吴杉 08_06

1 计算entities概率的函数：getEntsPrblts()

标准式子：

$$
\begin{aligned}
&p(y_d = l | w, z_{\neg d}, t_{\neg d}, y_{\neg d}, \Gamma) \propto \frac{\alpha_l + |\{y_{d'} = l, d \neq d'\}|}{\sum_{k=1}^{K}(\alpha_k + \alpha_{df} + |D| - 1)} \\
&\prod_{i=1}^{N_d}\Bigg(\frac{\gamma_1 + |\{t_{d'j} = 0, d \neq d'\}|}{\gamma_1 + \gamma_2 + \sum_{d \neq d'} N_{d'}} \frac{\beta_{bg} + |\{t_{d'j} = 0, w_{d'j} = w_{di}, d \neq d'\}|}{|W|\beta_{bg} + |\{t_{d'j} = 0, d \neq d'\}|} \\
&+ \frac{\gamma_2 + |\{t_{d'j} = 1, d \neq d'\}|}{\gamma_1 + \gamma_2 + \sum_{d \neq d'} N_{d'}} \frac{\beta_l + |\{z_{d'j} = l, w_{d'j} = w_{di}, d \neq d'\}|}{|W|\beta_l + |\{z_{d'j} = l, d \neq d'\}|}\Bigg)
\end{aligned}
\tag{1}
$$

因为在同一个句子中许多变量都是相同的，在归一化是将被约去，所以简化了一部分分母：

$$
\begin{aligned}
&p(y_d = l | w, z_{\neg d}, t_{\neg d}, y_{\neg d}, \Gamma) \propto (\alpha_l + |\{y_{d'} = l, d \neq d'\}|) \cdot \\
&\prod_{i=1}^{N_d}\Bigg((\gamma_1 + |\{t_{d'j} = 0, d \neq d'\}|) \cdot \frac{\beta_{bg} + |\{t_{d'j} = 0, w_{d'j} = w_{di}, d \neq d'\}|}{\beta_{bg} + \frac{|\{t_{d'j}=0, d \neq d'\}|}{|W|}} \\
&+ (\gamma_2 + |\{t_{d'j} = 1, d \neq d'\}|) \cdot \frac{\beta_l + |\{z_{d'j} = l, w_{d'j} = w_{di}, d \neq d'\}|}{\beta_l + \frac{|\{z_{d'j}=l, d \neq d'\}|}{|W|}}\Bigg)
\end{aligned}
\tag{2}
$$

亦即:

$$
\begin{aligned}
&p(y_d = l | w, z_{\neg d}, t_{\neg d}, y_{\neg d}, \Gamma) \propto (\alpha_l + |\{y_{d'} = l, d \neq d'\}|) \cdot \\
&\prod_{i=1}^{N_d}\Bigg(\frac{\gamma_1 + |\{t_{d'j} = 0, d \neq d'\}|}{\gamma_2 + |\{t_{d'j} = 1, d \neq d'\}|} \frac{\beta_{bg} + |\{t_{d'j} = 0, w_{d'j} = w_{di}, d \neq d'\}|}{\beta_{bg} + \frac{|\{t_{d'j}=0, d \neq d'\}|}{|W|}} \\
&+ \frac{\beta_l + |\{z_{d'j} = l, w_{d'j} = w_{di}, d \neq d'\}|}{\beta_l + \frac{|\{z_{d'j}=l, d \neq d'\}|}{|W|}}\Bigg)
\end{aligned}
\tag{3}
$$

更显式的写为:

$$
\begin{aligned}
&p(y_d = ent) \propto (alpha\_ent + Cnt\_ent_{(w.r.t\ peudoDocs)}) \cdot \\
&\prod_{w \in pseudoDoc} \left(\frac{gamma\_bg + Cnt\_bg}{gamma\_df + Cnt\_df} \cdot \frac{beta\_bg + Cnt(bg, w)}{|W| * beta\_bg + Cnt\_bg} \right. \\
&\left. + \frac{beta\_ent + Cnt(ent, w)}{|W| * beta\_ent + Cnt\_ent_{(w.r.t\ words)}} \right)
\end{aligned} \tag{4}
$$

抽取一部分共同的式子:

$$
\begin{aligned}
comPart = (gamma\_bg + Cnt\_bg)/(gamma\_df + Cnt\_df) \\
/(beta\_bg + Cnt\_bg/|W|)
\end{aligned} \tag{5}
$$

得:

$$
\begin{aligned}
&p(y_d = ent) \propto (alpha\_ent + Cnt\_ent_{(w.r.t\ peudoDocs)}) \cdot \\
&\prod_{w \in pseudoDoc} \left(comPart \cdot (beta\_bg + Cnt(bg, w)) \right. \\
&\left. + \frac{beta\_ent + Cnt(ent, w)}{beta\_ent + Cnt\_ent_{(w.r.t\ words)}/|W|} \right)
\end{aligned} \tag{6}
$$

对此处的$|w|$的具体所指存疑:

$$
\begin{aligned}
comPart = (gamma\_bg + Cnt\_bg)/(gamma\_df + Cnt\_df) \\
/(beta\_bg + \frac{Cnt\_bg}{|W|_{(:contextNum?currentContextNum)}})
\end{aligned} \tag{7}
$$

注：这里的words都视为context words。

$$
\begin{aligned}
&p(y_d = ent) \propto (alpha\_ent + Cnt\_ent_{(w.r.t\ \ peudoDocs)})\cdot \\
&\prod_{w \in pseudoDoc} \Bigg(comPart \cdot (beta\_bg + Cnt(bg, w)) \\
&+ \frac{beta\_ent + Cnt(ent, w)}{beta\_ent + \frac{Cnt\_ent_{(w.r.t\ \ words)}}{|W|_{(:contextNum?entity.wordid_cnt.size())}}} \Bigg)
\end{aligned} \tag{8}
$$

ANS:context_words.

另外对于mention词，t必为1，所以对于此项word只需求:

$$
\frac{beta\_ent + Cnt(ent, w)}{beta\_ent + Cnt\_ent_{(w.r.t\ \ words)}/|W|} \tag{9}
$$

2 选择进一步增量学习数据的函数：isToLearn()

$selectMode\ \ ==\ \ 3$时，使用置信度评定$(> \eta)$:

$$
\frac{\max_i p(y_d = i)}{\sum_i p(y_d = i)} > \eta \tag{10}
$$

$$
\frac{\max_i e^{p(y_d = i)}}{\sum_i e^{p(y_d = i)}} > \eta \tag{11}
$$

3 对unlabeled used set进行采样，sampleOnce(indexes):

3.1 问题是是否要对labeled data重采Indicators。

ANS:不用。

4 计算Indicator的函数：getIndicatorPrblt()

标准式子：

$$
\begin{aligned}
\frac{p(t_{di}=0|w,z_{\neg d},t_{\neg d},y,\Gamma)}{p(t_{di}=1|w,z_{\neg d},t_{\neg d},y,\Gamma)} =& \frac{\gamma_1+|\{t_{d'j}=0,d'\neq d\}|}{\gamma_2+|\{t_{d'j}=1,d'\neq d\}|}\cdot \\
& \frac{\beta_{bg}+|\{t_{d'j}=0,w_{d'j}=w_{di},d'\neq d\}|}{\beta_{df}+|\{z_{d'j}=y_d,w_{d'j}=w_{di},d'\neq d\}|}\cdot \\
& \frac{|W|\beta_{df}+|\{z_{d'j}=y_d,d'\neq d\}|}{|W|\beta_{bg}+|\{t_{d'j}=0,d'\neq d\}|}
\end{aligned}
\tag{12}
$$

更显式的写为:

$$
\frac{p(t_{di}=0)}{p(t_{di}=1)}=\frac{gamma\_bg+cnt\_bg}{gamma\_df+cnt\_df}\cdot\frac{beta\_bg+cnt(bg,w_{di})}{beta\_df+cnt(ent,w_{di})}\cdot\frac{|W|beta\_df+cnt\_ent}{|W|beta\_bg+cnt\_bg}
\tag{13}
$$