Title

Author

June 3, 2022

1 Section 1

Incomplete data log likelihood

$$
\begin{aligned}
\log \mathcal{L}(\theta) &= \ell(\theta) \\
&= \log(p(X|\theta)) \\
&= \log(\sum_{i=1}^{|S(X)|} p(S(X)_i|\theta)) \\
&= \log(\sum_{i=1}^{|S(X)|} \prod_{j=1}^{|S(X)_i|} p(T_{ij}|\theta))
\end{aligned}
$$

Where:

- X is the actual text random variable (RV). This is the only RV we observe.
- S is a deterministic function mapping X to the list of all tokenizations of X possible under a vocabulary $\mathcal{V}$. Note that the number of tokens in the k^{th} tokenization of X (i.e. $|S(X)_k|$) may or may not be equal to the number of tokens in a different tokenization of X. This is of course because the number of characters in each token of tokenizations can be much different.
- Each tokenization, $S(X)_i \quad i \in [1, |S(X)|]$, is a sequence of RVs, $T_{ij} \quad j \in [1, |S(X)_i|]$, which correspond to the indeces of the tokens in the vocabulary $\mathcal{V}$ that make up the tokenization. Thus, $T_{ij} \in [1, |\mathcal{V}|]$.

It is important to realize that T_{ij} is completely dependent on the observed variable X. In other words, if X is known then T_{ij} can be completely determinded by the use of the function S. However, what we need is to find the setting of θ to maximize this log-likelihood. In our tokenization model $p(T_{ij}|\theta) = \theta_{T_{ij}}$. If we view the T_{ij} variables as hidden variable, we can use EM to maximize this log-likelihood.

To use the EM algorithm, though, we must define a joint distribution $p(X, Z|\theta)$ and then derive the posterior distribution $p(Z|X, \theta)$. Our joint distribution will be very similar to the marginal $p(X|\theta)$, but walking through the details will help clarify the posterior.

$\mathbf{Z}$ is a sequence of latent token variables, $Z_j \ j \in [0, \infty]$. Each $\mathbf{z}_j$ is a scalar RV in $[1, |\mathcal{V}|]$ indicating which token in $\mathcal{V}$ was generated at timestep j. This differs from $S(X)_i$ in that $S(X)_i$ represents a unique full tokenization of the text X and can therefore only take on $|S(X)|$ values. Z_j on the other hand, is representing a single token and can therefore take on $|\mathcal{V}|$ values.

Complete data log likelihood

$$
\begin{aligned}
\log \mathcal{L}(\theta) &= \ell(\theta) \\
&= \log(p(X|\theta)) \\
&= \log(\sum_{i=1}^{|S(X)|} \prod_{j=1}^{|S(X)_i|} p(T_{ij}|\theta)) \\
&= \log(T_{i1} \sum_{i=1}^{|S(X)|} \prod_{j=2}^{|S(X)_i|} p(T_{ij}|\theta))
\end{aligned}
$$

Thus, the joint distribution $p(S(X)_i, \mathbf{Z}|\theta)$ is zero everywhere that $T_{ij} \neq Z_j$. As such,

$$
\begin{aligned}
\sum_{\mathbf{Z}} \sum_{i=1}^{|S(X)|} p(S(X)_i, \mathbf{Z}|\theta) &= \sum_{\mathbf{Z}} \sum_{i=1}^{|S(X)|} p(S(X)_i, \mathbf{Z} = S(X)_i)|\theta) \\
&= \sum_{j=1}^{\infty} \sum_{i=1}^{|S(X)|} p(T_i j, Z_j|\theta)
\end{aligned}
$$

and therefore,

$$
\begin{aligned}
\log \mathcal{L}(\theta) &= \log(\sum_i p(S(X)_i|\theta)) \\
&= \log(\sum_i \sum_{\mathbf{Z}} p(S(X)_i, \mathbf{Z}|\theta)) \\
&= \log(\sum_{\mathbf{Z}} p(S(X) = C(\mathbf{Z}), \mathbf{Z}|\theta)) \\
&= \log(\sum_{i,j} p(x_{ij} = C(\mathbf{Z})_{ij}, \mathbf{Z}_i|\theta))
\end{aligned}
$$

This merely states that when summing over all values of $\mathbf{Z}$ non-zero entries in the joint distribution will only occur when Z exactly corresponds to the original tokenizations derived when calculating $S(X)$.

As you may notice, the introduction of the hidden variable $\mathbf{Z}$ is merely a convenience. It serves to be precise in discerning between tokenization sequences vs. individual tokens, as we are now clearly summing over the time steps and tokens of $\mathbf{Z}$. It also allows for a clearer application of the EM algorithm since there are now observed (X) and hidden ($\mathbf{Z}$) variables.

2 Section 2

Lorem Ipsum