frac{(1 - 0.25) \times 0.4}{1 - 0.4} = \frac{0.75 \times 0.4}{0.6} = 0.5$. (Indeed, there are exactly 30 women and 30 men who are not wearing a red shirt.)
- **Notice**: $\Pr\{\text{'man'}|\text{'other color'}\} \neq \Pr\{\text{'other color'}|\text{'man'}\}$. The interpretation of conditional probabilities sometime leads to interesting paradoxes: search for, e.g., the 'Prosecutor's fallacy' and the 'Monty Hall problem'.

Conditional entropy

- Consider a pair of RVs (X, Y) with joint probability distribution $\Pr\{X = x_i \text{ and } Y = y_j\} = t_{ij}$ and marginal distributions $\Pr\{X = x_i\} = r_i$ and $\Pr\{Y = y_j\} = s_j$.
- The **conditional entropy** of X given Y is defined as:

$$H(X|Y) \overset{\text{def}}{=} \sum_j s_j H(X|Y = y_j)$$

  where
  $H(X|Y = y_j) = -\sum_i \Pr\{X = x_i | Y = y_j\} \log_2 \Pr\{X = x_i | Y = y_j\}$.
- **Theorem.** $H(X|Y) = H(X, Y) - H(Y)$.
  Proof. $H(X|Y) = -\sum_j s_j (\sum_i \frac{t_{ij}}{s_j} \log_2 \frac{t_{ij}}{s_j}) = -\sum_{ij} t_{ij} (\log_2 t_{ij} - \log_2 s_j) = -\sum_{ij} t_{ij} \log_2 t_{ij} + \sum_{ij} t_{ij} \log_2 s_j = H(X, Y) + \sum_j (\sum_i t_{ij}) \log_2 s_j = H(X, Y) + \sum_j s_j \log_2 s_j = H(X, Y) - H(Y)$. □

Properties of the conditional entropy

- By definition $H(X|Y)=\sum_j s_j H(X|Y=y_j)$, i.e., it is the average of positive quantities $H(X|Y=y_j)$.
- Hence, $H(X|Y)=0$ iff $H(X|Y=y_j)=0$ for all j, namely, iff the value of X is always certain given the value of Y.
- **Theorem.** $H(X|Y)\geq 0$, with equality iff there exists a function $f:\mathcal{Y}\to\mathcal{X}$ such that $X=f(Y)$.

  Proof. $H(X|Y)$ is non-negative because, by definition, it is the average of non-negative quantities. This, in particular, implies that $H(X|Y)=0$ iff $H(X|Y=y_j)=0$ for all j. This means that, for each j, there exists $\bar{k}(j)$ such that $\Pr\{X=x_i|Y=y_j\}=\delta_{i,\bar{k}(j)}$. In other words, if $Y=y_j$, then, with probability one, $X=x_{\bar{k}(j)}$. This is what we mean when saying that 'X is function of Y.' More formally: $X=f(Y)$, where $f:\mathcal{Y}\to\mathcal{X}$ is defined by $f(y_j)\overset{\text{def}}{=}x_{\bar{k}(j)}$. □
- **Theorem.** $H(X|Y)=H(X)$ iff X and Y are independent.

  Proof. Since $H(X|Y)=H(X,Y)-H(Y)$, the condition $H(X|Y)=H(X)$ holds iff $H(X,Y)-H(Y)=H(X)$ or, equivalently, iff $H(X,Y)=H(X)+H(Y)$. But before we showed that $H(X,Y)=H(X)+H(Y)$ iff X and Y are independent. □

Information

- We define the **mutual information** of X and Y as

$$\begin{aligned} I(X;Y) &\overset{\text{def}}{=} H(X)+H(Y)-H(X,Y)\\ &= H(X)-H(X|Y)\\ &= H(Y)-H(Y|X). \end{aligned}$$

- **Theorem.**
  1. $I(X;Y)=I(Y;X)$
  2. $0\leq I(X;Y)\leq\min\{H(X),H(Y)\}$
  3. $I(X;Y)=0$ iff X and Y are independent
  4. $I(X;Y)=H(X)$ iff $\exists f:\mathcal{Y}\to\mathcal{X}$ such that $X=f(Y)$
  5. $I(X;Y)=H(Y)$ iff $\exists g:\mathcal{X}\to\mathcal{Y}$ such that $Y=g(X)$

  Proof. Point (1): by definition. Point (2), lower bound: proved before (slide 8). Point (2), upper bound: because, due to positivity of the conditional entropy, $I(X;Y)\overset{\text{def}}{=}H(X)-H(X|Y)\leq H(X)$ and $I(X;Y)\overset{\text{def}}{=}H(Y)-H(Y|X)\leq H(Y)$. Point (3): proved before (slide 8). Point (4): because $I(X;Y)=H(X)$ iff $H(X|Y)=0$, i.e., iff $X=f(Y)$. Point (5): because $I(X;Y)=H(Y)$ iff $H(Y|X)=0$, i.e., iff $Y=g(X)$. □

Summary of lecture one

- **Q:** How uncertain is X? **A:** Compute the entropy

$$H(X) = -\sum_i \Pr\{X = x_i\} \log_2 \Pr\{X = x_i\}$$

- **Q:** How uncertain is X knowing Y? **A:** Compute the conditional entropy

$$\begin{aligned} H(X|Y) &= -\sum_{ij} \Pr\{X = x_i, Y = y_j\} \log_2 \frac{\Pr\{X = x_i, Y = y_j\}}{\Pr\{Y = y_j\}} \\ &= H(X,Y) - H(Y) \end{aligned}$$

- **Q:** How much information about X is contained in Y? How much information about Y is contained in X? How much dependent are X and Y? **A:** Compute the mutual information

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

Keywords for lecture one

random variable, entropy, bit, joint probability distribution, marginal probability distribution, dependent and independent random variables, conditional probability, conditional entropy, mutual information