UNIVERSITY of DELAWARE UNIVERSITY of DE

Basics of Information Theory

(Search and Data Mining)

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

UNIVERSITY of DELAWARE UNIVERSITY of DE

Information Theory

- Information theory deals with uncertainty and the transfer or storage of quantified information in the form of bits.
- Useful concepts for text analysis
  - Entropy
  - KL divergence
  - Mutual information

UNIVERSITY of DELAWARE UNIVERSITY of DE

Entropy

- Measuring uncertainty of a random variable

UNIVERSITY of DELAWARE UNIVERSITY of DE

Entropy: A Motivating Example

- We use random variable X to represent the outcomes of a coin flipping.

$$X = \begin{cases} 1 & if\ heads \\ 0 & if\ tails \end{cases}$$

- P(X=1)=0.5 is more random than P(X=1)=0.9.
- How does one quantitatively measure the randomness of a random variable like X?

UNIVERSITY of DELAWARE UNIVERSITY of DE

Entropy: Definition

- Entropy H(X) measures the uncertainty/randomness of random variable X.

$$H(X) = H(p) = \sum_{x \in \Omega} -p(x)\log p(x) \quad \Omega = all\ \ possible\ values$$
$$Define\ \ 0\log 0 = 0,\ \ \log = \log_2$$

- Interpretation of Entropy:
  - Minimum average number of bits to compress values of X
  - The more random X is, the harder to compress

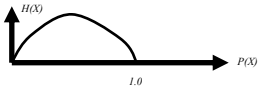
$$"Information\ of\ x" = "\#bits\ to\ code\ x" = -\log p(x) \quad H(X) = E_p[-\log p(x)]$$

UNIVERSITY of DELAWARE UNIVERSITY of DE

An example of Entropy (X: Head)

- For a fair coin, we have p(X=1) = P(X=0) =0.5.
  So, we get $H(X) = -0.5 * log0.5 - 0.5 * log0.5 = 1$
- For a complete biased coin with p(X=1) =1, P(X=0) =0.
  So, we get $H(X) = -0 * log0 - 1 * log1 = 0$

A fair coin has the maximum information, and is hardest to compress
A biased coin has some information, and can be compressed to <1 bit on average
A completely biased coin has no information, and needs only 0 bit



UNIVERSITY of DELAWARE UNIVERSITY of DE

Entropy of words

- Let W be the random variable that denotes whether a word occurs in a document.
  - W=1 if the word occurs
  - W=0 otherwise
- How is the $value\ of\ H(W_{the})$ compared with the value of $H(W_{computer})$?

UNIVERSITY of DELAWARE UNIVERSITY of DE

Mutual Information:

- measuring the correlation of two random variables

UNIVERSITY of DELAWARE UNIVERSITY of DE

Conditional Entropy

- Conditional entropy is used to quantify uncertainties of conditional probabilities.
- H(X|Y): the expected uncertainty of X given that we observe Y.
  - H(X|Y)=0, if X is completely determined by Y.
  - H(X|Y)=H(X), if X and Y are independent.
- Mathematically, we have

$$H(Y \mid X) = \sum_{x \in \Omega_X} p(x) H(Y \mid X = x)$$
$$= -\sum_{x \in \Omega_X} p(x) \sum_{y \in \Omega_Y} p(y \mid x) \log p(y \mid x)$$
$$= -\sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p(x, y) \log p(y \mid x) \ = -E(\log p(Y \mid X))$$

UNIVERSITY of DELAWARE UNIVERSITY of DE

Mutual Information

- It is defined as the reduction of entropy X due to knowledge Y, i.e, I(X;Y).

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

- It can measure the correlation of two random variables.
- Properties:
  - I(X;Y)≥0;
  - I(X;Y)=I(Y;X);
  - I(X;Y)=0 iff X & Y are independent

10

UNIVERSITY of DELAWARE UNIVERSITY of DE

Kullback-Leibler (KL) divergence

- comparing two distributions

UNIVERSITY of DELAWARE UNIVERSITY of DE

Cross Entropy H(p,q)

Assume X has the distribution p, but we encode X with a code optimized for a wrong distribution q, what it the expected number of bits?

$$H(p,q) = E_p[-\log q(x)] = -\sum_{x \in \Omega} p(x) \log q(x)$$

Intuitively, H(p,q) ≥ H(p).

12

Kullback-Leibler (KL) Divergence

Assume X has the distribution p, but we encode it with a code optimized for a wrong distribution q. How many bits would we waste?

$$D(p\,\|\,q) = H(p,q) - H(p) = \sum_{x\in\Omega} p(x)\log\frac{p(x)}{q(x)}$$

Properties:

- *D(p||q)≥0*
- *D(p||q)≠D(q||p)*
- *D(p||q)=0 iff p=q*

⇒ **KL-divergence is often used to measure the distance between two distributions**

13