实验2. 隐马尔科夫模型实践Extra

MF1733071，严德美，1312480794@qq.com

2017 年 12 月 6 日

综述

对于已完成的维特比算法，可以用来做一些比较有意义的事情，比如说中文词性标注。由于词性标注有标注的语料，属于有监督的学习，可以通过最大似然估计统计出模型所需参数λ=[A,B,π]，本次词性标注任务使用人民日报1998年2月到6月的标注语料作为训练数据，人民日报1998年1月的标注语料作为测试数据，隐状态数目(词性数)有46个，对于输入已经分好词的句子输出对应的词性。

Extra.

对于有标注的语料，可以使用最大似然估计求出模型参数。

$$a_{ij} = P(s_j|s_i) = \frac{Number\ of\ transitions\ from\ state\ s_i\ to\ state\ s_j}{Number\ of\ transtion\ out\ of\ stae\ s_i}, \quad 1 \le i, j \le N \tag{1}$$

$$b_i(v_k) = P(v_k|s_i) = \frac{Number\ of\ times\ observation\ v_k\ occurs\ in\ state\ s_i}{Number\ of\ times\ in\ state\ s_i}, 1 \le i \le N, 1 \le j \le M \tag{2}$$

$$\pi_i = P(y_1 = s_i), \quad 1 \le i \le N \tag{3}$$

Result.

测试语料使用人民日报1998年1月的标注语料，评判标准是句子中词语的词性预测正确的数目，总词数1121447个，预测正确的词数839632个，正确率74.8%

```
F:\tools\anaconda\python.exe F:/codes/py_space/ml_project/HMM/tagger.py
839632 1121447 0.7487041295754503

Process finished with exit code 0
```

图 1: 人民日报1998年1月的新闻标注语料测试结果

以下手动输入句子来进行测试预测结果，比如对以下句子进行词性预测，[向/p 广大/b 职工/n 祝贺/v 新年/t ，/w 对/p 节日/n 坚守/v 岗位/n 的/u 同志/n 们/k 表示/v 慰问/v]左斜杠后面的表示词性，以下是预测结果：

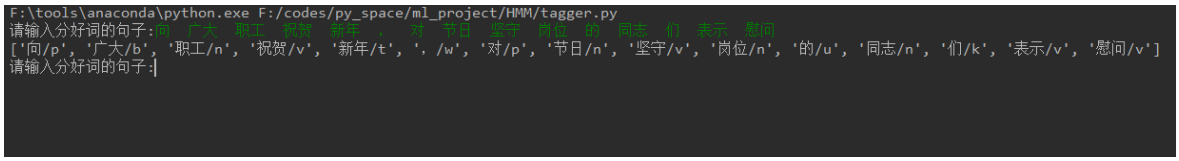

图 2: 输入句子进行预测