Statistical Natural Language Processing

Lecture 5: Language Model Evaluation

Dr. Momtazi
Amirkabir University of Technology

Outline

1. Entropy
2. Entropy and Linguistics
3. Language Model Evaluation
4. Parameter Tuning and Cross-validation

Outline

1. Entropy

2. Entropy and Linguistics

3. Language Model Evaluation

4. Parameter Tuning and Cross-validation

Entropy

- Entropy measures the amount of information in a RV
- Amount of information contained in a message (after removing all possible redundancy)
- number of bits that the message has after compression

Entropy

$$H(V) = E[-\log(p(V))]$$

$$H(V) = \sum_{w_i \in V} -p(w_i)\log(p(w_i))$$

Note: if you want the “unit” of the entropy to be “bit”, you have to use the log to the basis 2

Example 1

- Reporting the result of rolling an 8-sided die
- Entropy:

$$H(X) = -\sum_{i=1}^{8} p(i) \log(p(i))$$

$$H(X) = -\sum_{i=1}^{8} \frac{1}{8} \log(\frac{1}{8}) = -\log(\frac{1}{8}) = \log 8 = 3 bits$$

The average length of the message needed to transmit an outcome of that variable using the optimal code

Example 1

- Reporting the result of rolling an 8-sided die
- The most efficient way is to simply encode the result as a 3 digit binary message:

1	2	3	4	5	6	7	8
001	010	011	100	101	110	111	000

Example 2

Vocabulary with two words:
$V = a, b$

$$p(a) = x$$

$$p(b) = 1 - x$$

$$H = -x \log x - (1 - x) \log(1 - x)$$

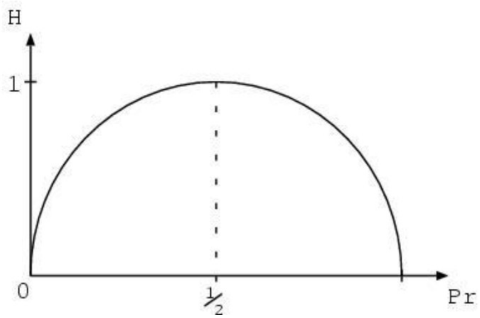
$$x = 0 \quad \rightarrow \quad H = 0$$

$$x = 1 \quad \rightarrow \quad H = 0$$

Example 2

Vocabulary with two words:
$V = a, b$

$$H = -x \log x - (1 - x) \log(1 - x)$$



Example 3

Vocabulary of W words w_i with uniform distribution $p(w_i) = 1/W$

$$H = \sum_{i=1}^{W} -p(w_i)\log(p(w_i)) = \sum_{i=1}^{W} -\frac{1}{W}\log(\frac{1}{W})$$

$$H = -W\frac{1}{W}\log(\frac{1}{W}) = -\log(\frac{1}{W}) = \log(W)$$

Entropy for uniform distribution: log of the number of symbols

Joint Entropy

- The joint entropy of 2 RV X, Y is the amount of the information needed on average to specify both their values

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

Conditional Entropy

- The conditional entropy of a RV Y given another X, expresses how much extra information one still needs to supply on average to communicate Y given that the other party knows X

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

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

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

Chain Rule

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

$$H(X_1, ...X_n) = H(X_1) + H(X_2|X_1) + ... + H(X_n|X_1, ...X_{n-1})$$

Mutual Information

- $I(X, Y)$ is the mutual information between X and Y.
- The reduction of uncertainty of one RV due to knowing about the other, or the amount of information one RV contains about the other

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

$$H(X) - H(X|Y) = H(Y) - H(Y|X) = I(X, Y)$$

Mutual Information

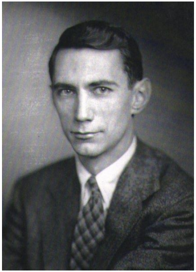
$$I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

- $I(X, Y)$ is 0 only when X and Y are independent:
  $H(X|Y) = H(X)$

Outline

Shannon Game

Shannon's Experiment to Calculate the Entropy of English



`http://www.math.ucsd.edu/~crypto/java/ENTROPY/`

Complete the Sentence

Th-r- -s -nly -n- w-y t- f-ll -n th- v-w-ls -n th-s s-nt-nc-

There is only one way to fill in the vowels in this sentence

Entropy of a Language: Shannons Approach

- Show somebody the beginning of a text
- Ask him/her to guess the next letter
- Count the number of trials

Entropy and Linguistics

- Entropy is measure of uncertainty. The more we know about something the lower the entropy
- If a language model captures more of the structure of the language, then the entropy should be lower
- We can use entropy as a measure of the quality of our models

Outline

Perplexity

- Definition:
  Perplexity is a measurement of how well a probability distribution or probability model predicts a sample.

- The perplexity of a discrete probability distribution p is defined as

$$2^{H(p)} = 2^{-\sum_{w_i \in V} p(w_i)\log(p(w_i))}$$

Perplexity

- In natural language processing, perplexity is a way of evaluating language models.
- A language model is a probability distribution over entire sentences or texts.

Branching Factor

- Branching factor is the number of possible words that can be used in each position of a text
  - Maximum branching factor for each language is V

John eats an ...

~~computer~~ apple ~~umbrella~~ ~~desk~~

~~book~~ ~~banana~~ orange

Branching Factor

- A good language model should be able to
  - minimize this number
  - give a higher probability to the words that occur in real texts

Branching Factor

Can we give the same knowledge
to a computer to predict the next character?

Perplexity

$$P(S) = P(w_1, w_2, ..., w_n)$$

$$Perplexity(S) = P(w_1, w_2, ..., w_n)^{-\frac{1}{N}} = \sqrt[N]{\frac{1}{P(w_1, w_2, ..., w_n)}}$$

$$Perplexity(S) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i|w_1, w_2, ..., w_{i-1})}}$$

Goal: giving higher probability to frequent texts
$\Rightarrow$ minimizing the perplexity of the frequent texts

Evaluation

- The evaluation must give an indication of how well the learner will do when it is asked to make new predictions for data it has not already seen.
  - Dividing the corpus into two parts

train	test

  - Building a language model from the training set
  - Estimating the probability of the test set
  - Calculate the perplexity of the test set

Perplexity

- Maximum branching factor for each language is $|V|$

$$Perplexity(S) = (\prod_{i=1}^{N} P(w_i|w_1, w_2, ..., w_{i-1}))^{-\frac{1}{N}}$$

- Example: predicting next characters instead of next words ($|V| = 26$)

$$\underset{26}{\_\_} \quad \underset{26}{\_\_} \quad \underset{26}{\_\_} \quad \underset{26}{\_\_} \quad \underset{26}{\_\_}$$

$$Perplexity(S) = ((\frac{1}{26})^5)^{-\frac{1}{5}} = 26$$

Perplexity

- Wall Street Journal
  - Training set: 38 million word tokens
  - Test set: 1.5 million words

	Unigram	Bigram	Trigram
Perplexity	962	170	109

Outline

Normal Evaluation

- Dividing the corpus into two parts

- Building a language model from the training set
- Estimating the probability of the test set
- Calculating the perplexity of the test set

Evaluation with Parameter Tuning

- Dividing the corpus into three parts

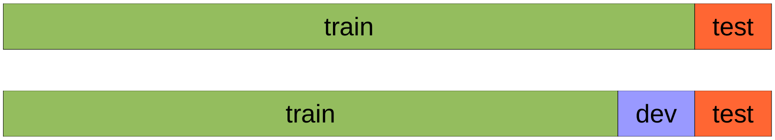
- Building a language model from the training set
- Calculating the perplexity of the development set with different parameter values
- Choosing the best parameter value and use it to estimate the probability of the test set
- Calculating the perplexity of the test set

Motivation

- There is no guarantee that the chosen test set is representative enough to model our data

- Solution:
  - Assessing how the results of a statistical analysis will generalize to an independent data set
  - Performing multiple rounds of cross-validation using different partitions, and the validation results are averaged over the rounds

Cross-validation



- *k*-fold cross-validation
- Leave-one-out cross-validation

Further Reading

- Speech and Language Processing
  - Chapter 4