ECS7024 Statistics for Artificial Intelligence and Data Science

Topic 19: Introduction to Information Theory

William Marsh

Outline

- Aim: Introduce ideas of information theory
  - Increasingly likely to encounter information theory in machine learning

- Entropy: measures surprise
- Relative entropy: measure information gain
- Mutual information: measures ‘correlation’

Motivation

- Previously: strength of correlation
  - Approach: average of $(x_i - \bar{x})(y_i - \bar{y})$
  - Covariance
  - Only applies to continuous variables (with a mean)
  - Linear relationships

- Is there a more-general approach to 'dependence'?

Entropy

Entropy: Definition

- X is a discrete variable
- Entropy (measured in 'bits')

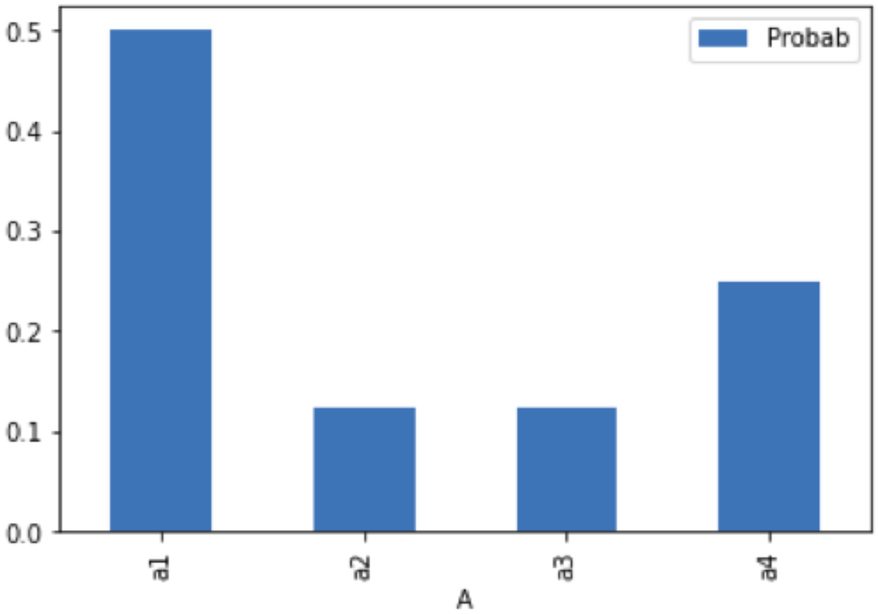
Log of number < 1 is -ve

$$H(X) = -\sum_{i}^{n} p(x_i).\,log_2\, p(x_i)$$

$$H(X) = \sum_{i}^{n} p(x_i).\,log_2\left(\frac{1}{p(x_i)}\right)$$

Log of 1/p is –log p

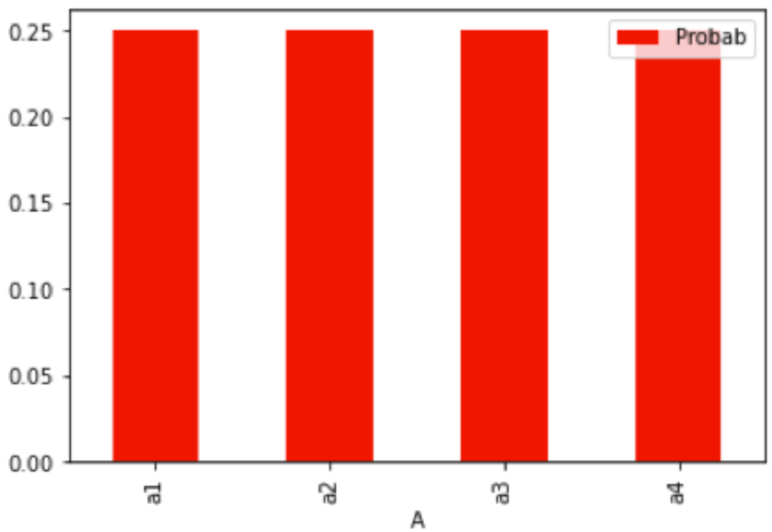
Example



A	p(a)	log(1/p(a))	H(a)
a1	1/2	1	1/2
a2	1/8	3	3/8
a3	1/8	3	3/8
a4	1/4	2	1/2

$$H(A) = \tfrac{1}{2} + \tfrac{3}{4} + \tfrac{1}{2} = 1\tfrac{3}{4}$$

Example: Uniform



A	p(a)	log(1/p(a))	H(a)
a1	1/4	2	1/2
a2	1/4	2	1/2
a3	1/4	2	1/2
a4	1/4	2	1/2

H(A) = 2

Origin and Interpretation

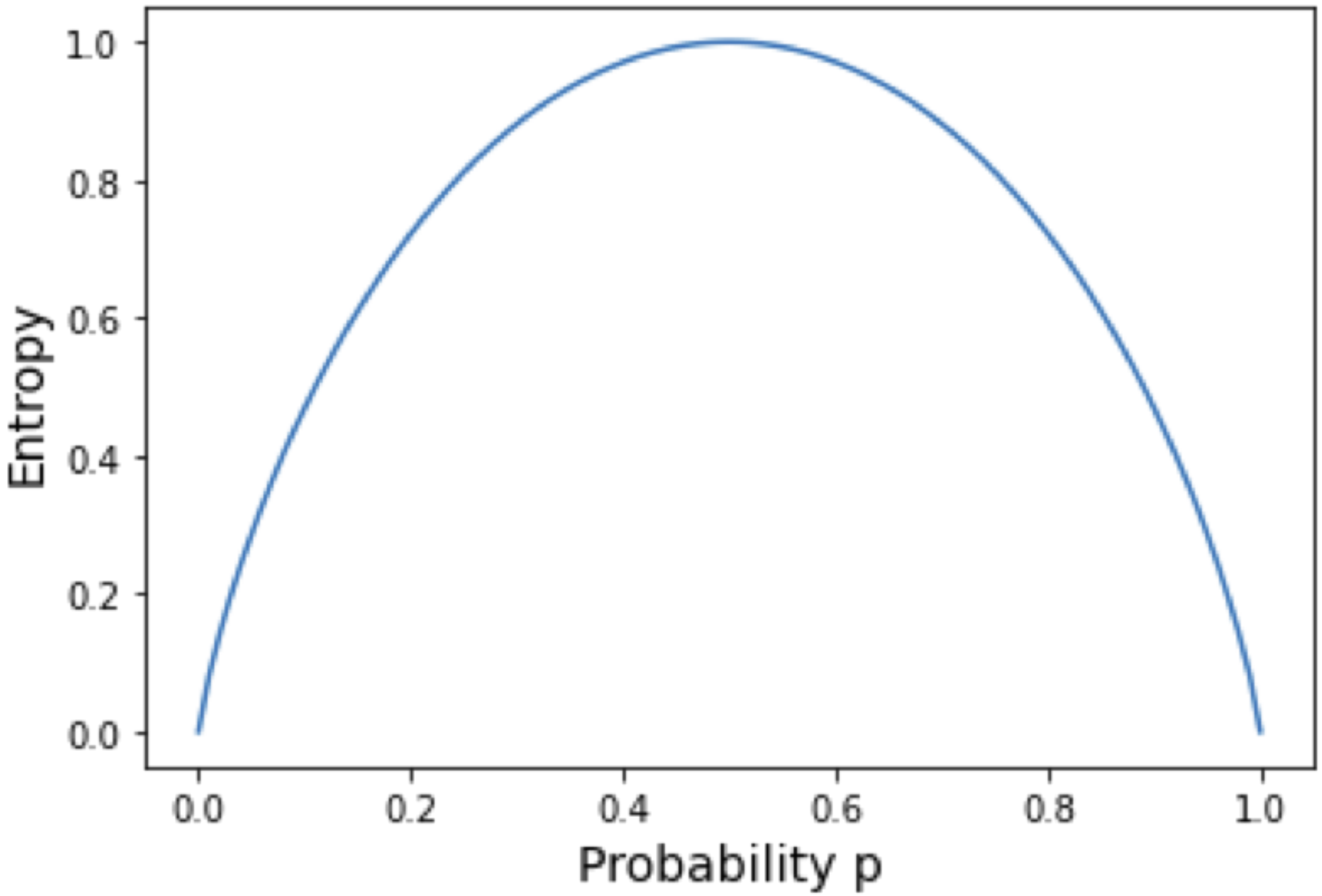
- Origin in communication
  - Best coding scheme for sending message
  - Length of code for ‘a’ given by H(a)

- Information content - surprise

Event	Probability	Entropy / Information
I did not win lottery	High	Low
I did win lottery	Low	High

Example: Biased Coins

- Consider a coin toss with P(‘heads’) = p



Entropy Properties

- Consider the information of an event of probability p
  - What do we learn when ‘e’ occurs?

a. info(p) >= 0

b. info(1) = 0

c. if $p_1 > p_2$ then $info(p_1) < info(p_2)$

d. $info(p_1 \text{ and } p_2) = info(p_1) + info(p_2)$

No negative information

Learn nothing when e certain

Learn more when e less probable

Information from separate events adds

Entropy of Two Variables

- Entropy applies to two (or more) variables
- H(X,Y)
  - Each case has probability e.g. P(x1, y2)
- Properties

Max when X and Y independent

$$H(X,Y) \leq H(X) + H(Y)$$

$$\max[H(X), H(Y)] \leq H(X,Y)$$

Min when X determines Y (or vice versa)

Relative Entropy (KL Divergence)

Compare two distributions over same outcomes

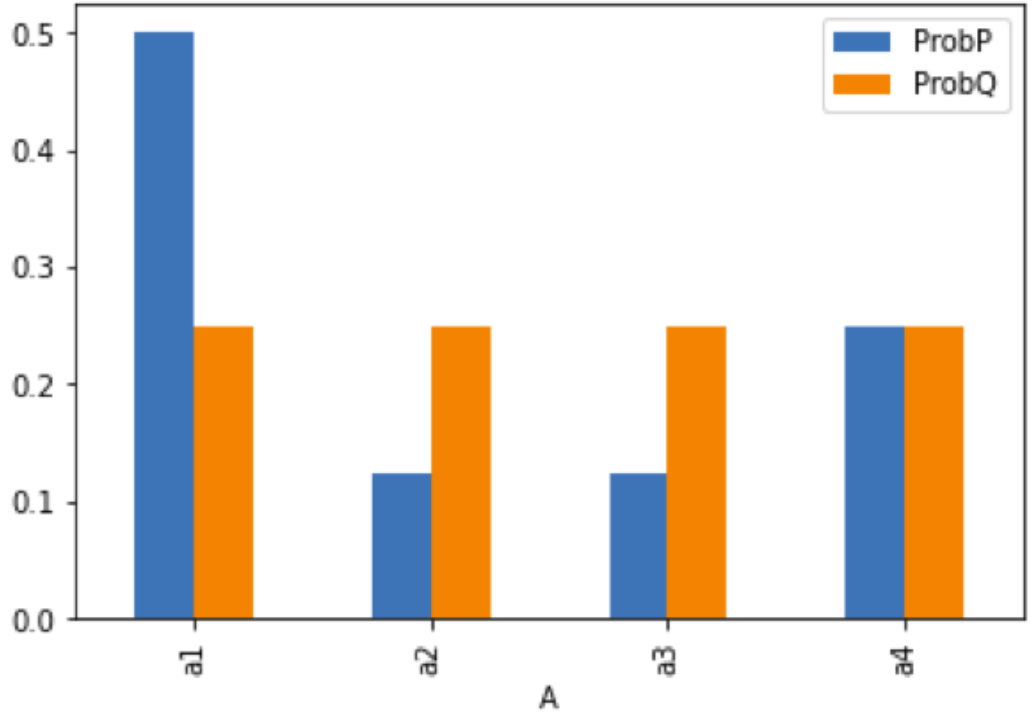
Definition of Relative Entropy

- Also known as KL-divergence
  - Kullback
  - Leibler
- Compares two probability distribution P, Q
  - Same states $x \in X$
  - How closely does Q approximate P

$$D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) log\left(\frac{P(x)}{Q(x)}\right)$$

Example

- How closely does Q
- ... approximate P



A	P(a)	Q(a)	P(a)/Q(a)	log(P(a)/Q(a))	P(a). log(...)
a1	1/2	1/4	2	1	1/2
a2	1/8	1/4	1/2	-1	-1/8
a3	1/8	1/4	1/2	-1	-1/8
a4	1/4	1/4	1	0	0

$$D_{KL}(P \parallel Q) = 1/4$$

Some Properties

$$D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) log\left(\frac{P(x)}{Q(x)}\right)$$

- $D_{KL}(P \parallel Q) \geq 0$
- Equals zero if P same as Q

- Not symmetric

$$D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$$

Interpretation as Information Gain

- How much information is gained by using P instead of Q

- Bayesian updating
  - Q is prior
  - P is posterior given new data (observations)
  - $D_{KL}(P \parallel Q)$ measure the information gained from the new data

Mutual Information

Measure of dependence, not just linear

Definition

- Mutual Information I(X;Y)
  - X, Y are probability distributions
  - Not necessarily same states

- Definition 1:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

- Equivalent definition 2:

$$I(X;Y) = D_{KL}(P(X,Y) \parallel P(X) \times P(Y))$$

Example (Definition 2)

- Joint probability P(A, B)

- Marginal distributions

A	B	Probability
a1	b1	1/18
a1	b2	3/18
a2	b1	1/18
a2	b2	5/18
a3	b1	7/18
a3	b2	1/18

marginalise B

B	Probability
b1	(1+1+7)/18 = 1/2
b2	(3+5+1)/18 = 1/2

marginalise A

A	Probability
a1	(1+3)/18 = 2/9
a2	(1+5)/18 = 3/9
a3	(7+1)/18 = 4/9

Product P(A) x P(B)

- If A and B independent then P(A,B) = P(A). P(B)

B	Probability
b1	(1+1+7)/18 = 1/2
b2	(3+5+1)/18 = 1/2

A	Probability
a1	(1+3)/18 = 2/9
a2	(1+5)/18 = 3/9
a3	(7+1)/18 = 4/9

A	B	Probability
a1	b1	2/18
a1	b2	2/18
a2	b1	3/18
a2	b2	3/18
a3	b1	4/18
a3	b2	4/18

Calculate I(A;B)

$$I(A;B) = D_{KL}(P(A,B) \parallel P(A)\times P(B))$$

A	B	P(A,B)	P(A) x P(B)	Ratio	Log2 (Ratio)	P(A,B) x Log2()
a1	b1	0.056	0.111	0.50	-1.000	-0.056
a1	b2	0.167	0.111	1.50	0.585	0.097
a2	b1	0.056	0.167	0.33	-1.585	-0.088
a2	b2	0.278	0.167	1.67	0.737	0.205
a3	b1	0.389	0.222	1.75	0.807	0.314
a3	b2	0.056	0.222	0.25	-2.000	-0.111

Total = 0.361

Use of I(A;B)

- Measure of dependence
  - Applied to discrete (and continuous) ✓
  - Not just linear ✓
  - Not normalised ✗
  - Not well-supported in Pandas ✗

- May encounter information gain in decision trees as a loss function

Summary

- Introduced idea from ‘information theory’
- Difficult concepts
- Be aware of possible use in ML