

Introduction to Introduction to Natural Language Processing / Úvod do zpracování přirozeného jazyka

Lekce / Class1-2

NPFL124

LS 2020/21

Part 1

prof. RNDr. Jan Hajič, Dr. / doc. RNDr. Pavel Pecina, Ph.D.

ÚFAL MFF UK

`{hajic,pecina}@ufal.mff.cuni.cz`

`http://ufal.mff.cuni.cz/jan-hajic`

Intro to NLP

- Instructor: Jan Hajič, Pavel Pecina
 - ÚFAL MFF UK, office: 420 / 422 MS
 - Hours: J. Hajic: Mon 10:00-11:00
 - preferred contact: {hajic,pecina}@ufal.mff.cuni.cz
- Room & time:
 - lecture: online, Thursday 12:20-13:50
 - seminar [cvičení] follows (Pavel Pecina, Zdeněk Žabokrtský, ...)
 - Other info: pls see at the seminar

Textbooks you need

- Manning, C. D., Schütze, H.:
 - *Foundations of Statistical Natural Language Processing*. The MIT Press. 1999. ISBN 0-262-13360-1. [**available at least at MFF / Computer Science School library, Malostranske nam. 25, 11800 Prague 1**]
- Jurafsky, D., Martin, J.H.:
 - *Speech and Language Processing*. Prentice-Hall. 2000. ISBN 0-13-095069-6 and newer editions. [**recommended**].
- Cover, T. M., Thomas, J. A.:
 - *Elements of Information Theory*. Wiley. 1991. ISBN 0-471-06259-6.
- Jelinek, F.:
 - *Statistical Methods for Speech Recognition*. The MIT Press. 1998. ISBN 0-262-10066-5

Other reading

- Journals:
 - Computational Linguistics
 - Transactions on Computational Linguistics
- Proceedings of major conferences:
 - ACL (Assoc. of Computational Linguistics)
 - EACL (European Chapter of ACL)
 - EMNLP (Empirical Methods in NLP)
 - CoNLL (Natural Language Learning in CL)
 - IJCNLP (Asian chapter of ACL)
 - COLING (Intl. Committee of Computational Linguistics)

Course segments (first three lectures)

- Intro & Probability & Information Theory
 - The very basics: definitions, formulas, examples.
- Language Modeling
 - n-gram models, parameter estimation
 - smoothing (EM algorithm)

Probability

Experiments & Sample Spaces

- Experiment, process, test, ...
- Set of possible basic outcomes: sample space Ω
 - coin toss ($\Omega = \{\text{head,tail}\}$), die ($\Omega = \{1..6\}$)
 - yes/no opinion poll, quality test (bad/good) ($\Omega = \{0,1\}$)
 - lottery ($|\Omega| \cong 10^7 .. 10^{12}$)
 - # of traffic accidents somewhere per year ($\Omega = \mathbb{N}$)
 - spelling errors ($\Omega = Z^*$), where Z is an alphabet, and Z^* is a set of possible strings over such and alphabet
 - missing word ($|\Omega| \cong \text{vocabulary size}$)

Events

- Event A is a set of basic outcomes
- Usually $A \subset \Omega$, and all $A \in 2^\Omega$ (the event space)
 - Ω is then the certain event, \emptyset is the impossible event
- Example:
 - experiment: three times coin toss
 - $\Omega = \{\mathbf{HHH}, \mathbf{HHT}, \mathbf{HTH}, \mathbf{HTT}, \mathbf{TTH}, \mathbf{THT}, \mathbf{TTH}, \mathbf{TTT}\}$
 - count cases with exactly two tails: then
 - $A = \{\mathbf{HTT}, \mathbf{THT}, \mathbf{TTH}\}$
 - all heads:
 - $A = \{\mathbf{HHH}\}$

Probability

- Repeat experiment many times, record how many times a given event A occurred (“count” c_1).
- Do this whole series many times; remember all c_i s.
- Observation: if repeated really many times, the ratios of c_i/T_i (where T_i is the number of experiments run in the i -th series) are close to some (unknown but) **constant** value.
- Call this constant a **probability of A** . Notation: **$p(A)$**

Estimating probability

- Remember: ... close to an *unknown* constant.
- We can only estimate it:
 - from a single series (typical case, as mostly the outcome of a series is given to us and we cannot repeat the experiment), set

$$p(A) = c_1/T_1.$$

- otherwise, take the weighted average of all c_i/T_i (or, if the data allows, simply look at the set of series as if it is a single long series).
- This is the **best** estimate.

Example

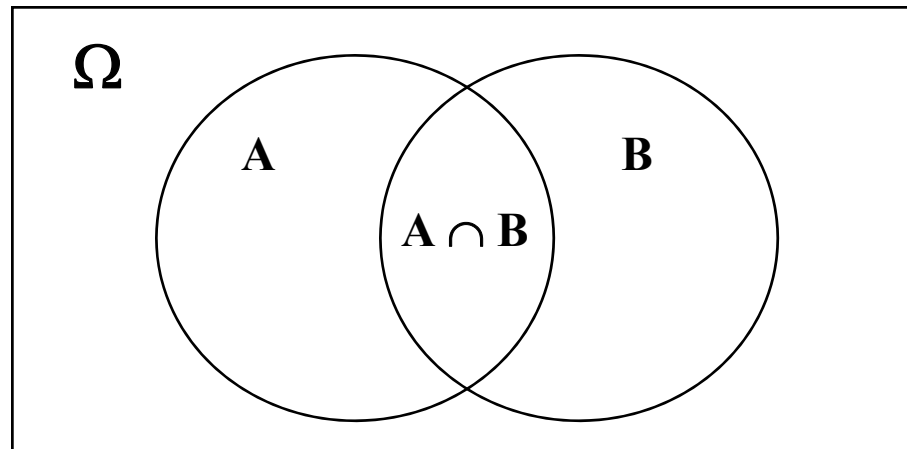
- Recall our example:
 - experiment: three times coin toss
 - $\Omega = \{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}, \text{TTH}, \text{THT}, \text{TTH}, \text{TTT}\}$
 - count cases with exactly two tails: $A = \{\text{HTT}, \text{THT}, \text{TTH}\}$
- Run an experiment 1000 times (i.e. 3000 tosses)
- Counted: 386 cases with two tails (HTT, THT, or TTH)
- estimate: $p(A) = 386 / 1000 = .386$
- Run again: 373, 399, 382, 355, 372, 406, 359
 - $p(A) = .379$ (weighted average) or simply $3032 / 8000$
- *Uniform* distribution assumption: $p(A) = 3/8 = .375$

Basic Properties

- Basic properties:
 - $p: 2^\Omega \rightarrow [0,1]$
 - $p(\Omega) = 1$
 - Disjoint events: $p(\cup A_i) = \sum_i p(A_i)$
- [NB: axiomatic definition of probability: take the above three conditions as axioms]
- Immediate consequences:
 - $p(\emptyset) = 0$, $p(\bar{A}) = 1 - p(A)$, $A \subseteq B \Rightarrow p(A) \leq p(B)$
 - $\sum_{a \in \Omega} p(a) = 1$

Joint and Conditional Probability

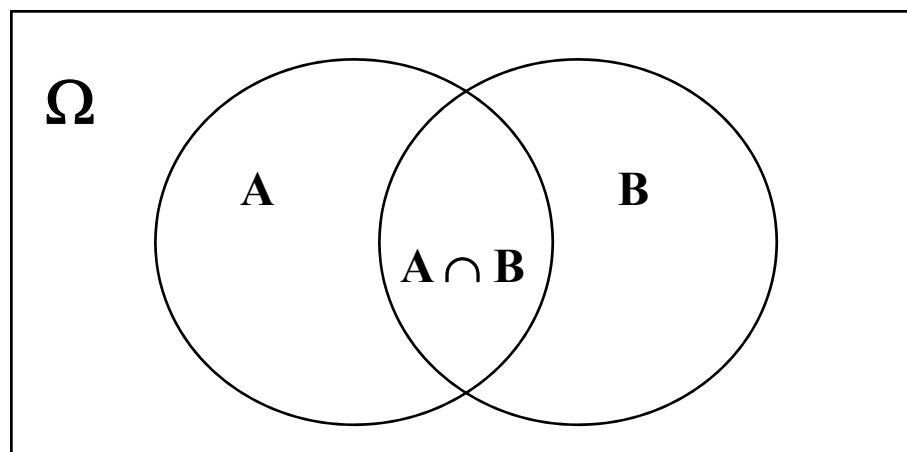
- $p(A,B) = p(A \cap B)$
- $p(A|B) = p(A,B) / p(B)$
 - Estimating form counts:
 - $p(A|B) = p(A,B) / p(B) = (c(A \cap B) / T) / (c(B) / T) = c(A \cap B) / c(B)$



Bayes Rule

- $p(A,B) = p(B,A)$ since $p(A \cap B) = p(B \cap A)$
 - therefore: $p(A|B) p(B) = p(B|A) p(A)$, and therefore

$$p(A|B) = p(B|A) p(A) / p(B)$$



Independence

- Can we compute $p(A,B)$ from $p(A)$ and $p(B)$?
- Recall from previous foil:

$$p(A|B) = p(B|A) p(A) / p(B)$$

$$p(A|B) p(B) = p(B|A) p(A)$$

$$p(A,B) = p(B|A) p(A)$$

... we're almost there: how $p(B|A)$ relates to $p(B)$?

– $p(B|A) = P(B)$ iff A and B are **independent**

- Example: two coin tosses, weather today and weather on March 4th 1789;
- Any two events for which $p(B|A) = P(B)$!

Chain Rule

$$p(A_1, A_2, A_3, A_4, \dots, A_n) =$$



$$p(A_1|A_2, A_3, A_4, \dots, A_n) \times p(A_2|A_3, A_4, \dots, A_n) \times \\ \times p(A_3|A_4, \dots, A_n) \times \dots p(A_{n-1}|A_n) \times p(A_n)$$

- this is a direct consequence of the Bayes rule.

The Golden Rule (of Classic Statistical NLP)

- Interested in an event A given B (when it is not easy or practical or desirable to estimate $p(A|B)$):
- take Bayes rule, max over all As:
- $\operatorname{argmax}_A p(A|B) = \operatorname{argmax}_A p(B|A) \cdot p(A) / p(B) =$

$$\operatorname{argmax}_A p(B|A) p(A) \quad !$$

- ... as $p(B)$ is constant when changing As

Random Variable

- is a function $X: \Omega \rightarrow Q$
 - in general: $Q = \mathbb{R}^n$, typically \mathbb{R}
 - easier to handle real numbers than real-world events
- random variable is *discrete* if Q is countable (i.e. also if finite)
- Example: *die*: natural “numbering” $[1,6]$, *coin*: $\{0,1\}$
- Probability distribution:
 - $p_X(x) = p(X=x) =_{\text{df}} p(A_x)$ where $A_x = \{a \in \Omega : X(a) = x\}$
 - often just $p(x)$ if it is clear from context what X is

Expectation

Joint and Conditional Distributions

- is a mean of a random variable (weighted average)
 - $E(X) = \sum_{x \in X(\Omega)} x \cdot p_X(x)$
- Example: one six-sided die: 3.5, two dice (sum) 7
- Joint and Conditional distribution rules:
 - analogous to probability of events
- Bayes: $p_{X|Y}(x,y) \stackrel{\text{notation}}{=} p_{XY}(x|y) \stackrel{\text{even simpler notation}}{=} p(x|y) = p(y|x) \cdot p(x) / p(y)$
- Chain rule: $p(w,x,y,z) = p(z) \cdot p(y|z) \cdot p(x|y,z) \cdot p(w|x,y,z)$

Essential Information Theory

The Notion of Entropy

- Entropy ~ “chaos”, fuzziness, opposite of order, ...
 - you know it:
 - **it is much easier to create “mess” than to tidy things up...**
- Comes from physics:
 - Entropy does not go down unless energy is applied
- Measure of *uncertainty*:
 - if low... low uncertainty; the higher the entropy, the higher uncertainty, but the higher “surprise” (information) we can get out of an experiment

The Formula

- Let $p_X(x)$ be a distribution of random variable X
- Basic outcomes (alphabet) Ω

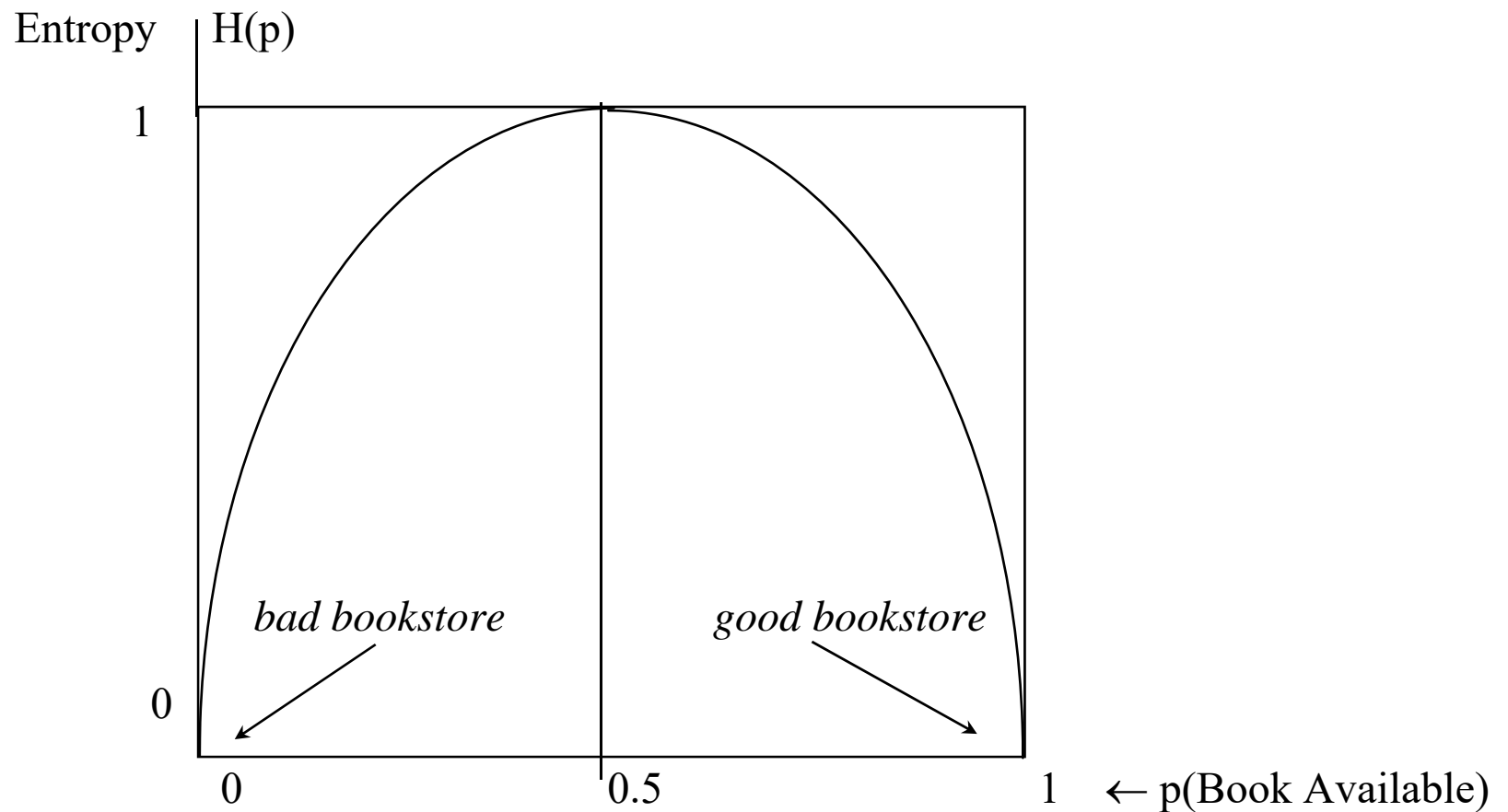
$$H(X) = - \sum_{x \in \Omega} p(x) \log_2 p(x) \quad !$$

- Unit: bits (\log_{10} : nats)
- Notation: $H(X) = H_p(X) = H(p) = H_X(p) = H(p_X)$

Using the Formula: Example

- Toss a fair coin: $\Omega = \{\text{head}, \text{tail}\}$
 - $p(\text{head}) = .5, p(\text{tail}) = .5$
 - $\mathbf{H(p)} = -0.5 \log_2(0.5) + (-0.5 \log_2(0.5)) = 2 \times ((-0.5) \times (-1)) = 2 \times 0.5 = \mathbf{1}$
- Take fair, 32-sided die: $p(x) = 1 / 32$ for every side x
 - $\mathbf{H(p)} = -\sum_{i=1..32} p(x_i) \log_2 p(x_i) = -32 (p(x_1) \log_2 p(x_1))$
(since for all i $p(x_i) = p(x_1) = 1/32$)
 $= -32 \times ((1/32) \times (-5)) = \mathbf{5}$ (now you see why it's called *bits*?)
- Unfair coin:
 - $p(\text{head}) = .2 \dots \mathbf{H(p)} = .722$; $p(\text{head}) = .01 \dots \mathbf{H(p)} = .081$

Example: Book Availability



The Limits

- When $H(p) = 0$?
 - if a result of an experiment is *known* ahead of time:
 - necessarily:

$$\exists \mathbf{x} \in \Omega; \mathbf{p}(\mathbf{x}) = 1 \ \& \ \forall \mathbf{y} \in \Omega; \mathbf{y} \neq \mathbf{x} \Rightarrow \mathbf{p}(\mathbf{y}) = 0$$

- Upper bound?
 - none in general
 - for $|\Omega| = n$: $H(p) \leq \log_2 n$
 - **nothing can be more uncertain than the uniform distribution**

Perplexity: motivation

- Recall:
 - 2 equiprobable outcomes: $H(p) = 1$ bit
 - 32 equiprobable outcomes: $H(p) = 5$ bits
 - 4.3 billion equiprobable outcomes: $H(p) \approx 32$ bits
- What if the outcomes are not equiprobable?
 - 32 outcomes, 2 equiprobable at .5, rest impossible:
 - **$H(p) = 1$ bit**
 - Any measure for comparing the entropy (i.e. uncertainty/difficulty of prediction) (also) for random variables with different number of outcomes?

Perplexity

- Perplexity:
 - $G(p) = 2^{H(p)}$
- ... so we are back at 32 (for 32 eqp. outcomes), 2 for fair coins, etc.
- it is easier to imagine:
 - NLP example: vocabulary size of a vocabulary with uniform distribution, which is equally hard to predict
- the “wilder” (biased) distribution, the better:
 - lower entropy, lower perplexity

Joint Entropy and Conditional Entropy

- Two random variables: X (space Ω), Y (Ψ)
- Joint entropy:
 - no big deal: ((X, Y) considered a single event):

$$H(X, Y) = - \sum_{x \in \Omega} \sum_{y \in \Psi} p(x, y) \log_2 p(x, y)$$

- Conditional entropy:

$$H(Y|X) = - \sum_{x \in \Omega} \sum_{y \in \Psi} \underline{p(x, y)} \log_2 p(y|x)$$

recall that $H(X) = E(\log_2(1/p_X(x)))$

(weighted average: weights are not conditional)

Conditional Entropy (Using the Calculus)

- other definition:

$$H(Y|X) = \sum_{x \in \Omega} p(x) H(Y|X=x) =$$

for $H(Y|X=x)$, we can use the
single-variable definition ($x \sim \text{constant}$)

$$= \sum_{x \in \Omega} p(x) \left(- \sum_{y \in \Psi} p(y|x) \log_2 p(y|x) \right) =$$

$$= - \sum_{x \in \Omega} \sum_{y \in \Psi} p(y|x) p(x) \log_2 p(y|x) =$$

$$= - \sum_{x \in \Omega} \sum_{y \in \Psi} p(x,y) \log_2 p(y|x)$$

Properties of Entropy I

- Entropy is non-negative:
 - $H(X) \geq 0$
 - proof: (recall: $H(X) = - \sum_{x \in \Omega} p(x) \log_2 p(x)$)
 - **$\log(p(x))$ is negative or zero for $x \leq 1$,**
 - **$p(x)$ is non-negative; their product $p(x)\log(p(x))$ is thus negative;**
 - **sum of negative numbers is negative;**
 - **and $-f$ is positive for negative f**
- Chain rule:
 - $H(X, Y) = H(Y|X) + H(X)$, as well as
 - $H(X, Y) = H(X|Y) + H(Y)$ (since $H(Y, X) = H(X, Y)$)

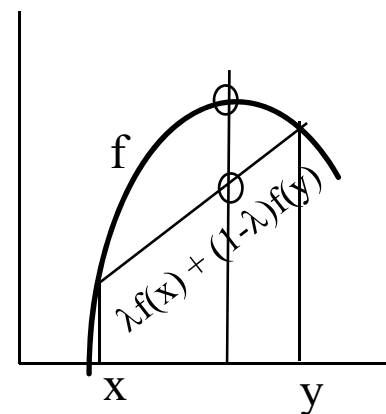
Properties of Entropy II

- Conditional Entropy is better (than unconditional):
 - $H(Y|X) \leq H(Y)$
- $H(X, Y) \leq H(X) + H(Y)$ (follows from the previous (in)equalities)
 - **equality iff X, Y independent**
 - **[recall: X, Y independent iff $p(X, Y) = p(X)p(Y)$]**
- $H(p)$ is concave (remember the book availability graph?)
 - concave function f over an interval (a, b) :

$$\forall x, y \in (a, b), \forall \lambda \in [0, 1]:$$

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y)$$

- **function f is convex if $-f$ is concave**



“Coding” Interpretation of Entropy

- The least (average) number of bits needed to encode a message (string, sequence, series,...) (each element having being a result of a random process with some distribution p): $= H(p)$
- Remember various compressing algorithms?
 - they do well on data with repeating (= easily predictable = low entropy) patterns
 - their results though have high entropy \Rightarrow compressing compressed data does nothing

Coding: Example

- How many bits do we need for ISO Latin 1?
 - \Rightarrow the trivial answer: 8
- Experience: some chars are more common, some (very) rare:
 - ...so what if we use more bits for the rare, and less bits for the frequent? [be careful: want to decode (easily)!]
 - suppose: $p('a') = 0.3$, $p('b') = 0.3$, $p('c') = 0.3$, the rest: $p(x) \cong .0004$
 - code: 'a' ~ 00, 'b' ~ 01, 'c' ~ 10, rest: $11b_1b_2b_3b_4b_5b_6b_7b_8$
 - code acbbécbaac: 001001011110000111111001000010
 a c b b é c b a a c
 - number of bits used: 28 (vs. 80 using “naive” coding)
- code length $\sim 1 / \text{probability}$; conditional prob OK!

Kullback-Leibler Distance (Relative Entropy)

- Remember:
 - long series of experiments... c_i/T_i oscillates around some number... we can only estimate it... to get a distribution \underline{q} .
- So we get a distribution \underline{q} ; (sample space Ω , r.v. X)
the true distribution is, however, \underline{p} . (same Ω , X)
 \Rightarrow how big error are we making?
- $D(\underline{p}||\underline{q})$ (the Kullback-Leibler distance):

$$D(\underline{p}||\underline{q}) = \sum_{x \in \Omega} \underline{p}(x) \log_2 (p(x)/q(x)) = E_{\underline{p}} \log_2 (p(x)/q(x))$$

Comments on Relative Entropy

- Conventions:
 - $0 \log 0 = 0$
 - $p \log (p/0) = \infty$ (for $p > 0$)
- Distance? (less “misleading”: Divergence)
 - not quite:
 - **not symmetric: $D(p||q) \neq D(q||p)$**
 - **does not satisfy the triangle inequality**
 - but useful to look at it that way
- $H(p) + D(p||q)$: bits needed for encoding p if q is used

Mutual Information (MI) in terms of relative entropy

- Random variables X, Y ; $p_{X \cap Y}(x,y)$, $p_X(x)$, $p_Y(y)$
- Mutual information (between two random variables X, Y):

$$I(X, Y) = D(p(x,y) \parallel p(x)p(y))$$

- $I(X, Y)$ measures how much (our knowledge of) Y contributes (on average) to easing the prediction of X
- or, how $p(x,y)$ deviates from (independent) $p(x)p(y)$

Mutual Information: the Formula

- Rewrite the definition: [recall: $D(r||s) = \sum_{v \in \Omega} r(v) \log_2 (r(v)/s(v))$;
substitute $r(v) = p(x,y)$, $s(v) = p(x)p(y)$; $\langle v \rangle \sim \langle x,y \rangle$]

$$\begin{aligned} I(X,Y) &= D(p(x,y) || p(x)p(y)) = \\ &= \sum_{x \in \Omega} \sum_{y \in \Psi} p(x,y) \log_2 (p(x,y)/p(x)p(y)) \end{aligned} \quad !$$

- Measured in bits (what else? :-)

From Mutual Information to Entropy

- by how many bits the knowledge of Y lowers the entropy $H(X)$:

$$\begin{aligned}
 I(X, Y) &= \sum_{x \in \Omega} \sum_{y \in \Psi} p(x, y) \log_2 \left(\frac{p(x, y)}{p(y)p(x)} \right) = \\
 &\quad \dots \text{use } p(x, y)/p(y) = p(x|y) \\
 &= \sum_{x \in \Omega} \sum_{y \in \Psi} p(x, y) \log_2 \left(\frac{p(x|y)}{p(x)} \right) = \\
 &\quad \dots \text{use } \log(a/b) = \log a - \log b \text{ (} a \sim p(x|y), b \sim p(x)\text{), distribute sums} \\
 &= \sum_{x \in \Omega} \sum_{y \in \Psi} p(x, y) \log_2 p(x|y) - \sum_{x \in \Omega} \sum_{y \in \Psi} p(x, y) \log_2 p(x) = \\
 &\quad \dots \text{use def. of } H(X|Y) \text{ (left term), and } \sum_{y \in \Psi} p(x, y) = p(x) \text{ (right term)} \\
 &= \frac{-H(X|Y)}{1} + \left(- \sum_{x \in \Omega} p(x) \log_2 p(x) \right) = \\
 &\quad \dots \text{use def. of } H(X) \text{ (right term), swap terms} \\
 &= H(X) - H(X|Y) \quad \dots \text{by symmetry, } = H(Y) - H(Y|X)
 \end{aligned}$$

Properties of MI vs. Entropy

- $I(X, Y) = H(X) - \underline{H(X|Y)}$ = number of bits the knowledge of Y lowers the entropy of X
 $= H(Y) - H(Y|X)$ (prev. foil, symmetry)

Recall: $H(X, Y) = H(X|Y) + H(Y) \Rightarrow -H(X|Y) = H(Y) - H(X, Y) \Rightarrow$

- $I(X, Y) = H(X) + \underline{H(Y) - H(X, Y)}$
- $I(X, X) = H(X)$ (since $H(X|X) = 0$)
- $I(X, Y) = I(Y, X)$ (just for completeness)
- $I(X, Y) \geq 0$... let's prove that now (as promised).

Other (In)Equalities and Facts

- Log sum inequality: for $r_i, s_i \geq 0$

$$\sum_{i=1..n} (r_i \log(r_i/s_i)) \leq \left(\sum_{i=1..n} r_i \right) \log\left(\sum_{i=1..n} r_i / \sum_{i=1..n} s_i \right)$$

- $D(p||q)$ is convex [in p, q] (\Leftarrow log sum inequality)
- $H(p_X) \leq \log_2|\Omega|$, where Ω is the sample space of p_X

Proof: uniform $u(x)$, same sample space Ω : $\sum p(x) \log u(x) = -\log_2|\Omega|$;

$$\log_2|\Omega| - H(X) = -\sum p(x) \log u(x) + \sum p(x) \log p(x) = D(p||u) \geq 0$$

- $H(p)$ is concave [in p]:

Proof: from $H(X) = \log_2|\Omega| - D(p||u)$, $D(p||u)$ convex $\Rightarrow H(x)$ concave

Cross-Entropy

- Typical case: we've got series of observations

$$T = \{t_1, t_2, t_3, t_4, \dots, t_n\} \text{ (numbers, words, ...; } t_i \in \Omega \text{);}$$

estimate (simple):

$$\forall y \in \Omega: \tilde{p}(y) = c(y) / |T|, \text{ def. } c(y) = |\{t \in T; t = y\}|$$

- ...but the true p is unknown; every sample is too small!
- Natural question: how well do we do using \tilde{p} [instead of p]?
- Idea: simulate actual p by using a different T'
(or rather: by using different observation we simulate the insufficiency of T vs. some other data (“random” difference))

Cross Entropy: The Formula

- $H_{p'}(\tilde{p}) = H(p') + D(p' \| \tilde{p})$

$$H_{p'}(\tilde{p}) = - \sum_{x \in \Omega} p'(x) \log_2 \tilde{p}(x) \quad !$$

- p' is certainly not the true p , but we can consider it the “real world” distribution against which we test \tilde{p}
- note on notation (confusing...): $p/p' \leftrightarrow \tilde{p}$, also $H_{T'}(p)$
- (Cross)Perplexity: $G_{p'}(p) = G_{T'}(p) = 2^{H_{p'}(\tilde{p})}$

Conditional Cross Entropy

- So far: “unconditional” distribution(s) $p(x)$, $p'(x)$...
- In practice: virtually always conditioning on context
- Interested in: sample space Ψ , r.v. Y , $y \in \Psi$;
context: sample space Ω , r.v. X , $x \in \Omega$;;
“our” distribution $p(y|x)$, test against $p'(y,x)$,
which is taken from some independent data:

$$H_{p'}(p) = - \sum_{y \in \Psi, x \in \Omega} p'(y,x) \log_2 p(y|x)$$

Sample Space vs. Data

- In practice, it is often inconvenient to sum over the sample space(s) Ψ, Ω (especially for cross entropy!)
- Use the following formula:

$$H_{p'}(p) = - \sum_{y \in \Psi, x \in \Omega} p'(y,x) \log_2 p(y|x) = - 1/|T'| \sum_{i=1..|T'|} \log_2 p(y_i|x_i) \quad !$$

- This is in fact the normalized log probability of the “test” data:

$$H_{p'}(p) = - 1/|T'| \log_2 \prod_{i=1..|T'|} p(y_i|x_i)$$

Cross Entropy: Some Observations

- $H(p)$?? $<, =, >$?? $H_{p'}(p)$: ALL!

- Previous example:

$[p(a) = .25, p(b) = .5, p(\alpha) = 1/64 \text{ for } \alpha \in \{c..r\}, = 0 \text{ for the rest: } s,t,u,v,w,x,y,z]$

$$H(p) = 2.5 \text{ bits} = H(p') \text{ (barb)}$$

- Other data: probable: $(1/8)(6+6+6+1+2+1+6+6) = 4.25$

$$H(p) < 4.25 \text{ bits} = H(p') \text{ (probable)}$$

- And finally: abba: $(1/4)(2+1+1+2) = 1.5$

$$H(p) > 1.5 \text{ bits} = H(p') \text{ (abba)}$$

- But what about: baby $-p'('y')\log_2 p('y') = -.25\log_2 0 = \infty$ (??)

Cross Entropy: Usage

- Comparing data??
 - NO! (we believe that we test on real data!)
- Rather: comparing distributions (vs. real data)
- Have (got) 2 distributions: p and q (on some Ω, X)
 - which is better?
 - better: has lower cross-entropy (perplexity) on real data S
- “Real” data: S
- $H_S(p) = -1/|S| \sum_{i=1..|S|} \log_2 p(y_i|x_i)$?? $H_S(q) = -1/|S| \sum_{i=1..|S|} \log_2 q(y_i|x_i)$

Comparing Distributions

Test data S: probable

- $p(\cdot)$ from prev. example:

$$H_S(p) = 4.25$$

$p(a) = .25, p(b) = .5, p(\alpha) = 1/64$ for $\alpha \in \{c..r\}, = 0$ for the rest: s,t,u,v,w,x,y,z

- $q(\cdot|\cdot)$ (conditional; defined by a table):

$q(\cdot \cdot) \rightarrow$ ↓	a	b	e	l	o	p	r	other
a	0	.5	0	0	0	.125	0	0
b	1	0	0	0	1	.125	0	0
e	0	0	0	1	0	.125	0	0
l	0	.5	0	0	0	.125	0	0
o	0	0	0	0	0	.125	1	0
p	0	0	0	0	0	.125	0	1
r	0	0	0	0	0	.125	0	0
other	0	0	1	0	0	.125	0	0

ex.: $q(o|r) = 1$

$q(r|p) = .125$

$$(1/8) (\log(p|oth.) + \log(r|p) + \log(o|r) + \log(b|o) + \log(a|b) + \log(b|a) + \log(l|b) + \log(e|l))$$

$$(1/8) (0 + 3 + 0 + 0 + 1 + 0 + 1 + 0)$$

$$H_S(q) = .625$$