NPFL147

Statistical Methods in Natural Language Processing

2. Essential information theory

Pavel Pecina, Jindřich Helcl

14 October, 2025

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Update on Homework Assignments

- **Three homework assignments**:
    1. Assigned **Nov 4**, submission **Nov 25, 8pm**
    2. Assigned **Nov 25**, submission **Dec 16, 8pm**
    3. Assigned **Dec 16**, submission **Jan 6, 8pm**
- The assignments will be awarded by 0–100 points each
- Late submissions up to 2 weeks → 50% point reduction
- Submissions received later than 2 weeks → 0 points
- **One two-week no-penalty extension will be granted upon request sent by email before the deadline.**

Passing Requirements

- Completion of both the homework assignments and exam is required
- Students need to earn **at least 50 points for each assignment** (before late submission penalization) and **at least 50 points for the test.**
- The points received for the assignments and test will be available in **SIS**.
- The **final grade** will be based on the average results of the exam test and the three homework assignments, **all four weighted equally**.

Course Segments

1. Introduction, probability, essential information theory
2. Statistical language modelling (n-gram)
3. Statistical properties of words
4. Word embeddings
5. Hidden Markov models, Tagging

Recap from Last Week

Why is NLP difficult?

- many “words”, many “phenomena” → many “rules”
  - **OED: 400k words; Finnish lexicon (of forms): ~2 . 107**
  - sentences, clauses, phrases, constituents, coordination, negation, imperatives/questions, inflections, parts of speech, pronunciation, topic/focus, and much more!
- irregularity (exceptions, exceptions to the exceptions, ...)
  - plural forms
    - **potato → potato es (tomato, hero,...); photo → photo s**
    - and even: **both mango -> mango s or → mango es**
  - Adjective / Noun order
    - **new book, electrical engineering, general regulations, flower garden, garden flower**
    - but **Governor General**

Estimating Probability

- True probability **unknown**
- We can estimate from our observations
  - Either from a single series,
  - .. or take (weighted) average from each series
  - .. or concatenate all series

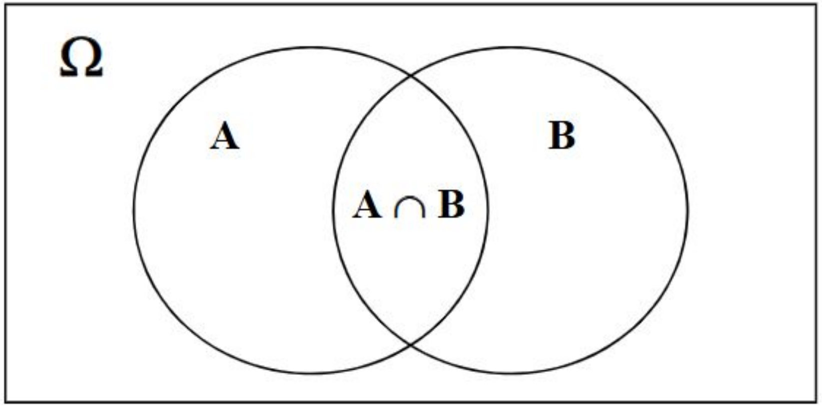
$$p(A) = \frac{1}{N}\sum_{i=1}^{N} c_i/T_i = \frac{\sum_{i=1}^{N} c_i}{\sum_{i=1}^{N} T_i}$$

- aka Maximum Likelihood Estimate

*... this is the **best** estimate*

Bayes Rule

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$



.. follows from symmetry of joint probability.

Chain Rule

Joint and conditional probabilities for **many events**

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

.. useful in NLP where we can approximate some of the terms

The "Golden Rule" of Classic Statistical NLP

P(A|B) in NLP applications

- Speech recognition, machine translation, language modeling
  - B = audio signal, source sentence, previous word
  - A = transcription, target sentence, next word

- Goal is to find A that **maximizes P(A|B)**

$$A^* = \text{argmax}_A \, p(A|B)$$

- When estimating P(A|B) directly is not desirable, use Bayes rule

$$A^* = \text{argmax}_A \, p(B|A)p(A)\cancel{/p(B)}$$

- Ignore the p(B) term which is **constant**

Essential Information Theory

The Notion of Entropy

- Entropy ~ “chaos”, fuzziness, opposite of order, ...
  - you know: **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

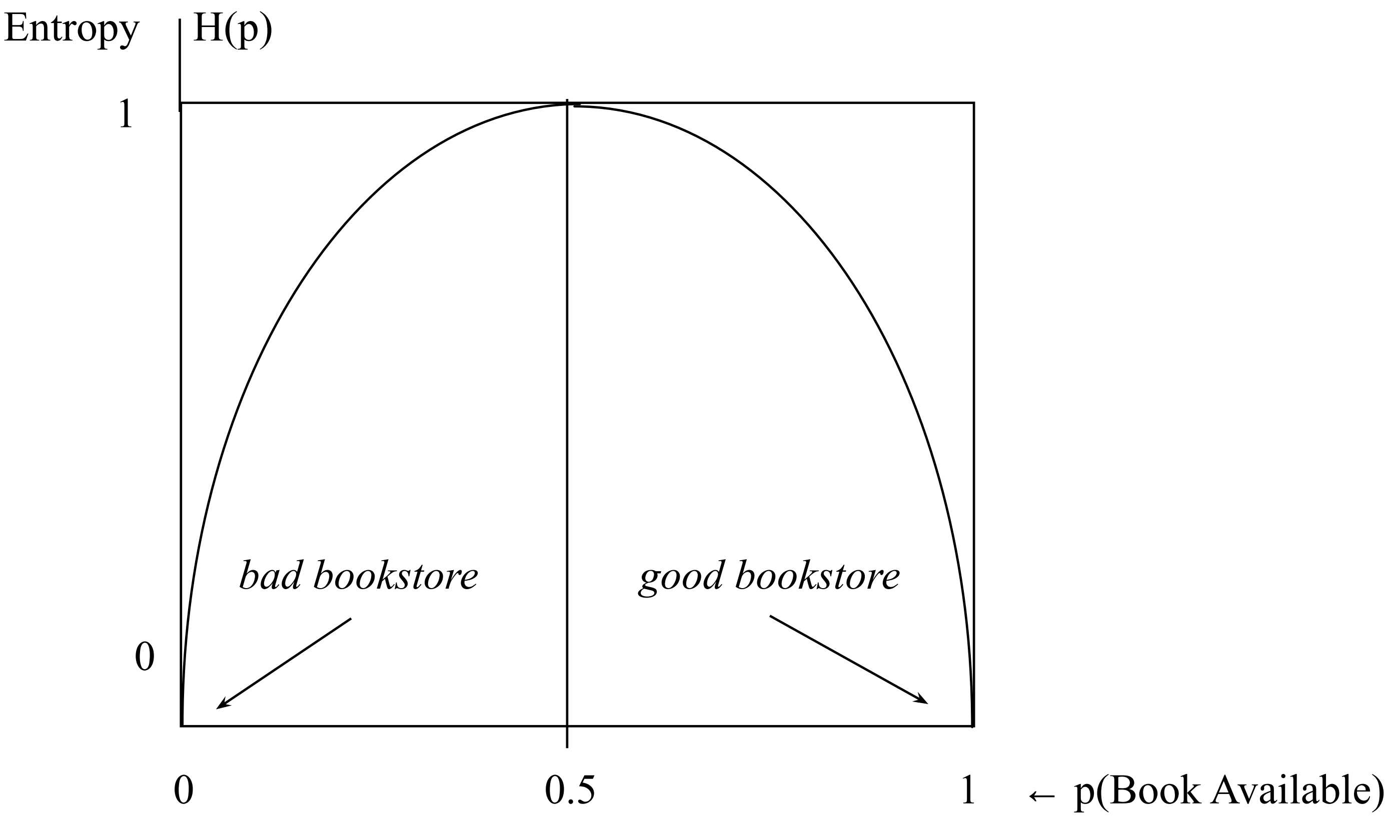
$$H(X) = - \sum_{x \in \Omega} p(x) \log_2 p(x) \quad !$$

- Unit: bits ($\log_e$: nats)
- Notation: $H(X) = H_p(X) = H(p) = H_X(p) = H(p_X)$

Using the Formula: Example

- Toss a fair coin: Ω = {head,tail}
  - $p(\text{head}) = 0.5,\ p(\text{tail}) = 0.5$
  - $H(p) = -\ 0.5 \log_2(0.5) + (-\ 0.5 \log_2(0.5)) = 2 \times ((-0.5) \times (-1)) = 2 \times 0.5 = 1$
- Take fair, 32-sided die: $p(x) = 1/32$ for every side x
  - $H(p) = -\sum_{i = 1..32} p(x_i) \log_2 p(x_i) = -\ 32\ (p(x_1) \log_2 p(x_1))$
    (since for all i: $p(x_i) = p(x_1) = 1/32$)
  - $H(p) = -32 \times ((1/32) \times (-5)) = 5$
    (now you see why it's called bits?)
- Unfair coin:
  - $p(\text{head}) = 0.2\ \ldots\ H(p) = 0.722$
  - $p(\text{head}) = 0.01\ \ldots\ H(p) = 0.081$

Example: Book Availability



The Limits

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

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

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

Entropy and Expectation

- Recall:
  - $E(X) = \sum_{x\in X(\Omega)} p_X(x) \times x$
- Then:

$$E(\log_2(1/p_X(x))) = \sum_{x\in X(\Omega)} p_X(x) \log_2(1/p_X(x)) =$$

$$= - \sum_{x\in X(\Omega)} p_X(x) \log_2 p_X(x) =$$

$$= H(p_X) =_{\text{notation}} H(p)$$

Perplexity: motivation

- Recall:
  - 2 equiprobable outcomes: $H(p) = 1$ bit
  - 32 equiprobable outcomes: $H(p) = 5$ bits
  - 4.3 billion equiprobable outcomes: H(p) ~= 32 bits
- What if the outcomes are not equiprobable?
  - 32 outcomes, 2 equiprobable at 0.5, 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:

$$\mathbf{G(p) = 2^{H(p)}}$$

  - 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", and weights are not conditional)

Conditional Entropy (Using the Calculus)

- Alternative 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 ~ constant)

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

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

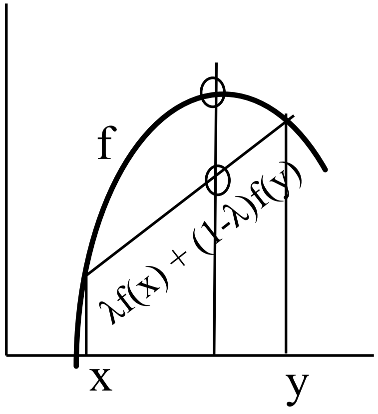
$$= - \sum_{x \in \Omega} \sum_{y \in \Psi} p(x,y) \log_2 p(y|x)$$

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 (the book availability graph?)
    - concave function <u>f</u> 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 <u>f</u> is convex if <u>-f</u> is concave



“Coding” Interpretation of Entropy

- $H(p)$: 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)
- Remember various compression algorithms?
  - they do well on data with repeating (= easily predictable = low entropy) patterns
  - their results though have high entropy ⇒ compressing compressed data does nothing

Coding: Example

- How many bits do we need for ISO Latin 1 character encoding?

  ⇒ 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)≅ .0004
  - **code: 'a' ~ 00, 'b' ~ 01, 'c' ~ 10, rest: 11$b_1b_2b_3b_4b_5b_6b_7b_8$**
  - **code acbbécbaac:** `0010010111000011111001000010`
  - `a c b b é c b a a c`
  - number of bits used: 28 (vs. 80 using "naive" coding)

- code length ~ 1 / probability

Entropy of a Language

- Imagine that we produce the next letter using

$$p(l_{n+1}|l_1,\ldots,l_n)$$

- where $l_1,\ldots,l_n$ is the sequence of **all** the letters which had been uttered so far (i.e. n is really big!); let's call $l_1,\ldots,l_n$ the **history** $h(h_{n+1})$, and all histories H:
- Then compute its entropy:

$$-\sum_{h \in H} \sum_{l \in A} p(l,h) \log_2 p(l|h)$$

- Not very practical, isn't it?

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 q.
- So we get a distribution q; (sample space Ω, r.v. X)
- The true distribution is, however, p (same Ω, X)

$\Rightarrow$ how big error are we making?

- $D(p||q)$ (the Kullback-Leibler distance):

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

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):

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

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

Mutual Information: the Formula

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

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

- Measured in bits (what else? :-)

From Mutual Information to Entropy

By how many bits the knowledge of Y ***<u>lowers</u>*** the entropy H(X):

$$I(X,Y) = \sum_{x \in \Omega} \sum_{y \in \Psi} p(x,y) \log_2 (\underline{p(x,y)/p(y)}p(x)) =$$

...use $p(x,y)/p(y) = p(x|y)$

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

...use $\log(a/b) = \log a - \log b$ $(a \sim p(x|y),\ b \sim p(x))$, distribute sums

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

...use def. of $H(X|Y)$ (left term), *and $\sum_{y \in \Psi} p(x,y) = p(x)$* (right term)

$$= - H(X|Y) + (- \sum_{x \in \Omega} p(x)\log_2 p(x)) =$$

...use def. of $H(X)$ (right term), *swap terms*

$$= H(X) - H(X|Y) \qquad \text{...by symmetry, } = H(Y) - H(Y|X)$$

Properties of MI vs. Entropy

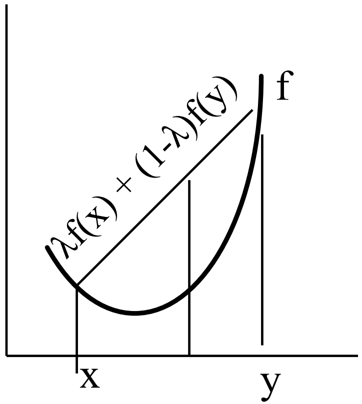
- $I(X,Y) = H(X) \underline{- H(X|Y)}$ = number of bits the knowledge of Y lowers the entropy of X

  $= H(Y) - H(Y|X)$ (symmetry, see prev. slide)

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).

Jensen's Inequality (JI)



- Recall: f is convex on interval (a,b) iff

  $\forall x,y \in (a,b),\ \forall \lambda \in [0,1]: f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$
- JI: for distr. $p(x)$, r.v. X on Ω, and convex f:

$$f\left(\sum_{x\in\Omega} p(x)\, x\right) \leq \sum_{x\in\Omega} p(x)\, f(x)$$

- Proof (idea): by induction on the number of basic outcomes;
- start with $|\Omega| = 2$ by:
  - $p(x_1)f(x_1) + p(x_2)f(x_2) \geq f(p(x_1)x_1 + p(x_2)x_2)$ ($\Leftarrow$ def. of convexity)
  - for the induction step ($|\Omega| = k \rightarrow k+1$), just use the induction hypothesis and def. of convexity (again).

Information Inequality

$$D(p||q) \geq 0$$

Proof:

$$0 = - \log 1 = - \log \sum_{x\in\Omega} q(x) = - \log \sum_{x\in\Omega} p(x)(q(x)/p(x) \leq$$

...apply Jensen's inequality here (- log is convex)...

$$\leq \sum_{x\in\Omega} p(x)(-\log(q(x)/p(x))) = \sum_{x\in\Omega} p(x)\log(p(x)/q(x)) = D(p||q)$$

Other (In)Equalities and Facts

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

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

- $D(p||q)$ 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 $H(X) = \log2|\Omega| - D(p||u)$, $D(p||u)$ convex $\Rightarrow H(x)$ concave

Cross Entropy

- Typical case: we've got series of observations

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

- estimate (simple):

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

- ... but the true p is unknown; every sample T is too small!
- 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

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

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

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

$$G_{p'}(\tilde{p}) = G_{T'}(\tilde{p}) = 2^{Hp'(\tilde{p})}$$

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:

$$\boxed{\begin{aligned} H_{p'}(p) = \; & - \textstyle\sum_{y \in \Psi,\, x \in \Omega} p'(y,x) \log_2 p(y|x) = \\ & - 1/|T'| \textstyle\sum_{i = 1 \ldots |T'|} \log_2 p(y_i|x_i) \end{aligned}}$$

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

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

Computation Example

- $\Omega = \{a,b,..,z\}$, prob. distr. (assumed/estimated from data):
  $p(a)=0.25$, $p(b)=0.5$, $p(\alpha)=1/64$ for $\alpha \in \{c..r\}$, $= 0$ for the rest: s,t,u,v,w,x,y,z
- Data (test): <u>barb</u> $p'(a) = p'(r) = 0.25$, $p'(b) = 0.5$
- Sum over Ω:

α	a	b	c	d	e	f	g	...	p	q	r	s	t	...	z		
$-p'(\alpha)\log_2 p(\alpha)$	0.5	+0.5	+0	+0	+0	+0	+0	+0	+0	+0	+1.5	+0	+0	+0	+0	+0	= <u>2.5</u>

- Sum over data:

$i\ /\ s_i$	1/b	2/a	3/r	4/b		$1/\lvert T'\rvert$
$-\log_2 p(s_i)$	1	+ 2	+ 6	+ 1	= 10	$(1/4) \times 10 =$ <u>2.5</u>

Cross Entropy: Some Observations

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

  $p(a)=0.25$, $p(b)=0.5$, $p(\alpha)=1/64$ for $\alpha \in \{c..r\}$, $= 0$ for the rest: s,t,u,v,w,x,y,z

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

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

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

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

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

- But what about: baby $-p'(y)\log_2 p(y) = -0.25 \log_2 0 = \infty$ (??)

Cross Entropy: Usage

- Comparing data??
  - **NO!** (we believe that we test on **<u>real</u>** data!)
- Rather: <u>comparing distributions</u> (**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

$$H_S(p) = -\,1/|S| \sum_{i=1..|S|} \log_2 p(y_i|x_i) \quad ?? \quad H_S(q) = -\,1/|S| \sum_{i=1..|S|} \log_2 q(y_i|x_i)$$

Comparing Distributions:

- Test data S: probable, p(.) from prev. example: $\boxed{H_S(p) = 4.25}$

  $p(a)=0.25$, $p(b)=0.5$, $p(\alpha)=1/64$ for $\alpha \in \{c..r\}$, $= 0$ for the rest: s,t,u,v,w,x,y,z

- q(.|.) conditional, defined by a table:

q(.|.)→ ↓	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.: $q(o|r) = 1$

ex.: $q(r|p) = 0.125$

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

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

$\boxed{H_S(q) = 0.625}$

Moodle Quiz

Moodle Quiz

https://dl1.cuni.cz/course/view.php?id=18547