Hypertext Retrieval and Mining (CS 610)
Homework 2
Out 2002-03-18, due 2002-03-22

The notation in McCallum's multi-label document classification paper can be a little confusing:

E-step:

$$\mathrm{P}(c|w \in d_i) \quad = \quad \frac{\vec{\lambda}_c^{(\vec{l}_i)}\mathrm{P}(w|c)}{\sum_{c'} \vec{\lambda}_{c'}^{(\vec{l}_i)}\mathrm{P}(w|c')} \tag{1}$$

M-step:

$$\vec{\lambda}_c^{(\vec{c})} \quad = \quad \frac{1 + \sum_{d_i \in \vec{c}} N(w, d_i)\mathrm{P}(c|w \in d_i)}{|\mathcal{C}| + \sum_{c' \in \vec{c}} \sum_{d_i \in \vec{c}} N(w, d_i)\mathrm{P}(c'|w \in d_i)} \tag{2}$$

$$\mathrm{P}(w_i|c) \quad = \quad \frac{\sum_{d \in \mathcal{D}} N(w_i, d)\mathrm{P}(c|w_i)}{\sum_{w' \in \mathcal{V}} \sum_{d \in \mathcal{D}} N(w', d)\mathrm{P}(c|w')} \tag{3}$$

We use the following notation:

$\mathcal{C}$	universe of classes
$C \subseteq \mathcal{C}$	a set of classes
d	a document
c, γ	a class label
C_d	set of class labels associated with d
t	a term
$n(d, t)$	number of times t occurs in d
$\lambda(c, C)$	selector probability for $c \in C$
$p(t\|c)$	class-conditional term probability, analogous to $\theta_{c,t}$
$p(c\|d, t)$	probability that c generated event (d, t)

The last three lines represent parameters that we will estimate. Complete the following mutually recursive equations for their estimation:

$$p(c|d, t) \quad = \quad \frac{\lambda(c, C_d)\boxed{}}{\sum_\gamma \boxed{}} \tag{4}$$

$$p(t|c) \quad = \quad \frac{\sum_d n(d, t)\boxed{}}{\boxed{}} \tag{5}$$

$$\lambda(c, C) \quad = \quad \frac{\sum_{d: C_d = C}\boxed{}}{\sum_\gamma \boxed{}} \tag{6}$$

Next, propose a formulation to approximate the prior for class set $C \subseteq \mathcal{C}$ from a training set in which there may not be any document labeled with class set C.

(**Project:** Implement your formulation and compare with naive Bayes and SVM classifiers.)