DATA.STAT.840 Statistical Methods for Text Data Analysis

Exercises for Lecture 5: N-grams

Answered to problems 5.1 & 5.2

Exercise 5.1: Theoretical n-gram properties.

a)

Dependency structure of words in M-sized sequence can be modelled as a joint distribution $p(w_1, \cdots, w_M)$.

Let *N* be a n-gram of size $N \geq M$. Therefore *N* has to model dependencies of the first *M*-words by using lower-degree n-grams (1-gram, 2-gram, …, M-gram) and corresponding probability distribution $p(w_1)p(w_2|\ w_1)p(w_3|\ w_2, w_1), \cdots, p(w_M|\ w_{M-1}, \cdots, w_1)$.

By the definition of the chain rule of probability, above representation is equal to joint probability $p(w_1, \cdots, w_M)$.

Therefore, n-gram *N* can represent all statistical dependencies needed to generate the document.

b)

Weighted average

Assumptions: $\alpha_{v|\ [w_1,\ldots,w_{n-1}]} = \alpha_{shared}$

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

$$= \frac{n_v + \alpha_{shared}}{n + \sum \alpha_{shared}} \text{ (for simplicity)}$$

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

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

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

Where $\frac{n_i}{n}$ represent the likelihood and $\frac{\alpha_{shared}}{\sum \alpha_{shared}}$ represent the prior (uniform) distribution.

Mixing weight

Since *V* is the size of the vocabulary, it holds that $\sum \alpha_{shared} = V\alpha_{shared}$. Therefore the mixing weight in weighted average can be re-expressed as

$$\frac{n}{n + \sum \alpha_{shared}} = \frac{n}{n + V\alpha_{shared}},$$

where $n = n_{[w_1,\cdots,w_{n-1}]}$, i.e. the number of occurrences of certain subsequence in a n-gram context.

Selection of α_{shared}

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

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

$$n > \frac{1}{2}(n+V\alpha_{shared})$$

$$\frac{1}{2}n > \frac{1}{2}V\alpha_{shared}$$

$$\frac{n}{V} > \alpha_{shared}$$

So α_{shared} must be smaller than $\frac{n}{V}$ so that the weight of the data is grater than the weight of the prior.

Exercise 5.2: Bigram probabilities.

a)

Given probabilities are not possible. This can be noted if we write out the probability of $p(w_1|\ w_2)$.

By Bayes rule:

$$\begin{aligned} p(w_1 = \text{'rock'}|\ w_2 = \text{'band'}) &= \frac{p(w_2 = \text{'band'}|\ w_1 = \text{'rock'})p(w_1 = \text{'rock'})}{p(w_2 = \text{'band'})} \\ &= \frac{0.4 * 0.01}{0.003} \\ &= \frac{0.4 * 0.01}{0.003} \\ &= \frac{4}{3} \end{aligned}$$

$p(w_1|\ w_2)$ must be smaller than one. Therefore they are not possible.

b)

We have to calculate following product:

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

Conditional probabilities can be factorized as follows (for example):

$$p(whole|the) = \frac{p(the|whole)p(whole)}{p(the)} = p(the|whole)\frac{1}{300}$$

By the definition of question, I assume that we need to set $p(the|whole)$ (and the rest) such that $p(whole|the) = p(the|whole)\frac{1}{300} \leq 1$. and $p(the|whole) \leq 1$. If this is the correct procedure, then the first product can be factorized as:

$$p(the) * p(the|whole) * \tfrac{1}{3} * p(whole|of) * 100 * p(of|science) * \tfrac{3}{100} * p(science|is) * \tfrac{200}{3} *$$
$$p(is|nothing) * \frac{1}{100} * p(nothing|more) * 5 * p(more|than) * \frac{9}{10} * p(than|refinement) * \frac{1}{450} *$$
$$p(refinement|of) * 5000 * p(of|everyday) * \frac{3}{5000} * p(everyday|thinking) * 5$$

When fixing the unknown conditional probabilities to some numbers that satisfy the inequality above, the bigram probability is:

$$0.03 * \tfrac{2}{3} * \tfrac{1}{3} * \tfrac{1}{1000} * 100 * \tfrac{9}{10} * \tfrac{3}{100} * \tfrac{2}{200} * \tfrac{200}{3} * \tfrac{9}{10} * \tfrac{1}{100} * \tfrac{1}{6} * 5 * \tfrac{3}{4} * \tfrac{9}{10} * \tfrac{9}{10} * \tfrac{1}{450} * \tfrac{1}{5010} * 5000 *$$
$$\tfrac{9}{10} * \tfrac{3}{5000} * \tfrac{1}{6} * 5$$

$$\approx 5.45 * 10^{-14}$$