DATA.STAT.840 Statistical Methods for Text Data Analysis

Exercises for Lecture 5: N-grams

Daniel Kusnetsoff

Exercise 5.1: Bigram probabilities.

a)

Are the following probabilities possible in an bigram model? $p(w_1='\text{rock}')=0.01$, $p(w_2='\text{band}')=0.003$, $p(w_2='\text{band}'|w_1='\text{rock}')=0.4$. Prove why/why not. Derive an inequality between $p(w_1)$, $p(w_2)$ and $p(w_2|w_1)$ for what probabilities are possible. Hint: consider the Bayes rule.

Let's consider the possibilities of $p(w_1 \mid w_2)$

The Bayes rule would be written here:

$p(w_1 = "rock" | \, w_2 = "band")$ = $\frac{p(w_2="band" | \, w_1="rock") * p(w_1="rock")}{p(w_2="band")}$

= $\frac{0,4*0,01}{0,003}$ = 1,333...

This is not possible, as the probability of $p(w_1 \mid w_2)$ should be between 0 and 1.

b)

Consider the sentence "The whole of science is nothing more than a refinement of everyday thinking." (Albert Einstein, *Physics and Reality*, 1936). Compute the probability of the sentence in a bigram model using the following unigram probabilities:

$p('\text{the}')=0.03$,
$p('\text{whole}')=0.0001$, $p('\text{of}')=0.01$, $p('\text{science}')=0.0003$, $p('\text{is}')=0.02$,
$p('\text{nothing}')=0.0002$, $p('\text{more}')=0.001$, $p('\text{than}')=0.0009$,
$p('\text{refinement}')=2 \cdot 10^{-6}$, $p('\text{everyday}')=6 \cdot 10^{-6}$, $p('\text{thinking}')=3 \cdot 10^{-5}$.

You need to choose some corresponding bigram probabilities so that they satisfy the condition you derived in (a).

The product we want to calculate is:

p(the) *p(whole|the) * p(of|whole) * p(science|of) * p(is|science) * (nothing|is) * p(more|nothing) * p(than|more) * p(refinement|than) * p(of|refinement) * p(everyday|of) * p(thinking|everyday)

We can calculate the conditional probabilities:

The first part we do not have is

$$p(\text{whole}|\text{the}) = \frac{p(the|whole)*p(whole)}{p(the)} = p(\text{the}|\text{whole})*\frac{1}{300}, \quad (0{,}0001/0{,}03)$$

This means we will get the equation p(whole |the) = p(the |whole)* $\frac{1}{300}$, which must be ≤ 1

Using this result we can factorize the first product of the main product equation below:

p(the)*p(whole|the)*p(of|whole)*p(science|of)*p(is|science)*(nothing|is)*p(more|nothing)*
p(than|more)*p(refinement|than)*p(of|refinement)*p(everyday|of)*p(thinking|everyday)

As we wish to resolve the sentence we wish to alter the equation to a form that shows the original sentence:

p(the)*p(the | whole)*p(whole | of)*p(of | science)*p(science | is)*(is | nothing)* p(nothing | more)* p(more | than)*p(than |refinement) *p(refinement |of)* p(of | everyday)* p(everyday | thinking)

And factorize:

p(the)*p(the | whole)*(1/300) * p(whole | of) *(100)*p(of | science) *(3/100)*p(science | is) * (200/3) *(is | nothing) * (1/100) * p(nothing | more) * 5 * p(more | than) * (9/10) *p(than |refinement) * (1/450) *p(refinement |of) * (5000)* p(of | everyday) * (3/5000) * p(everyday | thinking) * 5

choose some corresponding bigram probabilities so that they satisfy the condition in a.:

0,03 * (1/2)*(1/300) * (1/101) * 100 * (1/4) *(3/100)* (2/200)* (200/3) * (99/100)*(1/100) * (1/6)*5 * (8/10)*(9/10) * (4/5)*(1/450) * (1/5001)*5000 * (99/100)*(3/5000)* (1/6) * 5 =

1.2936E-15

Exercise 5.2: Theoretical n-gram properties.

(a) Suppose you need to generate a document of length M words. Show that if the n in an ngram model is at least as large as M, the n-gram model can represent all statistical dependencies that might exist in the language needed to generate the document. So that, for
example, a 5-gram model can represent all dependencies needed to generate sentences of 5
words.

(b) Consider a simplified version of the maximum a posteriori estimation of n-gram probabilities described on the lecture. Suppose all pseudocounts in the Dirichlet priors use
the same shared value, $\alpha_{v|[w_1,\ldots,w_{n-1}]}=\alpha_{shared}$ for all vocabulary terms v and all contexts $[w_1,\ldots, w_{n-1}]$ where α_{shared} is the shared value. This results in estimates that are simple smoothed proportional counts. This kind of smoothing is called **Laplace smoothing** when
$\alpha_{shared}=1$ and **Lidstone smoothing** otherwise.

◦ Show that in this setting, the maximum a posteriori estimate (as shown on the course slides) for a n-gram probability can be written as a weighted average of two terms: (1) the maximum likelihood estimate of the probability and (2) a uniform distribution over the vocabulary.

◦ Show that the mixing weight in the weighted average depends on the number of occurrences of a n-gram context compared to $V\,\alpha_{shared}$ where V is the vocabulary size.

◦ For an individual n-gram context, how should α_{shared} be chosen so that the weight of the data is greater than the weight of the prior?

Report your proofs.

a)

The dependency of words in M-sized document can be modelled using joint distribution.:

$$p(w_1, w_2, \ldots, w_m)$$

N is a n-gram with the assumption of size: N > M. This means that N models the dependencies of the first M-words in the distribution. This can be done using lower-degree n-grams and by the corresponding probability distribution:

p(w_1)*p(w_2| w_1) * p(w_3 | w_2, w_1) … p(w_M | w_{M-1}, w_{M-2}, … w_1)

As we can know the pattern is following the probability chain rule we can agree that it equals the previously presented joint distribution of probabilities.

⇨ Hence, the n-grams with size N can represent all dependencies necessary to create the wanted document.

b.

Suppose all pseudocounts in the Dirichlet priors use the same shared value,

$\alpha_v \mid [w_1,\ldots, w_{n-1}] = \alpha_{shared}$

Weighted Average:

$$\theta_v^{MAP} = \frac{n_{v \mid [w_1,\ldots,w_{n-1}]} + \alpha_{v \mid [w_1,\ldots,w_{n-1}]}}{n_{\mid [w_1,\ldots,w_{n-1}]} + \sum \alpha_{i \mid [w_1,\ldots,w_{n-1}]}}$$

$$= \frac{n_{v} + \alpha_{shared}}{n + \alpha_{shared}}$$

$$= \frac{n}{n + \alpha_{shared}} * \frac{n_v}{n} + \frac{\sum \alpha_{shared}}{n + \sum \alpha_{shared}} * \frac{\alpha_{shared}}{\sum \alpha_{shared}}$$

$$= \frac{n}{n + \alpha_{shared}} * \frac{n_v}{n} + \left(1 - \frac{n}{n + \sum \alpha_{shared}}\right) * \frac{\alpha_{shared}}{\sum \alpha_{shared}}$$

$\frac{n_i}{n}$ = likelihood

$\frac{\alpha_{shared}}{\sum \alpha_{shared}}$ = distribution (prior)

Mixing Weights:

V=vocabulary size

- ⇨ $\sum \alpha_{shared} = \alpha_{shared} * V$

Hence, we can write the mixing weight in the weighted average as:

$$\frac{n}{n + \alpha_{shared}} = \frac{n}{n + \alpha_{shared} * V}$$

How should α*shared* be chosen so that the weight of
the data is greater than the weight of the prior?

- ⇨ $\alpha_{shared} = ?$

$$\frac{n}{n + \alpha_{shared} * V} > 1 - \frac{n}{n + \alpha_{shared} * V}$$

=> $\frac{n}{n + \alpha_{shared} * V} > 1/2$

=> n > $(1/2) * (n + \alpha_{shared} * V)$

=> ½ > ½ V * α_{shared}

=> n/V > α_{shared}

How do we get the data weight to be greater than the prior weight?

- ⇨ We must α_{shared} so that it is a smaller value than (n/V).