manipal PROlearn

Foundations for NLP

By
Rahim Baig

Agenda

- Entropy
  - Entropy
  - Joint entropy
  - Conditional entropy
  - Mutual Information
- KL divergence
- Cross entropy
- Bayes Rule
- Expectation Maximization (tomorrow)
- Noisy channel model

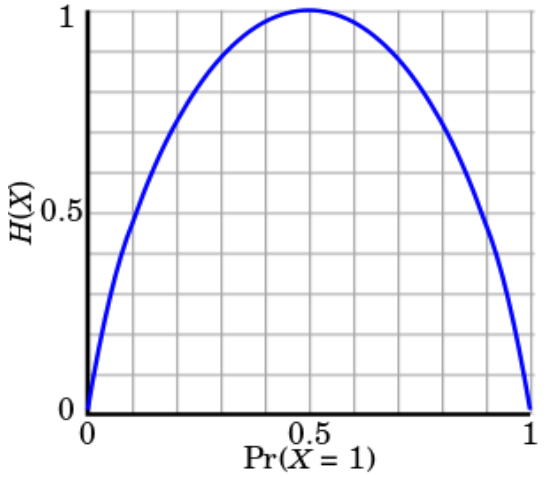
Entropy

Attempts to characterize the uncertainty of a random variable

High uncertainty => High entropy

$$H(X) = -\sum_{x \in \mathcal{X}} P(x) \log P(x).$$

Measured in 'bits' of log2, 'nats' if ln



Entropy of a fair coin?

Entropy of a 6 sided dice?

*Define 0log0 = 0

Joint Entropy

Measure of uncertainty associated with a set of variables

$$\mathrm{H}(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log_2[P(x,y)]$$

- We merely have to compute Equation (1) over all possible pairs of the two random variables
- Otherwise no different that regular entropy

In 'bits': the amount of information needed on average to specify the value of two discrete random variables.

Exercise: Find the joint entropy

	Hot	Cool
Sunny	10	5
Rainy	5	0

Conditional Entropy

Uncertainty in Y conditioned on X

- Amount of uncertainty in one given we already know the other

$$\mathrm{H}(Y|X) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)}$$

- 0 if Y is completely specified by X
- Unchanged from H(Y) if X and Y are completely independent

Note: this is not a symmetric metric!

	Hot	Cool
Sunny	5	5
Rainy	10	0

Mutual Information

Measures the relationship between two random variables

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}.$$

0 only if the variables are completely independent

Symmetric

Examples -

- X: roll of a fair die. Y: roll is even
- X: roll of a fair die, Z: roll of another fair die

A very nice explanation

KL Divergence

Measures the difference between probability distributions

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

Not symmetric

Also - Jensen Shannon distance

Bayes' Theorem

- Flipping probabilities
- Update prior beliefs based on new evidence

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}.$$

Provides a principled way to update hypotheses

$$P(H \mid E) = \frac{P(E \mid H)}{P(E)}\, P(H).$$

Widespread use in Machine Learning

- Posterior = Likelihood Ratio * Prior

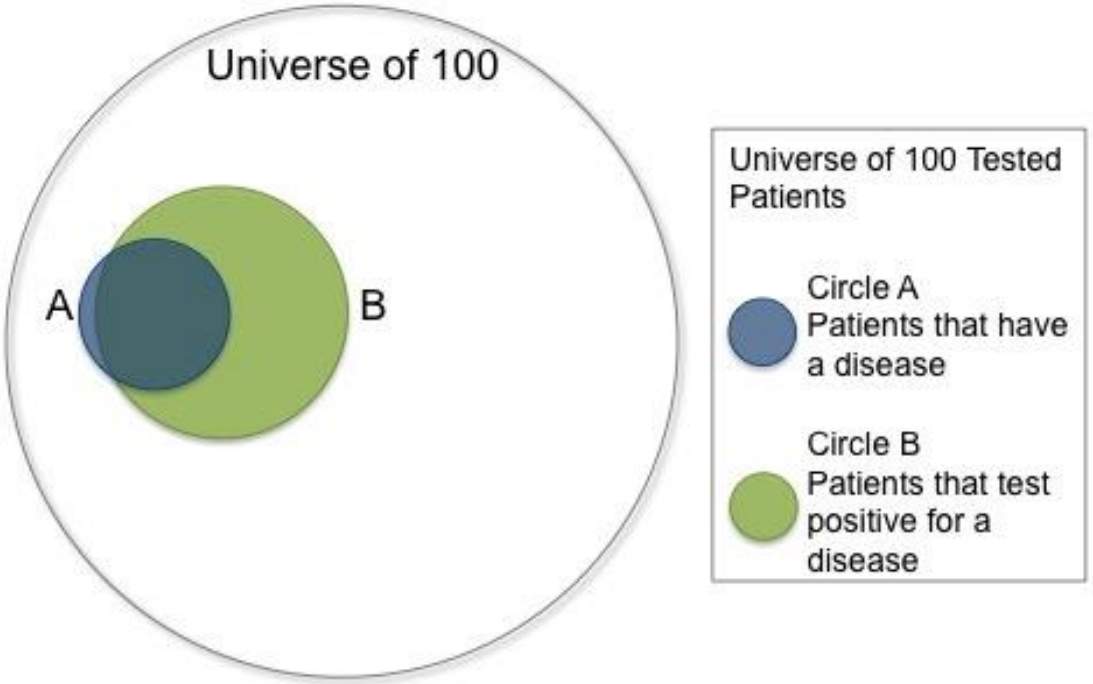
$$P(A_i|B) = \frac{P(B|A_i).P(A_i)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

$$f(\theta|X) = \frac{\pi(\theta).l(X|\theta)}{\sum \pi(\theta).l(X|\theta)}$$

$$f(\theta|X) = \frac{\pi(\theta).l(X|\theta)}{\int \pi(\theta).l(X|\theta).d\theta}$$

Bayes' Theorem - Visual example

$$P(H \mid E) = \frac{P(E \mid H)}{P(E)} P(H).$$



The waterfall diagram

- Another way of looking at it

Exercise: Disease test

Given information:
A new test for a disease has a 99% accuracy.
The disease affects 1% of the population.

$$P(H \mid E) = \frac{P(E \mid H)}{P(E)} P(H).$$

Calculate the probability that someone who tested positive for the disease actually has that disease

Exercise: Spam filtering

Given information:
The term 'Free' occurs in 20% of the emails marked as spam.
0.1% of non-spam emails include the term 'Free'.
50% of all emails are spam

$$P(H \mid E) = \frac{P(E \mid H)}{P(E)} P(H).$$

Calculate the probability that an email is spam if the word 'free' appears in it.

P (Spam) = 0.5
P (Free | Spam) = 0.2
P (Free | No spam) = 0.001
P (Spam | Free) = ?

Noisy Channel Model

Hypothetical system where -

Input: grammatically correct English (X)

Encoder: garbles the input (X => Y)

Output: English with mistakes (Y)

Or,

- Correct word to misspelled
- Turkish to English

E.g.: treat misspelled word as if correct word distorted by the channel

Objective: From observed word, find correct word that generated the observation

Noisy Channel Model - Bayesian Inference

- From observed word, find correct word that generated this
- 'Correct' word is the one that maximises the probability of getting observed word

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(w|x) \longrightarrow \hat{w} = \underset{w \in V}{\operatorname{argmax}} \frac{P(x|w)P(w)}{P(x)} \longrightarrow \hat{w} = \underset{w \in C}{\operatorname{argmax}} \overbrace{P(x|w)}^{\text{channel model}} \overbrace{P(w)}^{\text{prior}}$$

function NOISY CHANNEL SPELLING(*word x*, *dict D*, lm, *editprob*) **returns** *correction*

if $x \notin D$
candidates, edits ← All strings at edit distance 1 from x that are $\in D$, and their edit
for each c, e in candidates, edits
channel ← editprob(e)
prior ← lm(x)
score[c] = log *channel* + log *prior*
return argmax_c *score*[c]

Noisy Channel Model - Example

Error	Correction	Transformation: Correct Letter	Error Letter	Position (Letter #)	Type
acress	actress	t	—	2	deletion
acress	cress	—	a	0	insertion
acress	caress	ca	ac	0	transposition
acress	access	c	r	2	substitution
acress	across	o	e	3	substitution
acress	acres	—	s	5	insertion
acress	acres	—	s	4	insertion

w	count(w)	p(w)
actress	9,321	.0000231
cress	220	.000000544
caress	686	.00000170
access	37,038	.0000916
across	120,844	.000299
acres	12,874	.0000318

Candidate Correction	Correct Letter	Error Letter	x|w	P(x|w)
actress	t	-	c|ct	.000117
cress	-	a	a|#	.00000144
caress	ca	ac	ac|ca	.00000164
access	c	r	r|c	.000000209
across	o	e	e|o	.0000093
acres	-	s	es|e	.0000321
acres	-	s	ss|s	.0000342

Noisy Channel Model - Example

Candidate Correction	Correct Letter	Error Letter	x|w	P(x|w)	P(w)	10^9*P(x|w)P(w)
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	a|#	.00000144	.000000544	0.00078
caress	ca	ac	ac|ca	.00000164	.00000170	0.0028
access	c	r	r|c	.000000209	.0000916	0.019
across	o	e	e|o	.0000093	.000299	2.8
acres	-	s	es|e	.0000321	.0000318	1.0
acres	-	s	ss|s	.0000342	.0000318	1.0

Our guess would be **'across'**

. . . was called a "stellar and versatile **acress** *whose combination of sass and glamour has defined her. . . ".*

Solution: use larger language model instead of unigram

$$\text{P("versatile actress whose")} = .000021 * .0010 = 210 \times 10^{-10}$$
$$\text{P("versatile across whose")} = .000021 * .000006 = 1 \times 10^{-10}$$

Noisy Channel Model - Example

Candidate Correction	Correct Letter	Error Letter	x|w	P(x|w)	P(w)	10^9*P(x|w)P(w)
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	a|#	.00000144	.000000544	0.00078
caress	ca	ac	ac|ca	.00000164	.00000170	0.0028
access	c	r	r|c	.000000209	.0000916	0.019
across	o	e	e|o	.0000093	.000299	2.8
acres	-	s	es|e	.0000321	.0000318	1.0
acres	-	s	ss|s	.0000342	.0000318	1.0

Our guess would be **'across'**

. . . was called a "stellar and versatile **acress** *whose combination of sass and glamour has defined her. . . ".*

Solution: use larger language model instead of unigram

$$\text{P("versatile actress whose")} = .000021 * .0010 = 210 \times 10^{-10}$$
$$\text{P("versatile across whose")} = .000021 * .000006 = 1 \times 10^{-10}$$

manipal PROlearn

THANK YOU