Concepts of Probability and Information Theory

COMP41960

Félix Balado

School of Computer Science
University College Dublin

Information Sources

- Discrete information source: any entity that sequentially generates elements from a discrete alphabet (set Ω)
  - example: a person typing characters from the alphabet $\Omega = \{\text{a, b, c, \ldots, w, y, z}\}$
- The occurrence of an element (or of an event: a set an element belongs to) is not known before the source generates it, but we know that the element has to belong to Ω
  - a discrete information source can be modelled by assigning probabilities to all events from Ω
- A discrete random variable (r.v.) X can be defined by mapping events from Ω to a support set $\mathcal{X} \subset \mathbb{R}$
- A r.v. can model a discrete information source; examples:
  - map each letter in Ω to a number in $\mathcal{X} = \{0, 1, 2, \ldots, 25\}$
  - map vowels in Ω to 1 and consonants to 0: $\mathcal{X} = \{0, 1\}$

Random Variables

- Take $\mathcal{X} = \{x^{(1)}, \ldots, x^{(m)}\}$ to be the support set of a r.v. X
  - cardinality of $\mathcal{X}$ (support set size): $|\mathcal{X}| = m$
- Each $x \in \mathcal{X}$ can be assigned a probability, depending on the likelihood of the event from Ω that leads to x
  - the probability mass function (pmf) of r.v. X is the set of all probabilities $p_X(x) = \Pr(X = x)$, with $x \in \mathcal{X}$
  - notation: we just write $p(x)$ if r.v. X is understood
- Properties of the distribution of X (i.e., its pmf):
  - $0 \leq p(x) \leq 1$ for any $x \in \mathcal{X}$
  - $\sum_{x \in \mathcal{X}} p(x) = 1$
- Outcome (or realisation) of a r.v.: value $x \in \mathcal{X}$ drawn from X

Joint Distributions

- Two or more random variables are jointly described by means of their joint pmf
  - example: X, Y are jointly described by probabilities $0 \leq p_{X,Y}(x, y) \leq 1$, with $(x, y) \in \mathcal{X} \times \mathcal{Y}$
  - of course, it must hold that $\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x, y) = 1$

Joint Distributions

- Two or more random variables are jointly described by means of their joint pmf
  - example: X, Y are jointly described by probabilities $0 \leq p_{X,Y}(x,y) \leq 1$, with $(x,y) \in \mathcal{X} \times \mathcal{Y}$
  - of course, it must hold that $\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y) = 1$
- What is the pmf of X after observing an outcome y of Y?
  - pmf of X conditioned to $Y = y$ (*a posteriori* probability):

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

Joint Distributions

- Two or more random variables are jointly described by means of their joint pmf
  - example: X, Y are jointly described by probabilities $0 \leq p_{X,Y}(x,y) \leq 1$, with $(x,y) \in \mathcal{X} \times \mathcal{Y}$
  - of course, it must hold that $\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y) = 1$
- What is the pmf of X after observing an outcome y of Y?
  - pmf of X conditioned to $Y = y$ (*a posteriori* probability):

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

- If $p(x|y) = p(x)$ <u>for all x, y</u> then X and Y are independent
  - <u>example</u>: if X and Y model the simultaneous tossing of two dice, then $\mathcal{X} = \mathcal{Y} = \{1,2,3,4,5,6\}$ and $p(x,y) = \frac{1}{36}$; with fair dice $p(x|y) = p(x) = \frac{1}{6}$, so X and Y are independent

Statistical Independence of Random Variables

- Product rule of probability:

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x),$$

  - X and Y are independent iff (if and only if)

$$p(x, y) = p(x)p(y) \text{ for all } x \in \mathcal{X}, y \in \mathcal{Y}$$

    - i.e., the joint is the product of the priors (probabilities without conditioning are called *a priori* probabilities, or *priors*)
  - Bayes' law (from product rule):

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

Marginalisation

- We can recover the pmf of each individual r.v. from the joint pmf: this is called marginalisation
  - pmf of X from joint pmf of X and Y:

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y), \qquad x \in \mathcal{X}$$

  - (consequence of the law of total probabilities)
- Using the product rule, we can rewrite marginalisation as

$$p(x) = \sum_{y \in \mathcal{Y}} p(x|y)p(y), \qquad x \in \mathcal{X}$$

Example: Dependent Random Variables

- $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1\}$
- Assume the following joint pmf $p_{X,Y}(x, y)$:

X \ Y	0	1
0	$\frac{1}{2}$	$\frac{1}{4}$
1	$\frac{1}{8}$	$\frac{1}{8}$

are X and Y independent?

- $p_X(0) = \sum_{y \in \mathcal{Y}} p_{X,Y}(0, y) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$
  $p_Y(0) = \sum_{x \in \mathcal{X}} p_{X,Y}(x, 0) = \frac{1}{2} + \frac{1}{8} = \frac{5}{8}$

Example: Dependent Random Variables

- $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1\}$
- Assume the following joint pmf $p_{X,Y}(x, y)$:

X \ Y	0	1
0	$\frac{1}{2}$	$\frac{1}{4}$
1	$\frac{1}{8}$	$\frac{1}{8}$

are X and Y independent?

  - $p_X(0) = \sum_{y \in \mathcal{Y}} p_{X,Y}(0, y) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$
    $p_Y(0) = \sum_{x \in \mathcal{X}} p_{X,Y}(x, 0) = \frac{1}{2} + \frac{1}{8} = \frac{5}{8}$
- So the r.v.s X and Y are not independent, because $p_X(0)p_Y(0) = \frac{15}{32} \neq p_{X,Y}(0, 0) = \frac{1}{2}$

Example: Dependent Random Variables

- $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1\}$
- Assume the following joint pmf $p_{X,Y}(x, y)$:

X \ Y	0	1
0	$\frac{1}{2}$	$\frac{1}{4}$
1	$\frac{1}{8}$	$\frac{1}{8}$

are X and Y independent?

  - $p_X(0) = \sum_{y \in \mathcal{Y}} p_{X,Y}(0, y) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$
    $p_Y(0) = \sum_{x \in \mathcal{X}} p_{X,Y}(x, 0) = \frac{1}{2} + \frac{1}{8} = \frac{5}{8}$
- So the r.v.s X and Y are not independent, because $p_X(0)p_Y(0) = \frac{15}{32} \neq p_{X,Y}(0, 0) = \frac{1}{2}$
- Equivalently, using conditional probabilities:
  - $p_{Y|X}(0|0) = \frac{p_{X,Y}(0,0)}{p_X(0)} = \frac{2}{3} \neq p_Y(0) = \frac{5}{8}$

Multivariate Joint Distributions

- We can apply the product rule recursively to a joint pmf modelling more than two r.v.s; for instance, for three r.v.s:

  $$p_{X,Y,Z}(x, y, z) = p(z|x, y)p(x, y) = p(z|x, y)p(y|x)p(x)$$

  where $(x, y, z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$

  - we can apply the product rule recursively in other equivalent ways, e.g.: $p(x, y, z) = p(x|y, z)p(y, z)$
  - we can marginalise $p(x, y, z)$ to get $p(x)$, $p(y)$, $p(x, z)$, etc

Multivariate Joint Distributions

- We can apply the product rule recursively to a joint pmf modelling more than two r.v.s; for instance, for three r.v.s:

  $$p_{X,Y,Z}(x,y,z) = p(z|x,y)p(x,y) = p(z|x,y)p(y|x)p(x)$$

  where $(x,y,z) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$
  - we can apply the product rule recursively in other equivalent ways, e.g.: $p(x,y,z) = p(x|y,z)p(y,z)$
  - we can marginalise $p(x,y,z)$ to get $p(x)$, $p(y)$, $p(x,z)$, etc
- In general, for n random variables $X_1, \ldots, X_n$,i. a valid decomposition the joint pmf is

  $$p(x_1, \ldots, x_n) = p(x_1) \prod_{k=2}^{n} p(x_k|x_1, \ldots, x_{k-1})$$

  with $(x_1, \ldots, x_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ (i.e., $x_k \in \mathcal{X}_k$)
  - $n!$ possible decompositions

Multivariate Joint Distributions: i.i.d. Case

- In many cases the n random variables we are interested in are identically distributed as some r.v. X with support $\mathcal{X}$
  - in this case, $(x_1, \ldots, x_n) \in \mathcal{X} \times \cdots \times \mathcal{X} = \mathcal{X}^n$

Multivariate Joint Distributions: i.i.d. Case

- In many cases the n random variables we are interested in are identically distributed as some r.v. X with support $\mathcal{X}$
  - in this case, $(x_1, \ldots, x_n) \in \mathcal{X} \times \cdots \times \mathcal{X} = \mathcal{X}^n$
- n identically distributed random variables are independent iff

$$p(x_1, \ldots, x_n) = \prod_{k=1}^{n} p(x_k)$$

  - i.e., the joint pmf is the product of the n priors
  - $\rightarrow$ a discrete information source modellable through independent and identically distributed (i.i.d.) r.v.s is called memoryless

Multivariate Joint Distributions: i.i.d. Case

- In many cases the n random variables we are interested in are identically distributed as some r.v. X with support $\mathcal{X}$
  - in this case, $(x_1, \ldots, x_n) \in \mathcal{X} \times \cdots \times \mathcal{X} = \mathcal{X}^n$
- n identically distributed random variables are independent iff

$$p(x_1, \ldots, x_n) = \prod_{k=1}^{n} p(x_k)$$

  - i.e., the joint pmf is the product of the n priors

  → a discrete information source modellable through independent and identically distributed (i.i.d.) r.v.s is called memoryless
- Drawing one outcome from each of the r.v.s $X_1, \ldots, X_n$ i.i.d. as X is the same as drawing n independent outcomes from X
  - let n_x be the number of outcomes equal to $x \in \mathcal{X}$
  - then $\mathrm{fr}(x) = \frac{n_x}{n} \to p_X(x)$ as $n \to \infty$ (frequency interpretation of probability)
  - a normalised histogram empirically approximates the pmf of X

Expectation

- The expectation of r.v. X is the sum of all its possible outcomes weighted by their likelihoods

$$E(X) = \sum_{x \in \mathcal{X}} x \; p(x)$$

  - synonyms: average, mean, expected value

Expectation

- The expectation of r.v. X is the sum of all its possible outcomes weighted by their likelihoods

$$E(X) = \sum_{x \in \mathcal{X}} x \; p(x)$$

  - synonyms: average, mean, expected value
- The rationale for $E(X)$ comes from the law of large numbers
  - if r.v.s $X_1, \ldots, X_n$ are i.i.d. as X, then

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to E(X) \quad \text{as } n \to \infty$$

  - intuition: if $x_1, \ldots, x_n$ are outcomes of $X_1, \ldots, X_n$ then

$$\frac{1}{n} \sum_{i=1}^{n} x_i$$

(where n_x is the number of x_i equal to x)

Expectation

- The expectation of r.v. X is the sum of all its possible outcomes weighted by their likelihoods

$$E(X) = \sum_{x \in \mathcal{X}} x \, p(x)$$

  - synonyms: average, mean, expected value
- The rationale for $E(X)$ comes from the law of large numbers
  - if r.v.s $X_1, \ldots, X_n$ are i.i.d. as X, then

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to E(X) \quad \text{as } n \to \infty$$

  - intuition: if $x_1, \ldots, x_n$ are outcomes of $X_1, \ldots, X_n$ then

$$\frac{1}{n} \sum_{i=1}^{n} x_i = \sum_{x \in \mathcal{X}} x \frac{n_x}{n}$$

    (where n_x is the number of x_i equal to x)

Expectation

- The expectation of r.v. X is the sum of all its possible outcomes weighted by their likelihoods

$$E(X) = \sum_{x \in \mathcal{X}} x \, p(x)$$

  - synonyms: average, mean, expected value
- The rationale for $E(X)$ comes from the law of large numbers
  - if r.v.s $X_1, \ldots, X_n$ are i.i.d. as X, then

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to E(X) \quad \text{as } n \to \infty$$

  - intuition: if $x_1, \ldots, x_n$ are outcomes of $X_1, \ldots, X_n$ then

$$\frac{1}{n} \sum_{i=1}^{n} x_i = \sum_{x \in \mathcal{X}} x \, \mathrm{fr}(x) \to E(X) \quad \text{as } n \to \infty$$

(where n_x is the number of x_i equal to x)

Expectation (Functions of r.v.s)

Functions of r.v.s are also r.v.s, so they have expectations too:

- if $g : \mathbb{R} \to \mathbb{R}$ and X is a r.v., then $Y = g(X)$ is a r.v.

$$E(Y) = E(g(X)) = \sum_{x \in \mathcal{X}} g(x)\, p(x)$$

- if $g : \mathbb{R}^2 \to \mathbb{R}$ and X, Y are r.v.s then $Z = g(X, Y)$ is a r.v.

$$E(Z) = E(g(X, Y)) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} g(x, y)\, p(x, y)$$

Concepts of Information Theory: Entropy

- How surprising is the outcome x of a single r.v. X?
  - a lot if x is not likely, but not too much if x is likely
  - therefore the surprise associated to observing $x \in \mathcal{X}$ is inversely related to its probability, that is, $\frac{1}{p(x)}$

Concepts of Information Theory: Entropy

- How surprising is the outcome x of a single r.v. X?
  - a lot if x is not likely, but not too much if x is likely
  - therefore the surprise associated to observing $x \in \mathcal{X}$ is inversely related to its probability, that is, $\frac{1}{p(x)}$
  - let us define the amount of surprise associated to x as

$$\log \frac{1}{p(x)} = -\log p(x)$$

    - $\frac{1}{p(x^{(1)})} < \frac{1}{p(x^{(2)})} \leftrightarrow \log \frac{1}{p(x^{(1)})} < \log \frac{1}{p(x^{(2)})}$ (as log is increasing)
    - also, if $p(x) = 1$ then $-\log p(x) = 0$ (zero surprise)

Concepts of Information Theory: Entropy

- How surprising is the outcome x of a single r.v. X?
  - a lot if x is not likely, but not too much if x is likely
  - therefore the surprise associated to observing $x \in \mathcal{X}$ is inversely related to its probability, that is, $\frac{1}{p(x)}$
  - let us define the amount of surprise associated to x as

$$\log \frac{1}{p(x)} = -\log p(x)$$

    - $\frac{1}{p(x^{(1)})} < \frac{1}{p(x^{(2)})} \leftrightarrow \log \frac{1}{p(x^{(1)})} < \log \frac{1}{p(x^{(2)})}$ (as log is increasing)
    - also, if $p(x) = 1$ then $-\log p(x) = 0$ (zero surprise)
- Entropy of discrete r.v. X: average surprise about X

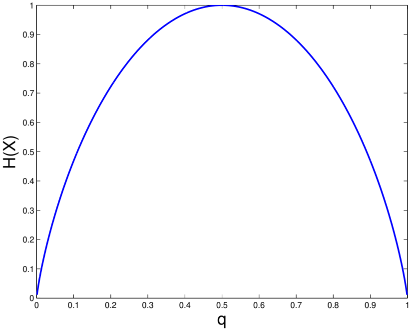
$$H(X) = E(-\log p(X)) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

  - equivalently, average uncertainty about X
  - expectation of r.v. $Y = -\log p(X)$

Concepts of Information Theory: Entropy

- The units of $H(X)$ depend on the base of the logarithm used
  - units are bits if base 2 logarithm is used (**default hereafter**)
  - notation: $\log a = \log_2 a$ and $\ln a = \log_e a$ (base e logarithm)
- Example: if $\mathcal{X} = \{0, 1\}$, with $p_X(1) = q$ and $p_X(0) = 1 - q$,

$$H(X) = q \log \frac{1}{q} + (1-q) \log \frac{1}{1-q} \quad \text{bits}$$



(note: $0 \log 0 = 0$)

Concepts of Information Theory: Entropy

- A "bit" (binary digit) can only be 0 or 1, but we have seen that $H(X)$ can take non-integer values; why?
  - answer: importantly, entropy can also be interpreted as the average information content of X
- Intuitive explanation: assume r.v. X in previous example
  - if X is deterministic ($p(1) = 1$, $p(0) = 0$), $H(X) = 0$ bits
    - example of successive outcomes of X: 1,1,1,1,1,1,1,1...
    - 0 bits per outcome asymptotically needed to represent this sequence (as we know the outcomes of X beforehand)
  - if X is completely random ($p(1) = p(0) = \frac{1}{2}$), $H(X) = 1$ bit
    - example of successive outcomes of X: 1,0,0,1,0,1,1,0...
    - 1 bit per outcome needed to represent this sequence (i.e., either "0" or "1" per outcome)

Concepts of Information Theory: Entropy

- A "bit" (binary digit) can only be 0 or 1, but we have seen that $H(X)$ can take non-integer values; why?
  - answer: importantly, entropy can also be interpreted as the average information content of X
- Intuitive explanation: assume r.v. X in previous example
  - if X is deterministic ($p(1) = 1$, $p(0) = 0$), $H(X) = 0$ bits
    - example of successive outcomes of X: 1,1,1,1,1,1,1,1...
    - 0 bits per outcome asymptotically needed to represent this sequence (as we know the outcomes of X beforehand)
  - if X is completely random ($p(1) = p(0) = \frac{1}{2}$), $H(X) = 1$ bit
    - example of successive outcomes of X: 1,0,0,1,0,1,1,0...
    - 1 bit per outcome needed to represent this sequence (i.e., either "0" or "1" per outcome)
  - If X is in between, then we need a number of bits per outcome somewhere between 0 and 1, <u>also given by $H(X)$</u>
    - on average, not possible to describe a source of information represented by X with less than $H(X)$ bits/outcome

Concepts of Information Theory (II)

- Joint entropy of two discrete random variables X and Y

$$
\begin{aligned}
H(X, Y) &= E(-\log p(X, Y)) \\
&= -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x, y) \log p(x, y)
\end{aligned}
$$

Concepts of Information Theory (II)

- Joint entropy of two discrete random variables X and Y

$$
\begin{aligned}
H(X, Y) &= E(-\log p(X, Y)) \\
&= -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x, y)\log p(x, y)
\end{aligned}
$$

- Conditional entropy, $H(X|Y)$
  - how surprised are we about X when we observe Y first?
  - $H(X|Y)$ is the average of $H(X|Y = y)$ for all outcomes y of Y

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

Concepts of Information Theory (II)

- Joint entropy of two discrete random variables X and Y

$$\begin{aligned} H(X,Y) &= E(-\log p(X,Y)) \\ &= -\sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p(x,y)\log p(x,y) \end{aligned}$$

- Conditional entropy, $H(X|Y)$
  - how surprised are we about X when we observe Y first?
  - $H(X|Y)$ is the average of $H(X|Y=y)$ for all outcomes y of Y

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

- Chain rule: $H(X,Y) = H(Y|X) + H(X) = H(X|Y) + H(Y)$

Properties of Entropy

1. $H(X) \geq 0$
   - $H(X) = 0$ for deterministic variables only (i.e, when there is $x \in \mathcal{X}$ such that $p(x) = 1$)

Properties of Entropy

1. $H(X) \geq 0$
   - $H(X) = 0$ for deterministic variables only (i.e, when there is $x \in \mathcal{X}$ such that $p(x) = 1$)
2. $H(X) \leq \log |\mathcal{X}|$; *proof*: (let $m = |\mathcal{X}|$)
   1. first, consider the inequality $\ln x \leq x - 1$: if $\sum_{i=1}^{m} a_i = 1$ and $\sum_{i=1}^{m} b_i = 1$, with $a_i \geq 0, b_i \geq 0$, the inequality implies

$$-\sum_{i=1}^{m} a_i \log a_i \leq -\sum_{i=1}^{m} a_i \log b_i \quad (\text{Gibbs inequality})$$

   2. then, use the particular case $b_i = \frac{1}{m}$ in the inequality above

Properties of Entropy

1. $H(X) \geq 0$
   - $H(X) = 0$ for deterministic variables only (i.e, when there is $x \in \mathcal{X}$ such that $p(x) = 1$)
2. $H(X) \leq \log |\mathcal{X}|$; *proof*: (let $m = |\mathcal{X}|$)
   1. first, consider the inequality $\ln x \leq x - 1$: if $\sum_{i=1}^{m} a_i = 1$ and $\sum_{i=1}^{m} b_i = 1$, with $a_i \geq 0, b_i \geq 0$, the inequality implies

      $$-\sum_{i=1}^{m} a_i \log a_i \leq -\sum_{i=1}^{m} a_i \log b_i \quad \text{(Gibbs inequality)}$$

   2. then, use the particular case $b_i = \frac{1}{m}$ in the inequality above
3. The uniform distribution ($p(x) = \frac{1}{|\mathcal{X}|} = \frac{1}{m}$ for all $x \in \mathcal{X}$) yields $H(X) = \log |\mathcal{X}|$; so, it maximises entropy for $|\mathcal{X}| = m$
   - no other distribution (pmf) yields greater entropy
   - it is the most "random", most surprising, least compressible distribution

Properties of Entropy

4. $H(X, Y) \leq H(X) + H(Y)$
   - *proof*: $E(-\log p(X, Y)) \leq E(-\log(p(X)p(Y)))$ because of Gibbs inequality and the fact that $g(x, y) = p(x)p(y)$ is a pmf
   - $H(X, Y) = H(X) + H(Y)$ iff X and Y are independent
5. Conditioning cannot increase entropy:

$$H(X|Y) \leq H(X)$$

   proof: use the chain rule for $H(X, Y)$ and the inequality above

   - $H(X|Y) = H(X)$ iff X and Y are independent

Concepts of Information Theory (IV)

- Mutual information:

$$I(X;Y) = E\left(\log \frac{p(X,Y)}{p(X)p(Y)}\right)$$

- In terms of entropies

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

- Interpretation of $I(X;Y)$:
  - the reduction in uncertainty about X due to knowledge of Y

Concepts of Information Theory (and V)

- Properties of the mutual information
  - $I(Y; X) = I(X; Y)$ (so interpretation is valid both ways)
  - $I(X; Y) \geq 0$; *proof*: conditioning cannot increase entropy
  - $I(X; Y) = 0$ iff X and Y are independent
  - $I(X; X) = H(X)$ (entropy can be called "self-information")

