A mathematical theory of Language

Heinrich Hartmann, Rene Pickhardt

April 30, 2014

Contents

1 Preliminaries

1.1 Spaces of Sequences

Let A be a finite set. In Examples, this will be the set of characters or the set of words. We define the set of Sequences over A as

$$\Sigma A = \{(a_1, \ldots, a_l) \,|\, a_i \in A, l \geq 0\}.$$

We introduce the following notation:

- The empty seuqence is denoted by $\epsilon \in \Sigma A$.
- The length function is denoted by

$$length : \Sigma A \longrightarrow \mathbb{N}_0, \quad (a_1, \ldots, a_l) \mapsto l.$$

- We have the following canonical decomposition by sequence length

$$\Sigma A = A^0 \cup A^1 \cup A^2 \dots$$

and denote by $i_N : A^N \to \Sigma A$ the inclusion of the length-i sequences into ΣA.

Furthermore for each index $i \geq 0$ we have a projection

$$\pi_i : \Sigma A \longrightarrow A_+, \quad (a_1, \dots, a_l) \mapsto \begin{cases} a_i & l \geq i \\ \langle \text{undef} \rangle & l < i \end{cases}.$$

Here $A_+ := A \cup \{\langle \text{undef} \rangle\}$. This space allows π_i to be defined on whole of ΣA.

Furthermore if $N \geq 0$ and $i \geq 1$ are given, we define the i-th N-gram projection to be:

$$\pi_i^N : \Sigma A \longrightarrow A_+^N, \quad (a_1, \dots, a_l) \mapsto \begin{cases} (a_{i+0}, \dots, a_{i+N-1}) & l \geq i + N - 1 \\ \langle \text{undef} \rangle & l < i + N - 1 \end{cases}.$$

Note that we get back π_i as π_i^1. Moreover π_i^0 is the canonical projection of ΣA to the one point space $A^0 \subset A_+^0$.

We have the following two filtrations of ΣA

$$A^0 = \Sigma^{\leq 0} A \subset \Sigma^{\leq 1} A \subset \Sigma^{\leq 2} A \subset \dots \subset \Sigma A$$

and

$$\Sigma A = \Sigma_{\geq 0} A \supset \Sigma_{\geq 1} A \supset \Sigma_{\geq 2} A \supset \dots \supset \bigcap \Sigma_{\geq i} A = \emptyset$$

where

$$\Sigma^{\leq i} A = A^0 \cup \dots A^i$$

and

$$\Sigma_{\geq i} A = A^i \cup A^{i+1} \cup \dots$$

1.2 Spaces of probability measures

Let X be an at most countable[1] set. We denote by

$$\mathcal{M}(X) = \{\mu : X \longrightarrow \mathbb{R}_{\geq 0} \mid \sum_{x \in X} \mu(x) < \infty\}$$

[1] The assumption of countability could be dropped, at the expense of a slightly more technical treatment of infinite sums.

the space of all finite measures on X. For $A \subset X$ we write $\mu[A] = \sum_{x\in A} \mu(x)$. This definition agrees with the usual definition, of measures in the case of discrete spaces with maximal σ-algebra.

The set of probability measures is defined as

$$\mathcal{P}(X) = \{\mu \,|\, \mu[X] = 1\} \subset \mathcal{M}(X).$$

We get a normailzation map,

$$\mathcal{M}(X) \setminus \{0\} \longrightarrow \mathcal{P}(X), \quad \mu \mapsto \frac{1}{\mu[X]}\mu$$

which is defined for non-zero measures μ.

Let $f : X \longrightarrow Y$ be a map of sets, then we get a natural map

$$f_* : \mathcal{M}(X) \longrightarrow \mathcal{M}(Y), \quad f_*(\mu)(y) = \mu[f^{-1}(\{y\})] = \sum_{x:f(x)=y} \mu(x)$$

as well as $f_* : \mathcal{P}(X) \to \mathcal{P}(Y)$.

If f has finite fibers, then we also have the following map:

$$f^* : \mathcal{M}(Y) \longrightarrow \mathcal{M}(X), \quad \mu \mapsto \mu \circ f,$$

however the total volume of Y is not preserved, so that no map on $\mathcal{P}$ is induced. In particular case, that $\iota : A \subset X$ is the inclusion of a subspace we write $\mu|_A$ for $\iota^*(\mu)$. If for $P \in \mathcal{P}(X)$ the restriction $P|_A$ is not necessary a probability measure. If $P[A] \neq 0$, then $P|_A$ can be normalized to

$$P[\_|A] = \frac{1}{P[A]} P|_A,$$

the conditional probability measure on A.

If $\mu \in \mathcal{M}_f(X)$ and $g : X \to \mathbb{R}$, we define the *expectation* of g as:

$$E_\mu[f] := \sum_{x\in X} g(x)\mu(x).$$

This sum is well defined since μ is finitely supported.

1.3 Probability measures on sequences

In this section we study the relationship between $\mathcal{P}(A)$ and $\mathcal{P}(\Sigma A)$.

For $i \geq 1$ and $N \geq 0$ we have the following maps:

$$\pi_{i*} : \mathcal{P}(\Sigma A) \longrightarrow \mathcal{P}(A_+), \quad {\pi_i^N}_* : \mathcal{P}(\Sigma A) \longrightarrow \mathcal{P}(A_+^N),$$

as well as

$$length_* : \mathcal{P}(\Sigma A) \longrightarrow \mathcal{P}(\mathbb{N}_0).$$

1.3.1 from Language Models to unigram measures

Hence, for each probability measure P on ΣA, we have $\pi_{i*}P$ which is a measure on A_+. Note that

$$\pi_{i*}P(\langle\text{undef}\rangle) = P[\{s \mid \mathit{length}(s) < i\}] = P[\mathit{length} < i].$$

In the case, that $\pi_{i*}P(\langle\text{undef}\rangle) \neq 1$ we can normailze $\pi_{i*}P$ to the i-th element distribution

$$D_i P = \frac{1}{P[\mathit{length} \geq i]} \pi_{i*}P = \pi_{i*}P[\_|\mathit{length} \geq i] \ \in \ \mathcal{P}(A)$$

on A.

To define a total distribution of all elements, we want to take the following infinite sum of i-th element distributions

$$\sum_{i \geq 1} \pi_{i*}P$$

However, this sum is not necessarily finite for a general measure $P \in \mathcal{P}(\Sigma A)$, therefore we make the additional assumption that P is finitely supported and define

$$M^1 : \mathcal{M}_f(\Sigma A) \longrightarrow \mathcal{M}(A), \quad \mu \mapsto \sum_{i \geq 1} \pi_{i*}\mu.$$

Note that the measure is only defined on $A \subset A_+$, since we have

$$\sum_{i \geq 1} \pi_{i*}(\langle\text{undef}\rangle) = \sum_{i \geq 1} P[\mathit{length} < i] = \infty.$$

We calculate the total volume to be

$$M^1\mu(A) = \sum_{i \geq 1} \mu[\mathit{length} \geq i] = E_\mu[\mathit{length}].$$

Hence, for $P \in \mathcal{P}_f(\Sigma A)$ with $P[\mathit{length} = 0] \neq 1$, we can normalize the measure $M^1 P$, and define *unigram distribution* on A as

$$D^1 P := \frac{1}{E_P[\mathit{length}]} \sum_{i \geq 1} \pi_{i*}P \ \in \ \mathcal{P}(A).$$

1.3.2 From unigram measures to Language Models

Let $P \in \mathcal{P}(A)$ be a probability measure. $\forall l \in \mathbb{N}$ we can pull back P via π_i to $P(A^l)$ and define

$$B^l P = \prod_{i=1}^{l} \pi_i^* P = \prod_{i=1}^{l} P \circ \pi_i$$

Explicitly for $s = (a_1, \dots, a_l) \in A^l$ this means $P(s) = \prod_{i=1}^{l} P(\pi_i(s))$. Note that in these cases $\pi_i(s) \in A$ is well defined.

In the case of $l = 1$ it is trivial to see that we receive a probability measure. For $l = 2$ we have:

$$\sum_{s \in A^2} B^2 P = \sum_{s \in A^2} \prod_{i=1}^{2} P(\pi_i(s)) = \sum_{a_1, a_2} P(a_1)P(a_2) = \sum_{a_1} P(a_1) \sum_{a_2} P(a_2) = \sum_{a_1} P(a_1) = 1$$

A similar argument will hold for all $l \in \mathbb{N}$ showing that the pulled back probability measure $B^l P$ is indeed a probability measure on A^l.

In order to construct a measure in $\mathcal{P}(\Sigma(A))$ starting from $P \in \mathcal{P}(A)$ we have a lot of choice (indicating that $\mathcal{P}(\Sigma(A))$ is indeed bigger than $\mathcal{P}(A)$). Note that simply adding up $B^l P$ by setting

$$B_{naiveTry} P = \sum_{l \geq 1} B^1 P$$

will not work as

$$\sum_{s \in \Sigma A} B_{naiveTry} P(s) = \infty \neq 1$$

We can simply fix this by weighting the sum with an arbitrary chosen probability distribution $P_{weight} \in \mathcal{P}(\mathbb{N})$. So we receive a Langugage Model from a Unigram Model by setting:

$$BP = \sum_{l \geq 1} P_{weight}(l) B^1 P = \sum_{l \geq 1} P_{weight}(l) \prod_{i=1}^{l} P \circ \pi_i$$

When applying statistics one could estimate the length distribution on sentences as a weighting distribution. **We think one should investigate smoothing techniques for language models by changing this choice** Another open end which I did not include yet is the idea of pushing forward $B^l P$ via $i_N : A^N \longrightarrow \Sigma(A)$. Well I think I did this implicetly by not being totally clean when using π_i of stating if it was defined on ΣA or A^l. **is it possible to show that the above mentioned choice is up to a probability measure from $\mathcal{P}(\mathbb{N})$?**

1.3.3 N-gram measures

More generally, for integers $i \geq 1$ and $N \geq 0$ we get a measure ${\pi_i^N}_*(P)$ on A_+^N, for which

$$ {\pi_i^N}_* P(\langle \text{undef} \rangle) = P[\mathit{length} < i + N - 1]. $$

If this number is not equal to one, we define the i-th N-gram distribution to be

$$ D_i^N P = \frac{1}{P[\mathit{length} \geq i + N - 1]} \, {\pi_i^N}_* P \;\in\; \mathcal{P}(A^N) $$

For the global N-gram distributions we take

$$ M^N : \mathcal{M}_f(\Sigma A) \longrightarrow \mathcal{M}(A^N), \quad \mu \mapsto \sum_{i \geq 1} {\pi_i^N}_* \mu. $$

Again, this measure is only defined on $A^N \subset A_+^N$. We calculate the total volume as

$$ \begin{aligned} M^N(\mu)(A^N) &= \sum_{i \geq 1} \mu[\mathit{length} \geq i + (N-1)] && (1) \\ &= \sum_{j = N} (j - (N-1)) \mu[\mathit{length} = j] && (2) \\ &= E_\mu[\max\{0, \mathit{length} - (N-1)\}] && (3) \end{aligned} $$

Note that this number depends only on the length distribution $\mathit{length}_*(\mu) \in \mathcal{M}_f(\mathbb{N}_0)$ and is non-zero if and only if $\mu[\mathit{length} \geq N] \neq 0$.

Hence, for $P \in \mathcal{P}_f(\Sigma A)$, with $P[\mathit{length} \geq N] > 0$ we can define the *total N-gram distribution* as

$$ D^N P = \frac{1}{E[\max\{0, \mathit{length} - (N-1)\}]} \sum_{i \geq 1} {\pi_i^N}_* P \;\in\; \mathcal{P}(A^N). $$

Note, that the special case of $N = 1$ reduces to our earlier definition.

1.4 Markov measures

A probability measure P on ΣA is called K-Markov if for all $l \geq K$, $b_0, \ldots, b_l \in A$ and $n > l$ the condition

$$ \begin{aligned} &P[\pi_n = b_0 \,|\, \pi_{n-1} = b_1, \ldots, \pi_{n-K} = b_K, \ldots, \pi_{n-l} = b_l] \\ &= P[\pi_n = b_0 \,|\, \pi_{n-1} = b_1, \ldots, \pi_{n-K} = b_K] \end{aligned} $$

holds whenever both sides are well-defined, i.e. $P[\pi_{n-1} = b_1, \ldots, \pi_{n-l} = b_l]$ is non-zero.

1.4.1 The case of $\langle\text{undef}\rangle$

The above definition, does not specify a condition in the case one ore more of the b_i are $\langle\text{undef}\rangle$. For $b_0 = \langle\text{undef}\rangle$ is unproblematic. In the case that $b_0 \in A$ and $b_j = \langle\text{undef}\rangle$ for one $j > 0$, the condition is empty since $\pi_{i-j} = \langle\text{undef}\rangle$ implies $\pi_i = \langle\text{undef}\rangle$. For the remaining case of $b_0 = \langle\text{undef}\rangle$ and $b_j = \langle\text{undef}\rangle$ for one or more $j > 0$, the naive-condition does not extend. To see this, we take $K = 0$ and $l = 1$ with $b_1 = \langle\text{undef}\rangle$, so that the extended condition reads

$$P[\pi_n = \langle\text{undef}\rangle \,|\, \pi_{n-1} = \langle\text{undef}\rangle] = P[\pi_n = \langle\text{undef}\rangle]. \tag{4}$$

This implies $P[\pi_{n-1} = \langle\text{undef}\rangle] = 1$, which does not always hold.

1.4.2 0-Markov measures

In the special case of $K = 0$ we find

$$\begin{aligned} & P[\pi_n = b_0] = P[\pi_n = b_0 \,|\, \pi_{n-1} = b_1] \\ \Leftrightarrow \quad & P[\pi_n = b_0, \pi_{n-1} = b_1] = P[\pi_n = b_0] P[\pi_{n-1} = b_1] \end{aligned}$$

which is the definition of P-independent between π_n and π_{n-1} random variables, except that the case $\langle\text{undef}\rangle$ is excluded. We can account for that by using conditional probabilities. Assume that $P[\mathit{length} \geq n] > 0$ then, π_n, π_{n-1} are $P[\_|\mathit{length} \geq n]$ independent random variables on $\Sigma_{\geq n} A$.

Similarly, we see that the full collection $\{\pi_i\}_{i \leq n}$ is $P[\_|\mathit{length} \geq n]$-independent on $\Sigma_{\geq n} A$. Indeed,

$$P[\pi_0 = a_0, \ldots, \pi_n = a_n|\mathit{length} \geq n] = \frac{1}{P[\mathit{length} \geq n]} P[\pi_0 = a_0, \ldots, \pi_n = a_n]$$

and

$$\begin{aligned} P[\pi_0 = a_0, \ldots, \pi_n = a_n] =& P[\pi_0 = a_0, \ldots, \pi_n = a_n] \\ =& P[\pi_n = a_n|\pi_{n-1} = a_{n-1}, \ldots, \pi_0 = a_0] \\ & P[\pi_{n-1} = a_{n-1}|\pi_{n-2} = a_{n-1}, \ldots, \pi_0 = a_0] \\ & \ldots \\ & P[\pi_0 = a_0] \end{aligned}$$