Closeness of Distributions

- Example: Authorship Attribution

- Given a document X
- Two sets of documents written by Alice and Bob
- Decide which one wrote X

- A basic algorithm:
  - Author representation: Distribution of bigrams
  - $B_1 = (p_1, p_2 \ldots p_{729})$
  - $B_2 = (p'_1, p'_2 \ldots p'_{729})$

  - X: $(q_1 q_2 \ldots q_{729})$
  - If $d(X, B_1) > d(X, B_2)$, X is written by B2

Relative Entropy

- Kullback-Leibler divergence: difference of one probability distribution to another

$$D(p \| q) = \sum_{x_i \in X} p(x_i) \log \frac{p(x_i)}{q(x_i)}$$

$$= E_p \log \frac{p(x)}{q(x)}$$

Example: Authorship Attribution

"Given author candidates *A = {a1 . . . aj},* it is straightforward to build a model for each author by aggregating the training documents. We can build a model for an unattributed document in the same way. We can then determine the author model that is most similar to the model of the unknown document, by calculating KLD values between author models and unknown documents to identify the target author for which the KLD value is the smallest."

Using Relative Entropy for Authorship Attribution
Ying Zhao, Justin Zobel, and Phil Vines

Application

- Design a Huffman code: given samples of information source

- estimate probabilities of source alphabet
  → Estimate from a sample of the source

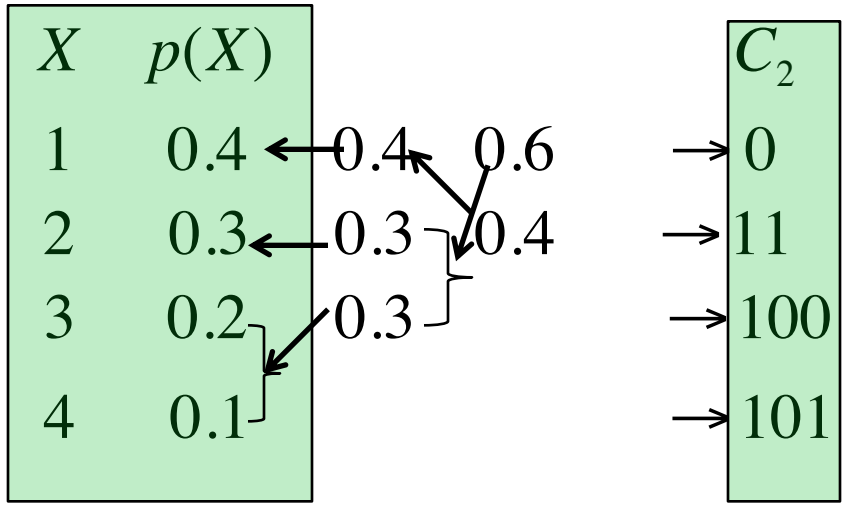
The Spanish conquest of Guatemala was a protracted conflict during the Spanish colonization of the Americas in which Spanish colonizers gradually incorporated the territory that became the modern country of Guatemal the colonial viceroyalty of New Spain. The Maya kingdoms resisted integration into the Spanish Empire with s tenacity that their defeat took almost two centuries. Pedro de Alvarado arrived in Guatemala from the newly conquered Mexico in early 1524, commanding a mixed force of Spanish conquistadors and native allies, mos The Itza Maya and other lowland groups in the Petén Basin were first contacted by Hernán Cortés but rem independent and hostile to the encroaching Spanish until 1697, when a concerted Spanish assault led by Mart Urzúa y Arizmendi finally defeated the last independent Maya kingdom. The indigenous peoples of Guatemal key elements of Old World technology such as a functional wheel, horses, steel and gunpowder; they were also extremely susceptible to Old World diseases, against which they had no resistance.And when the hole P The Spanish conquest of Guatemala was a protracted conflict during the Spanish colonization of the Americas in which Spanish colonizers gradually incorporated the territory that became the modern country of Guatemal the colonial viceroyalty of New Spain. The Maya kingdoms resisted integration into the Spanish Empire with s tenacity that their defeat took almost two centuries. Pedro de Alvarado arrived in Guatemala from the newly conquered Mexico in early 1524, commanding a mixed force of Spanish conquistadors and native allies, mos

- Suppose our estimate is q(x) – real distribution is p(x)
- How good would be our Huffman code?

Error in using incorrect distributions

- X={1,2,3,4,}; p(x) is not known -replaced by q(x).

X	$q(X)$	C_1
1	0.25	00
2	0.25	01
3	0.25	10
4	0.25	11



l_{av} = 1.9 bits

$$D(p \| q) = 0.4 \times \log \frac{0.4}{0.25} + 0.3 \times \log \frac{0.3}{0.25} + 0.2 \times \log \frac{0.2}{0.25} + 0.1 \times \log \frac{0.1}{0.25}$$

$$\approx 0.1 \text{ bits}$$

I(X;Y) = H(X) –H(X|Y)

- I(X;Y) can be seen as relative entropy of two distributions:

  p(x,y) and p(x)p(y)

Relative Entropy

- How
  one distribution diverges
  from a second

- Not a distance function
  → Not Symmetric

- D(p||q) is a measure of inefficiency of estimating p as q

  – "Average" number of extra bits when a source with distribution p is encoded using distribution q

→ Expected increase in code length is D(p||q)

Plan

- Models with two variable
- Joint distribution
- Entropy measures
  - Joint entropy, conditional entropy, mutual information, relative entropy
- Entropy as an information measure
- Min-entropy
- From data to distribution

Information measure:

Matching Intuition & formalization

- Intuition: information is not negative
- Formal proof: $H(X) \geq 0$
  - follows from definition

- Intuition: Relative entropy measures divergence– so non-negative
- Is the expression D(p||q) non-negative?
  - Formal proof: $D(p||q) \geq 0$

- Intuition: Mutual information is non-negative - information that Y gives about X
- Is the expression I(X;Y) non-negative?
  - Formal proof: $I(X;Y) \geq 0$

Mathematical proofs

1. $D(X||Y) \geq 0$
   - Equality if and only if p=q
   - Theorem 2.6.3 CT

2 and 3 follow from 1.

2. $I(X;Y) \geq 0$
   - Follow from non-negativity of relative entropy
   - $H(X|Y) \leq H(X)$

3. $H(X) \leq \log N$;
   $H(X) = \log (N)$ iff $p(x_i) = 1/N$, for all x_i
   - N is the number of possible values (size of the set)

Proving H(X) ≤ log N

Find relative entropy of p(x) and uniform distribution:

$$D(p(x) \,\|\, u_N) = \sum_{(x)} p(x) \log \frac{p(x)}{(1/N)}$$

$$= \sum_{(x,y)} p(x)[\log p(x) - \log(1/N)]$$

$$= \sum_{(x,y)} p(x)[\log p(x) + \log N]$$

$$= -H(X) + \log N \geq 0$$

Since $D(p(x), u_N) \geq 0$, ➔ $\log N \geq H(X)$

H(X|Y) ≤ H(X) *in expectation*

	$X=1$	$X=2$
$Y=1$	0	3/4
$Y=2$	1/8	1/8

- p(X=1) = 1/8 , p(X=2) = 7/8
  →H(X) = 0.544 bits
- p (Y=1) = 3/4 , p (Y=2) = 1/4
  →H(Y) = 1/4 ×2+ 3/4 log 3/4 = 1/2 + 3/4 ×.42 = .8 bit

- H(X|Y = 1) = 0 bit
- H(X|Y = 2) = 1 bit
- H(X|Y) =3/4 × 0 +1/4 × 1 = 0.25 bits

- Uncertainty increases when Y=2 is observed, but decreases on “average”.

Independence bound

- Independence bound on entropy

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

Proof: $H(X,Y) = H(X) + H(Y|X)$

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

Equality holds if and only if X and Y are independent.

Deterministic functions

- X= { 1,..n}, $p(X=i) = p_i$, i=1,…n
- $H(X) = -\Sigma\, p_i \log p_i$

- Y= f(X)
  f: { 1,..n} → { 1,..m} (deterministic fun)

- What is H(Y)?
  - Find $p(Y=i) = q_i$, i=1,…m
  - $H(Y) = -\Sigma\, q_i \log q_i$

- $H(Y) \leq H(X)$ (proof?)

Min-entropy

- In some applications other types of entropy is more appropriate.

- X, p(X)
  Min-entropy of a distribution is defined as:
  $H_\infty(X) = -\log \max_x p(x)$

- Example:
- $p(x_1)=2^{-1}$
- $p(x_2)=p(x_3)\ldots\ldots p(x_{256})= 2^{-8}$

Min-entropy is an important security measure for random number generators.

- H(X)~ 7.8 bits
- $H_\infty(X) = -\log_2(2^{-1})\sim 1$ bit

- Min-entropy measures the success chance of the best guess.

Min-entropy

- Shannon entropy is expected uncertainty:
  Not a good measure for success chance of best guess

- Example: $\mathcal{X} = \{0,1\}^3$,
- p(000)=9/16, p(001)=…p(111)=1/16= 2^{-4}
- H(X) = -(7/16) log(2^{-4}) - 9/16 log (9/16)
  = 2.22 bit

- $H_\infty(X)$ ~1 bit

- Entropy relations for min-entropy are different..