Yannick Wattenberg: Assignment 04

ywattenberg@inf.ethz.ch, 19-947-464.

14/01/2023 - 14:04h

Question 1: Calculating Prefix Probabilities

a)

We prove by The following equality is given

$$\sum_{w \in \Sigma} p(w) = 1 \tag{1}$$

We first prove the following equivalence:

$$\sum_{w \in \Sigma^*, |w|=n} \tilde{p}(w) = 1 \tag{2}$$

Using induction over n and the above equation as induction hypothesis.

$$\sum_{w \in \Sigma^*, |w|=1} \tilde{p}(w) = \sum_{w \in \Sigma} p(w) = 1 \tag{3}$$

$$n \to n+1$$

$$\sum_{w \in \Sigma^*, |w|=n+1} \tilde{p}(w) = \sum_{w' \in \Sigma} \sum_{w \in \Sigma^*, |w|=n} \tilde{p}(w' \circ w) \tag{4}$$

$$= \sum_{w' \in \Sigma} \sum_{w \in \Sigma^*, |w|=n} p(w') \tilde{p}(w) \qquad \text{def. } \tilde{p} \tag{5}$$

$$= \sum_{w' \in \Sigma} p(w') \cdot \sum_{w \in \Sigma^*, |w|=n} \tilde{p}(w) \tag{6}$$

$$= \sum_{w' \in \Sigma} p(w') \qquad (IH) \tag{7}$$

$$= 1 \qquad 1 \tag{8}$$

With this we can reformulate the sum as follows:

$$\sum_{w\in\Sigma^*}\tilde{p}(w) = \sum_{i=0}^{\infty}\sum_{w\in\Sigma^*,|w|=i}\tilde{p}(w) \tag{9}$$

$$= \sum_{i=0}^{\infty}\sum_{w\in\Sigma^*,|w|=i}\tilde{p}(w) \tag{10}$$

$$= \sum_{i=0}^{\infty} 1 \tag{11}$$

$$\sum_{i=1}^{\infty} 1 \to \infty \tag{12}$$

b)

The following equality is given

$$\sum_{w\in\Sigma\cup EOS} p(w) = 1 \Rightarrow \sum_{w\in\Sigma} p(w) = 1 - p(EOS) \tag{13}$$

We first prove the following equivalence:

$$\sum_{w\in\Sigma^*,|w|=n}\tilde{p}(w) = (1-p(EOS))^n \tag{14}$$

Using induction over n and the above equation as induction hypothesis.

$$\sum_{w\in\Sigma^*,|w|=1}\tilde{p}(w) = \sum_{w\in\Sigma} p(w) = 1 - p(EOS) \qquad 13 \tag{15}$$

$$n \to n+1$$

$$\sum_{w\in\Sigma^*,|w|=n+1}\tilde{p}(w) = \sum_{w'\in\Sigma}\sum_{w\in\Sigma^*,|w|=n}\tilde{p}(w'\circ w) \tag{16}$$

$$= \sum_{w'\in\Sigma}\sum_{w\in\Sigma^*,|w|=n} p(w')\tilde{p}(w) \qquad \text{def. } \tilde{p} \tag{17}$$

$$= \sum_{w'\in\Sigma} p(w') \cdot \sum_{w\in\Sigma^*,|w|=n}\tilde{p}(w) \tag{18}$$

$$= \sum_{w'\in\Sigma} p(w') \cdot (1-p(EOS))^n \qquad (IH) \tag{19}$$

$$= (1-p(EOS)) \cdot (1-p(EOS))^n \qquad 13 \tag{20}$$

$$= (1-p(EOS))^{n+1} \tag{21}$$

With this we can reformulate the sum as follows:

$$\sum_{w\in\Sigma^*} p(w) = \sum_{w\in\Sigma^*} p(EOS)\tilde{p}(w) = p(EOS)\sum_{i=0}^{\infty}\sum_{w\in\Sigma^*,|w|=i}\tilde{p}(w) \tag{22}$$

$$= p(EOS)\left(\sum_{i=0}^{\infty}\sum_{w\in\Sigma^*,|w|=i}\tilde{p}(w)\right) \tag{23}$$

$$= p(EOS)\left(\sum_{i=0}^{\infty}(1-p(EOS))^i\right) \tag{24}$$

$$p(EOS)\left(\sum_{i=0}^{\infty}(1-p(EOS))^i\right) \to p(EOS)\frac{1}{p(EOS)} = 1 \quad \text{(geom. series)} \tag{25}$$

c)

$$p_{pre}(w) \stackrel{!}{=} \sum_{u\in\Sigma^*} p(wu) \tag{26}$$

$$\sum_{u\in\Sigma^*} p(wu) = \sum_{u\in\Sigma^*} p(EOS|wu)p_{pre}(u|w)p_{pre}(w) \tag{27}$$

$$= p_{pre}(w)\left(\sum_{u\in\Sigma^*} p(EOS|wu)p_{pre}(u|w)\right) \tag{28}$$

$$= \frac{1}{p(EOS|w)}p(w)\left(\sum_{u\in\Sigma^*} p(u|w)\right) \quad \text{(def.}P\text{)} \tag{29}$$

$$= \frac{1}{p(EOS|w)}\left(\sum_{u\in\Sigma^*} p(w)\frac{p(w|u)p(u)}{p(w)}\right) \quad \text{(Bayes.)} \tag{30}$$

$$= \frac{1}{p(EOS|w)}\left(\sum_{u\in\Sigma^*} p(w|u)p(u)\right) \tag{31}$$

$$= \frac{1}{p(EOS|w)}p(w) \quad \text{(Bayes.)} \tag{32}$$

$$= p_{pre}(w) \quad \text{(def.}P\text{)} \tag{33}$$

In (30) we apply baysens rule which holds as the sum iterates over all possible suffixes. This means we sum over the whole probability space which gives probability one for the event we condition on.

d)

One can use the normal CKY algorithm where we define the score-function as the natural logarithm of the probability of the applied rule. This means the CKY algorithm will

calculate

$$chart[i,k,X] \mathrel{+}= exp(score(X \to YZ)) \cdot chart[i,j,Y] \cdot chart[j,k,Z] \tag{34}$$
$$= p(YZ|X) \cdot p(Y|w_i,...,w_j) \cdot p(Z|w_j,...,w_k) \tag{35}$$
$$\Rightarrow chart[i,k,X] = p(X|w_i,...,w_k) \tag{36}$$

Which gives us $p(w_1,...,w_n|S)$ for the top most cell. This is the Prefix probability of the sentence we can multiply this by $p(EOS|S)$ to get the probability of the sentence.

e)

$$\sum_{u \in \Sigma^*} p(wu) \overset{!}{=} p(S \overset{*}{\Rightarrow} wv) \tag{37}$$
$$\sum_{u \in \Sigma^*} p(wu) = \sum_{u \in \Sigma^*} p_{inside}(wu|S) \qquad \text{(S is the starting symbol)} \tag{38}$$
$$= \sum_{u \in \Sigma^*} p(S \overset{*}{\Rightarrow} wu) \qquad \text{(def. } p_{inside}\text{)} \tag{39}$$
$$= p(S \overset{*}{\Rightarrow} wv) \qquad \text{(for some arbitrary } v \in \Sigma^*\text{)} \tag{40}$$

In the last step we replace the sum by some arbitrary suffix v. This is possible as the sum iterates over all possible suffixes and ($\overset{!}{\Rightarrow}$) produces all possible derivations meaning it also generates all possible suffixes.

f)

First we define the probability matrix W of size $|\mathcal{N}| \times |\mathcal{N}|$ as follows: Every entry $W[A,B]$ gets the probability of $p[A \Rightarrow B\alpha]$ where α is the rest of the rule. We can calculate this probability by adding the probabilities of all rules which produce B as a first symbol. There can be at most $|\mathcal{N}|$ such rules as our grammar is in Chomsky normal form. This means it is possible to calculate all entries of the matrix W in $\mathcal{O}(|\mathcal{N}|^3)$.

Now we have to calculate W^*. We can do this using Lehmann's algorithm as seen in the lecture notes. For this to work it has to hold that entries on the diagonal of W are smaller than one, otherwise $W[A,A]^*$ would be infinity/diverge. However all entries on the diagonal have to be smaller than one as otherwise this would mean all derivations from A produce another A thus every sentence would be infinite, which we assume to not be the case. So we can use Lehmann's algorithm to calculate W^* which runs in $\mathcal{O}(|\mathcal{N}|^3)$. Now the entry $W^*[A,B]$ is the probability of $p[A \overset{*}{\Rightarrow} B\alpha] = p_{lc}(B|A)$. This follows from the definition of W^* and the definition of the matrix product. We already proved in Assignment 3 exercise 2 b) that M^n encodes the sum of path lengths of length n in the graph M, this translates to our example where the path length will be the probabilities of a specific derivation path. Then it follows from the definition of W^* that $W^*[A,B]$ is the sum of probabilities of a derivation path from A to B of length n for n from 0 to ∞. This is the same as $p_{lc}(B|A)$. Now with all $p_{lc}(B|A)$ calculated we can calculate $p_{lc}(YZ|X) = \sum_{X' \in \mathcal{N}} p_{lc}(X'|X)p(X' \Rightarrow YZ)$ which is in $\mathcal{O}(|\mathcal{N}|)$. Doing this for all entries will take $\mathcal{O}(|\mathcal{N}|^4)$ as we have to calculate the sum for all possible X, Y and Z.

g)

$$p_{pre}(w_i, ..., w_k|X) \stackrel{!}{=} \sum_{j=1}^{k-1} \sum_{Y,Z \in \mathcal{N}} p_{lc}(YZ|X) \cdot p_{inside}(w_i, ..., w_j|Y) \cdot p_{pre}(w_j + 1, ..., w_k|Z)$$

$$\sum_{j=1}^{k-1} \sum_{Y,Z \in \mathcal{N}} p_{lc}(YZ|X) \cdot p_{inside}(w_i, ..., w_j|Y) \cdot p_{pre}(w_j + 1, ..., w_k|Z) \tag{41}$$

$$= \sum_{j=1}^{k-1} \sum_{Y,Z \in \mathcal{N}} p(X \stackrel{*}{\Rightarrow} YZ\alpha) \cdot p(Y \stackrel{*}{\Rightarrow} w_i, ..., w_j) \cdot p(Z \stackrel{*}{\Rightarrow} w_{j+1}, ..., w_k, \mathbf{v}) \tag{42}$$

$$\tag{43}$$

Further we have:

$$p(X \stackrel{*}{\Rightarrow} Y'Z'\alpha) \cdot p(Y' \stackrel{*}{\Rightarrow} w_i, ..., w_{j'}) \cdot p(Z' \stackrel{*}{\Rightarrow} w_{j'+1}, ..., w_k, \mathbf{v'}) \tag{44}$$

$$= p(X \stackrel{*}{\Rightarrow} w_i, ..., w_k, \mathbf{v}) \tag{45}$$

$$- \left(\sum_{j=1, j \neq j'}^{k-1} \sum_{Y,Z \in \mathcal{N}; Y \neq Y', Z \neq Z'} p(X \stackrel{*}{\Rightarrow} YZ\alpha) \cdot p(Y \stackrel{*}{\Rightarrow} w_i, ..., w_j) \cdot p(Z \stackrel{*}{\Rightarrow} w_{j+1}, ..., w_k, \mathbf{v}) \right) \tag{46}$$

$$\Rightarrow \sum_{j=1}^{k-1} \sum_{Y,Z \in \mathcal{N}} p(X \stackrel{*}{\Rightarrow} YZ\alpha) \cdot p(Y \stackrel{*}{\Rightarrow} w_i, ..., w_j) \cdot p(Z \stackrel{*}{\Rightarrow} w_{j+1}, ..., w_k, \mathbf{v}) \tag{47}$$

$$= p(X \stackrel{*}{\Rightarrow} w_i, ..., w_k, \mathbf{v}) \tag{48}$$

$$= p_{pre}(w_i, ..., w_k|X) \tag{49}$$

The important step here is equality of (45) and the previous equation. Which is valid as $p(X \stackrel{*}{\Rightarrow} w_i, ..., w_k, \mathbf{v})$ is the sum of the probabilities of all derivation trees of $w_i, ..., w_k, , \mathbf{v}$. Further the previous equation is the probability of deriving $w_i, ..., w_j$ from Y', $w_i, ..., w_k, \mathbf{v}$ from Z' and Z'Y'α from X. This corresponds to the probabilities of the trees seen in figure 1: b). Thus in (45) we calculate the difference between all possible derivation trees and the derivation trees that we are not intrested in i.e. the ones that do not the ones just mentioned.

h)

Using f) we can calculate the left corner probabilities in $\mathcal{O}(|\mathcal{N}|^4)$. After that we can calculate $p_{pre}(w_k|X)$ as follows:

$$p_{pre}(w_k|X) = p_{pre}(X \stackrel{*}{\Rightarrow} w_k) = \sum_{Y \in \mathcal{N}} p_{lc}(Y|X) p(Y \Rightarrow w_k) \tag{50}$$

This is possible in $\mathcal{O}(|\mathcal{N}|)$ as we can have at most one rule per nonterminal that derives the terminal w_k. Thus we can calculate $p_{pre}(w_k|X)$ for all $X \in \mathcal{N}$ in $\mathcal{O}(|\mathcal{N}|^2)$. We can also

compute the inside probabilities for all substrings of $w_i, ..., w_{k-1}$ and tags in $\mathcal{O}(N^3|\mathcal{R}|) = \mathcal{O}(N^3|\mathcal{N}|^3)$ using the CKY algorithm with the score function defined as the ln of probability of the derivation rule. We also return the whole chart instead of just $chart[1, N+1, S]$. Thus in $chart[j, i, X]$ we have the probability of deriving $w_j, ..., w_i$ from X.

Now we can calculate the prefix probabilities $p_{pre}(w_l, w_k|X)$ for all X for l from $k-1$ to i. We have all probabilities in the sums precomputed for $l = k-1$. For each we need to iterate over all j in range $[l, k-1]$ and all Y, Z. The outer sum iterates over at most N elements while the inner sum iterates over less than $|N|^2$ elements. Thus the entire sum will take at most $\mathcal{O}(N|\mathcal{N}|^2)$ to calculate. Calculating for all X gives us a runtime of $\mathcal{O}(N|\mathcal{N}|^2)$. From this follows that the whole suppart runs in $\mathcal{O}(N^2|\mathcal{N}|^3)$ as we need to compute this sum N times.

Then the whole algorithm runs in $\mathcal{O}(N^2|\mathcal{N}|^3) + \mathcal{O}(N^3|\mathcal{N}|^3) + \mathcal{O}(|\mathcal{N}|^4) = \mathcal{O}(\mathcal{O}(N^3|\mathcal{N}|^3) + \mathcal{O}(|\mathcal{N}|^4))$

Calling this algorithm for all prefixes of **w** then gives a runtime of $\mathcal{O}(N^4|\mathcal{N}|^3 + N|\mathcal{N}|^4)$.

i)

To make the previous algorithm more efficient we can instead use the CKY algorithm to calculate the prefix probabilities. The precomputations are the same as before, we calculate the left corner probabilities, the inside probabilities. We also now have to precompute the prefix probabilities for all of the terminals in **w** thus instead of $\mathcal{O}(|\mathcal{N}|^2)$ we get $\mathcal{O}(N|\mathcal{N}|^2)$. Next we use the CKY algorithm to calculate the rest of the prefix probabilities, $chart[n, n+1, X]$ will be initialized to $p_{pre}(w_n|X)$. Then given the previous levels of prefix probabilities we calculate the current level as follows:

$$chart[n, l, X] = \sum_{j=i}^{l-1} \sum_{Y,Z \in \mathcal{N}} p_{lc}(YZ|X) p_{inside}(w_n, ..., w_j|Y) chart[j+1, l, Z] \tag{51}$$

It is obvious that $chart[n, l, X] = p_{pre}(w_n, ..., w_l|X)$. We can prove this by strong induction. The base case of $l = n+1$ is trivial as we initialized the chart to hold the prefix probabilities. For our induction assumption we assume that $chart[i, j, X] = p_{pre}(w_i, ..., w_j|X)$ for all $i, j; j - i < l - n$ and all $X \in \mathcal{N}$. Then we can calculate $chart[n, l, X]$ as stated:

$$chart[n, l, X] = \sum_{j=i}^{l-1} \sum_{Y,Z \in \mathcal{N}} p_{lc}(YZ|X) p_{inside}(w_n, ..., w_j|Y) chart[j+1, l, Z] \tag{52}$$

$$= \sum_{j=i}^{l-1} \sum_{Y,Z \in \mathcal{N}} p_{lc}(YZ|X) p_{inside}(w_n, ..., w_j|Y) p_{pre}(w_{j+1}, ..., w_l|Z) \quad \text{(IH)} \tag{53}$$

$$= p_{pre}(w_n, ..., w_l|X) \quad (\text{def } p_{pre}) \tag{54}$$

Then as we have used the CKY algorithm the runtime is $\mathcal{O}(N^3|\mathcal{N}|^3)$. This gives us a runtime of $\mathcal{O}(N^3|\mathcal{N}|^3 + |\mathcal{N}|^4)$.

j)

$$p_{pre}(w) = \prod_{n=1}^{N} p(w_n|w_0, ..., w_{n-1}) \tag{55}$$

$$\Rightarrow p_{pre}(w_j|w_0, ..., w_{j-1}) = \frac{p_{pre}(w)}{\prod_{n=1, n\neq j}^{N} p(w_n|w_0, ..., w_{n-1})} \tag{56}$$

$$= \frac{p_{pre}(w)}{p_{pre}(w_0, ..., w_{j-1}) \frac{p_{pre}(w)}{p_{pre}(w_0, ..., w_{j+1})}} \tag{57}$$

$$= \frac{p(S \overset{*}{\Rightarrow} \mathbf{wv})}{p(S \overset{*}{\Rightarrow} w_0, ..., w_{j-1}\mathbf{v}) \frac{p(S \overset{*}{\Rightarrow} \mathbf{wv})}{p(S \overset{*}{\Rightarrow} w_0, ..., w_{j+1}\mathbf{v})}} \tag{58}$$

$$= \frac{p(S \overset{*}{\Rightarrow} w_0, ..., w_{j+1}\mathbf{v})}{p(S \overset{*}{\Rightarrow} w_0, ..., w_{j-1}\mathbf{v})} \tag{59}$$

$$= \frac{p_{pre}(w_0, ..., w_{j+1})}{p_{pre}(w_0, ..., w_{j-1})} \tag{60}$$