GLMTK — Notes

Lukas Schmelzeisen
lukas@uni-koblenz.de

August 11, 2014

Contents

1 Attribution

I learned about most matter described in this document during my conversations with Rene Pickhardt. Most of the formulas described here can be attributed to his research.

Besides that, much stems from other language modeling research papers.

2 Notation

Σ^n	Set of all n-Grams
$\_$	Skipped Word
$\mathcal{S}_i$	i-th Word in Sequence
$\lvert\mathcal{S}\rvert$	Number of Words in Sequence
$c(\mathcal{S})$	Absolute Count of Sequence
$N_{1+}(\mathcal{S})$	Continuation Count of Sequence
$N = c(\_)$	Number of Words
$V = N_{1+}(\_)$	Vocabulary Size

3 n-Gram Probability Estimators

A n-Gram Probability Estimator is a function $P : \Sigma^n \to [0, 1]$ which returns the probability of a n-Gram *Sequence* $\mathcal{S}$ for a fixed n-Gram *History* $\mathcal{H}$.

Rene: not to be confused with a $P_{\mathcal{H}} : \Sigma \to [0, 1]$ which is what many papers describe as an n-gram model. I do not have a word for this object yet. Maybe one could call it NextWord-n-gram Model because it only focuses on the next word. This would be only one factor in the chainrule for marginal probabilities (I know the term is not defined here)

Rene: Even though I think that you have a clear understanding of the following it is not 100% obvious from the text: I am not clear how you define a Sequence here. Is $\lvert\mathcal{S}\rvert = n$ in your case or is $\lvert\mathcal{S}_F\rvert = n$ in your case? In any case we need to define that an n-gram is an n-gram is a sequence plus and additional count how often the sequence occured.

For easier handling we define the *Full Sequence* as the concatenation of history and sequence ($\mathcal{S}_F = \mathcal{H} * \mathcal{S}$), and the *Full History* as the concatenation of history and skipped sequence ($\mathcal{H}_F = \mathcal{H} * \underbrace{\_ \dots}_{|\mathcal{S}| \text{ many}}$)

An observation on the counts of n-Grams:

$$c(\mathcal{H}) = 0 \implies c(\mathcal{H}_F) = 0 \implies c(\mathcal{S}_F) = 0 \tag{1}$$

For histories we define the predicate of a (un-)seen history. Note that this defines the empty history as "seen", which is a choice that was made in order to make the definitions and implementations of estimators more natural.

$$\begin{aligned} \mathcal{H} \text{ seen} &\iff \mathcal{H} = \varnothing \vee c(\mathcal{H}) \neq 0 \\ \mathcal{H} \text{ unseen} &\iff \mathcal{H} \neq \varnothing \wedge c(\mathcal{H}) = 0 \end{aligned} \tag{2}$$

3.1 Kinds of Probability Estimators

n-Gram probability estimators can be separated into two categories, according to which mathematical type of probability they implement **Rene: I did not know there are different kinds of mathematical probabilities**: *Conditional Probabilities* or *Marginal Probabilities*. How you actually estimate the probability of a sequence depends on what kind of estimator you are using.

n-Gram probability with conditional probability estimators:

$$P(w_1^n) = P(w_n|w_1^{n-1}) \cdot P(w_{n-1}\_|w_1^{n-2}) \cdots P(w_1 \_\_ \dots) \tag{3}$$

Rene: in the following it becomes crucial to name all probability functions differently. Even though they might all be calculated as a fraction or MLE they are indeed very different functions from different spaces. I corrected this in the following. n-Gram probability with marginal probability estimators:

$$P_{marginal}(w_1^n) = P_n(w_n|w_1^{n-1}) \cdot P_{n-1}(w_{n-1}|w_1^{n-2}) \cdots P_1(w_1) \tag{4}$$

Conditional and marginal probabilities differ on how they handle the case of an unseen history. Conditional probabilities have $P(\mathcal{S} \mid \mathcal{H} \text{ unseen}) = 0$ while marginal probabilities have $P(\mathcal{S} \mid \mathcal{H} \text{ unseen}) = P_{\text{Substitute}}(\mathcal{S}|\mathcal{H})$.

Rene: Here we need to discuss what is really happening. I understand where the substitute probability function comes from. But I do not like this as a definition. If above in the marginal probability case was defined that each factor has to be a probability function in $\mathcal{S}$ it would be clear that if the functions are calculated as fractions and $\mathcal{H}$ was unseen that this case cannot be handled like the conditional probability case. I don't have internet here but I guess there is a theorem explaining exactly what has to be

required for a product measure to be a probability function (id est explaining the chain rule of probability in the marginal case)

Rene: I still do not fully like the fact that we have a sequence and not a word as an argument we will predict. I understand that this is necessary for theory building especially with skips and rescuing conditional probabilities. Also I see that this comes from implementation details yet it feels somewhat unnatural. It could be though that I am just biased by the sheer amount of papers going only for a word as an argument

3.2 Tests for Probability Estimators

In order for probability estimators to be probability measures, the following equations / tests should hold:

`NGramProbabilitiesSumTest`:

$$\sum_{\mathcal{S} \in \Sigma^n} P(\mathcal{S}) = 1 \tag{5}$$

Rene: note for myself. this should be very soon theoretically "proofed" by me. Just also to have a proper formulation of everything in a probabilistic way `FixedHistoryProbabilitiesSumTest`:

Conditional:

$$\begin{aligned} \forall \mathcal{H} \in \Sigma^n : \; & (\mathcal{H}\_ \text{ seen} \implies \sum_{\mathcal{S} \in \Sigma} P(\mathcal{S}|\mathcal{H}) = 1) \; \wedge \\ & (\mathcal{H}\_ \text{ unseen} \implies \sum_{\mathcal{S} \in \Sigma} P(\mathcal{S}|\mathcal{H}) = 0) \end{aligned} \tag{6}$$

Marginal:

$$\forall \mathcal{H} \in \Sigma^n : \; \sum_{\mathcal{S} \in \Sigma} P(\mathcal{S}|\mathcal{H}) \; = 1 \tag{7}$$

3.3 Substitute Probability Estimators

Substitute Probability Estimators are used in a context where other probability estimators cannot use their usual algorithm to estimate the probability of a sequence. They then instead use $P_{\text{Substitute}}$ to calculate that probability. **Rene: better: In the case of marginal probabilities we have to define a probability function for each value the history can take. It now can happen that if a certain history value was unssen that the**

formula for calculation breaks and we do not receive a probability function. In those cases we just guess (as we btw do in all other cases too) a probability function

Let $P_{\text{Substitute}} \in \left\{P_{\text{Uniform}}, P_{\text{AbsUnigram}}, P_{\text{ContUnigram}}\right\}$ fixed globally at program start.

$$P_{\text{Uniform}}\big(\mathcal{S}|\mathcal{H}\big) = \frac{1}{V} \tag{8}$$

$$P_{\text{AbsUnigram}}\big(\mathcal{S}|\mathcal{H}\big) = \frac{c\big(\mathcal{S}_1\big)}{N} \tag{9}$$

$$P_{\text{ContUnigram}}\big(\mathcal{S}|\mathcal{H}\big) = \frac{N_{1+}\big(\_\mathcal{S}_1\big)}{N_{1+}\big(\_\_\big)} \tag{10}$$

Rene: here isbackoff missing. AbsUnigram is a special case of backoff where we backoff all the way to the unigram distribuion. The main reason why backoff methods work and why people can play with this in implementations seems to be that they are in the marginal setting and need a substitute function. Backing off seems most accurate since it comprises more knowledge / information thatn AbsUnigram or even uniform

All substitute probability estimators are marginal probability estimators. **Rene: I do not understand how this statement is true. It could be that we have a different definition of marginal probability estimator in mind... btw. we need terms for the full product and the factors in the marginal settings. I also like the following defintion in the marginal setting:**

$$P_n(w_n|w_i^{n-1}) = \begin{cases} e.g. MLE & \mathcal{H}_{seen} \\ substitute & \mathcal{H}_{unseen} \end{cases} \tag{11}$$

3.4 Fraction Estimators

Rene: the following is very nice from the point of view of abstraction and imiplementation detail. I am not sure if I would build up the theory from this angle. It seem s tempting though to be driven by implementation details because this ensures that every side talks about the same thing.

Fraction Estimators are probability estimators that have the form $\frac{n}{d}$.

Conditional:

$$P_{\text{Frac}[n,d]}(\mathcal{S}|\mathcal{H}) = \begin{cases} 0 & \text{if } \mathcal{H} \text{ unseen} \vee d = 0 \\ \frac{n}{d} & \text{else} \end{cases} \tag{12}$$

Marginal:

$$P_{\text{Frac}[n,d]}(\mathcal{S}|\mathcal{H}) = \begin{cases} P_{\text{Substitute}}(\mathcal{S}|\mathcal{H}) & \text{if } \mathcal{H} \text{ unseen} \vee d = 0 \\ \frac{n}{d} & \text{else} \end{cases} \tag{13}$$

3.4.1 MaximumLikelihoodEstimator

$$P_{\text{MLE}}(\mathcal{S}|\mathcal{H}) = P_{\text{Frac}[c(\mathcal{S}_F),c(\mathcal{H}_F)]}(\mathcal{S}|\mathcal{H}) \tag{14}$$

3.4.2 "False"MaximumLikelihoodEstimator

FMLE only works in the marginal probability setting.

$$P_{\text{FMLE}}(\mathcal{S}|\mathcal{H}) = P_{\text{Frac}[c(\mathcal{S}_F),c(\mathcal{H})]}(\mathcal{S}|\mathcal{H}) \tag{15}$$

3.4.3 ContinuationMaximumLikelihoodEstimator

$$P_{\text{CMLE}}(\mathcal{S}|\mathcal{H}) = P_{\text{Frac}[N_{1+}(\_\mathcal{S}_F),N_{1+}(\_\mathcal{H}_F)]}(\mathcal{S}|\mathcal{H}) \tag{16}$$

3.5 Discount Estimators

A Discount Estimator takes any kind of fraction estimator and subtracts some discount from the numerator, in order to free probability mass to be used for smoothing. Obviously this means, that discount estimators are no longer probability estimators, and will not pass tests. Instead they have to be used in conjunction with an Interpolation Estimator. Discount estimators are still fraction estimators though.

Rene: I would not be sure if we should say it like this. I understand that we basically always have fraction estimators in mind since everything starts with this and modifies and combines this in some way. still I guess discounting an many other methods could work with other models too

$$P_{\text{Discount}[D,P_{\text{Frac}}[n,d]]}(\mathcal{S}|\mathcal{H}) = P_{\text{Frac}[\max(0,n-D),d]}(\mathcal{S}|\mathcal{H}) \tag{17}$$

With $D : \mathcal{H} \to [0, 1]$.

3.5.1 AbsoluteDiscountEstimator

$$P_{\text{AbsDiscount}[D,P_{\text{Frac}}[n,d]]}(\mathcal{S}|\mathcal{H}) = P_{\text{Discount}[D,P_{\text{Frac}}[n,d]]}(\mathcal{S}|\mathcal{H}) \tag{18}$$

With $D \in [0, 1]$.

3.6 Backoff Estimators

Conditional:

$$P_{\text{Backoff}[P_\alpha,P_\beta]}(\mathcal{S}|\mathcal{H}) = \begin{cases} 0 & \text{if } \mathcal{H} = \varnothing \\ \gamma(\mathcal{H}) \cdot P_\beta(\mathcal{S}|\hat{\mathcal{H}}) & \text{if } c(\mathcal{S}_F) = 0 \\ P_\alpha(\mathcal{S}|\mathcal{H}) & \text{else} \end{cases} \tag{19}$$

Marginal:

$$P_{\text{Backoff}[P_\alpha,P_\beta]}(\mathcal{S}|\mathcal{H}) = \begin{cases} P_{\text{Substitute}}(\mathcal{S}|\mathcal{H}) & \text{if } \mathcal{H} = \varnothing \\ \gamma(\mathcal{H}) \cdot P_\beta(\mathcal{S}|\hat{\mathcal{H}}) & \text{if } c(\mathcal{S}_F) = 0 \\ P_\alpha(\mathcal{S}|\mathcal{H}) & \text{else} \end{cases} \tag{20}$$

With P_α, P_β any probability estimators and *backoff coefficient* γ:

$$\gamma(\mathcal{H}) = \frac{1 - \sum_{\mathcal{S}\in\Sigma:c(\mathcal{H}\mathcal{S})>0} P_\alpha(\mathcal{S}|\mathcal{H})}{\sum_{\mathcal{S}\in\Sigma:c(\mathcal{H}\mathcal{S})=0} P_\beta(\mathcal{S}|\hat{\mathcal{H}})} \tag{21}$$

3.7 Interpolation Estimators

Rene: why is there no difference in conditional and marginal case. I just see there is latex comments where you tried this...

$$P_{\text{Interpol}[P_{\text{Discount}}[D,P_{\text{Frac}}[n,d]],P_\beta]}(\mathcal{S}|\mathcal{H}) = \tag{22}$$
$$\begin{cases} P_{\text{Frac}[n,d]}(\mathcal{S}|\mathcal{H}) & \text{if } \mathcal{H} = \varnothing \ \vee \ \mathcal{H} \text{ contains only } \_ \\ P_{\text{Discount}[D,P_{\text{Frac}}[n,d]]}(\mathcal{S}|\mathcal{H}) \ + \ \gamma(D,d,H) \cdot P_\beta(\mathcal{S}|\mathcal{H}) & \text{else} \end{cases}$$

With the P_β any probability estimator and *interpolation coefficient* γ:

$$\gamma(D,d,H) = \begin{cases} 0 & \text{if } d = 0 \\ \frac{D \cdot N_{1+}(\mathcal{H}\_)}{d} & \text{else} \end{cases} \tag{23}$$

3.8 Combination Estimator

A Combination Estimator mixes two other probability estimators.

$$P_{\text{Comb}[\lambda,P_\alpha,P_\beta]}(\mathcal{S}|\mathcal{H}) = \lambda \cdot P_\alpha(\mathcal{S}|\mathcal{H}) \;\; + \;\; (1-\lambda) \cdot P_\beta(\mathcal{S}|\mathcal{H}) \tag{24}$$

With P_α, P_β any probability estimators and $\lambda \in [0, 1]$.

4 TODO

- Interpol estimator still fails test for conditional case.
- FMLE still doesn't pass any tests.
- How to make `FixedHistoryProbabilitiesSumTest` with Continuation Estimators not be a mess?
- How to do `FixedHistoryProbabilitiesSumTest` with Combination Estimator.
- **Rene: how to introduce skips on a general basis (also with POS?)**