GLMTK — Notes

Lukas Schmelzeisen
lukas@uni-koblenz.de

August 10, 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* $\mathcal{H}$.

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 ($\mathcal{H}_F = \mathcal{H} * \underbrace{\_ \ldots}_{|\mathcal{S}| \text{ many}}$)

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: *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:

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

n-Gram probability with marginal probability estimators:

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

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 $P(\mathcal{S} \mid \mathcal{H} \text{ unseen}) = P_{\text{Substitute}}(\mathcal{S}|\mathcal{H})$.

3.2 Tests for Probability Estimators

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

`NGramProbabilitiesSumTest`:

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

`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.

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

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

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

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

All substitute probability estimators are marginal probability estimators.

3.4 Fraction Estimators

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

Conditional:

$$P_{\text{Frac}[n,d]}\big(\mathcal{S}|\mathcal{H}\big) = \begin{cases} 0 & \text{if } \mathcal{H} \text{ unseen} \vee d = 0 \\ \frac{n}{d} & \text{else} \end{cases} \tag{11}$$

Marginal:

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

3.4.1 MaximumLikelihoodEstimator

$$P_{\text{MLE}}\big(\mathcal{S}|\mathcal{H}\big) = P_{\text{Frac}[c(\mathcal{S}_F),c(\mathcal{H}_F)]}\big(\mathcal{S}|\mathcal{H}\big) \tag{13}$$

3.4.2 "False"MaximumLikelihoodEstimator

FMLE only works in the marginal probability setting.

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

3.4.3 ContinuationMaximumLikelihoodEstimator

$$P_{\text{CMLE}}(\mathcal{S}|\mathcal{H}) = P_{\text{Frac}[N_{1+}(\_\mathcal{S}_F),N_{1+}(\_\mathcal{H}_F)]}(\mathcal{S}|\mathcal{H}) \tag{15}$$

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.

$$P_{\text{Discount}[D,P_{\text{Frac}}[n,d]]}(\mathcal{S}|\mathcal{H}) = P_{\text{Frac}[\max(0,n-D),d]}(\mathcal{S}|\mathcal{H}) \tag{16}$$

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

3.5.1 AbsoluteDiscountEstimator

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

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

3.6 Backoff Estimators

Conditional:

$$P_{\text{Backoff}[P_\alpha,P_\beta]}(\mathcal{S}|\mathcal{H}) = \begin{cases} 0 & \text{if } \mathcal{H} = \varnothing \\ \gamma(\mathcal{H}) \cdot P_\beta(\mathcal{S}|\hat{\mathcal{H}}) & \text{if } c(\mathcal{S}_F) = 0 \\ P_\alpha(\mathcal{S}|\mathcal{H}) & \text{else} \end{cases} \tag{18}$$

Marginal:

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

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

$$\gamma(\mathcal{H}) = \frac{1 - \sum_{\mathcal{S}\in\Sigma:c(\mathcal{H}\mathcal{S})>0} P_\alpha(\mathcal{S}|\mathcal{H})}{\sum_{\mathcal{S}\in\Sigma:c(\mathcal{H}\mathcal{S})=0} P_\beta(\mathcal{S}|\hat{\mathcal{H}})} \tag{20}$$

3.7 Interpolation Estimators

$$P_{\text{Interpol}[P_{\text{Discount}}[D,P_{\text{Frac}}[n,d]],P_\beta]}\big(\mathcal{S}|\mathcal{H}\big) = \tag{21}$$
$$\begin{cases} P_{\text{Frac}[n,d]}\big(\mathcal{S}|\mathcal{H}\big) & \text{if } \mathcal{H} = \varnothing \ \vee \ \mathcal{H} \text{ contains only } \_ \\ P_{\text{Discount}[D,P_{\text{Frac}}[n,d]]}\big(\mathcal{S}|\mathcal{H}\big) \ + \ \gamma(D,d,H) \cdot P_\beta\big(\mathcal{S}|\mathcal{H}\big) & \text{else} \end{cases}$$

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

$$\gamma(D,d,H) = \begin{cases} 0 & \text{if } d = 0 \\ \frac{D \cdot N_{1+}(\mathcal{H}\_)}{d} & \text{else} \end{cases} \tag{22}$$

3.8 Combination Estimator

A Combination Estimator mixes two other probability estimators.

$$P_{\text{Comb}[\lambda,P_\alpha,P_\beta]}\big(\mathcal{S}|\mathcal{H}\big) = \lambda \cdot P_\alpha\big(\mathcal{S}|\mathcal{H}\big) \ + \ (1-\lambda) \cdot P_\beta\big(\mathcal{S}|\mathcal{H}\big) \tag{23}$$

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.