中文信息处理
隐马尔科夫模型，序列标注问题

李正华

苏州大学

2015 年 10 月 29 日

加 α 平滑：emission probability

平滑前：

$$e(w|t) = \frac{Count(w,t)}{Count(t)}$$

$$e(base|Vt) = \frac{Count(base,Vt)}{Count(Vt)}$$

平滑后：

$$e(w|t) = \frac{Count(w,t)+\alpha}{Count(t)+\alpha \times |\mathcal{V}|}$$

$$e(base|Vt) = \frac{Count(base,Vt)+\alpha}{Count(Vt)+\alpha \times |\mathcal{V}|}$$

$\mathcal{V}$ 表示训练数据中统计出来的词典，即不同词构成的一个集合。

$0 \leq \alpha \leq 1$; $\alpha = 1$ 又称为加一平滑，或 Laplace 平滑

编程作业：有监督的隐马尔科夫词性标注（共 15 分）

- 实现一个二元（一阶）隐马尔科夫模型，做词性标注任务（如果实现三元模型，分值可以适当增加）
- 在 train.conll 上使用极大似然估计方法确定模型参数
  - 使用加 α 平滑方法（你也可以使用或自己提出其他平滑方法），估计词性生成词的概率（发射概率）（3 分）
  - 直接估计词性转移概率（2 分）
- 实现 Viterbi 算法，对 dev.conll 进行词性标注（7 分）。
- 在 dev.conll 上评价模型的词性准确率（3 分）。

$$\textit{Tagging Accuracy} = \frac{\textit{\#words with correct tags}}{\textit{\#words in total}}$$

编程作业：无监督的隐马尔科夫词性标注（共 15 分）

- 实现一个二元（一阶）隐马尔科夫模型，做词性标注任务
- 在 train.conll 上使用极大似然估计方法确定模型参数：Hard EM（4 分）；Soft EM (8 分)；如果两个都完成，分数可以累加
  - 先从 train.conll 中统计词表，得到每一个词所有的可能词性，作为 EM 运行时的约束。
  - 迭代 100 次后停止，每一次迭代后报出目前数据的 log-likelihood
- 在 dev.conll 上评价模型的词性准确率（3 分）。
- 用 5 个不同的初始化种子，训练得到不同的模型，分别汇报准确率。

具体 EM 的相关介绍，包括前向后向算法，会尽快公布一个 pdf 讲义。