Linear Model（用于分类问题）讲课要点

李正华

2015 年 11 月 19 日

1 符号定义

$\mathcal{D} = \{S^j, Y^j\}_{j=1}^N$：表示一个数据集，包含 N 个句子和对应的 N 个人工标注的词性序列。

$S^j = w_1^j ... w_i^j ... x_{n_j}^j$：表示第 j 个句子，由 n_j 个词语组成。

$Y^j = y_1^j ... y_i^j ... y_{n_j}^j$：表示第 j 个句子对应的词性序列。

$\mathcal{T}$：表示词性集合，即隐状态的所有可能取值，$y_i^j \in \mathcal{T}$。

$\mathcal{V}$：表示词表 (vocabulary)，即数据 $\mathcal{D}$ 所有词语的集合，$w_i^j \in \mathcal{V}$。

2 将词性标注作为一个多元分类问题

注意，本节课内容和 HMM 不同，本节课中假设词语之间的词性相互独立预测（分类）。

一个分类任务为：给定一个句子 $S = w_0 ... w_n$ 和句子中的一个焦点词 w_i，要求确定其词性。如："我喜欢我的手机/NN"，即给定句子，要求模型预测第 4 个词 telephone 的词性。

线性模型中，定义句子 S 中的词 w_i 标为词性 t 的分值（不是概率）为：

$$\mathit{Score}(S, i, t) = \mathbf{w} \cdot \mathbf{f}(S, i, t) \tag{1}$$

其中，$\mathbf{f}(S, i, t)$ 表示将句子 S 中的词 w_i 标为词性 t 时对应的特征向量，$\mathbf{f}(.)$ 也可以看成一个特征抽取函数，返回一个特征向量（注意是稀疏特征向量，即只有很少的特征对应的值为 1，绝大多数都为 0）。请思考：如何表示稀疏特征向量？$\mathbf{w}$ 是特征权重向量，和 $\mathbf{f}(.)$ 的维度一致。每一个特征都有一个唯一个权重。由于模型之间利用特征权重向量和特征向量的点击，所以称为线性模型（Linear Model），即分值与特征向量之间为线性关系。

进而获得分值最高的词性：

$$t^* = \arg\max_{t \in \mathcal{T}} \mathit{Score}(S, i, t) \tag{2}$$

3 特征模板

在做分类的时候，我们希望把有用的信息都加入到模型中，这些信息即特征。特征模板的作用是描述我们想用哪些信息。表1给出了一个目前学术界比较常用的特征模板列表，即把 w_i 标为词性 t 时用到的特征集合。可以看到，我们主要用到了词信息，以及词内部的汉字信息。

	02: $t \circ w_i$
03: $t \circ w_{i-1}$	04: $t \circ w_{i+1}$
05: $t \circ w_i \circ c_{i-1,-1}$	06: $t \circ w_i \circ c_{i+1,0}$
07: $t \circ c_{i,0}$	08: $t \circ c_{i,-1}$
09: $t \circ c_{i,k}, 0 < k < \#c_i - 1$	
10: $t \circ c_{i,0} \circ c_{i,k}, 0 < k < \#c_i - 1$	
11: $t \circ c_{i,-1} \circ c_{i,k}, 0 < k < \#c_i - 1$	
12: **if** $\#c_i = 1$ **then** $t \circ w_i \circ c_{i-1,-1} \circ c_{i+1,0}$	
13: **if** $c_{i,k} = c_{i,k+1}$ **then** $t \circ c_{i,k} \circ$ *"consecutive"*	
14: $t \circ \text{prefix}(w_i, k), 1 \leq k \leq 4, k \leq \#c_i$	
15: $t \circ \text{suffix}(w_i, k), 1 \leq k \leq 4, k \leq \#c_i$	

表 1: POS tagging feature templates $\mathbf{f}(S, i, t)$。$\circ$ means string concatenation; $c_{i,k}$ denotes the k^{th} Chinese character of w_i; $c_{i,0}$ is the first Chinese character; $c_{i,-1}$ is the last Chinese character; $\#c_i$ is the total number of Chinese characters contained in w_i; prefix/suffix(w_i, k) denote the k-Character prefix/suffix of w_i.

4 特征、特征空间、特征向量

我们针对一些具体的训练实例，如："我喜欢我的手机/NN"，这个句子中"手机"标为NN，将特征模板实例化（instantialize）后，得到的一些具体的特征（字符串），称为**特征**。

所有训练数据 $\mathcal{D}$ 中出现的所有特征构成的集合，即特征空间，记为 $\mathcal{E}$，那么特征空间的维度为 $|\mathcal{E}|$，特征空间的维度一般很大，如百万级别。

对于一个实例，如"我喜欢我的手机/NN"，其对应（触发、根据模板实例化出）的特征集合记为 $B_{4,NN}$。那么这个集合也可以表示为一个特征向量的形式 $\mathbf{f}(S, 4, NN)$，这个向量的维度为 $|\mathcal{E}|$。这个向量中，大部分元素（下标）对应的值都为 0；只有那些包含在集合 $B_{4,NN}$ 中的元素为 1。所以，特征向量一般都非常稀疏。思考：如何表示稀疏的高维向量？

5 线性模型的编码流程

5.1 训练阶段

给定数据集 $\mathcal{D}$，设计好特征模板列表，训练阶段的结果是输出模型：一个特征集合（字符串）和每个特征的权重，格式如：

02:NN○ 手机　2
03:NN○ 我的　5
...
07:NN○ 手　1
...

前面是具体的一个特征，后面是特征对应的权重（如果特征权重为 0，可以不输出）。

1) 确定 feature space，即收集训练数据中所有的特征。

具体做法是，顺序处理 $\mathcal{D}$ 中的所有训练实例（instance）。对于每一个实例，根据特征模板列表，得到具体的特征，添加到特征空间（集合）中。

2) 估计特征权重向量 $\mathbf{w}$。

后面会介绍一种在线学习方法。

3) 输出模型（特征集合和权重），写到磁盘上。

5.2 测试阶段

目标是，给一个新的句子和其中的一个词，确定这个词的词性。

利用公式 (2) 求解。即尝试所有的词性分别构成特征向量，计算对应分值，最后选择分值最大的词性。

6 线性模型的训练过程

本部分介绍一种学术界常用的简单有效的训练方法，称为在线学习 Online training。其思想是，每次取一个训练实例，用当前模型进行预测，然后根据预测结果对特征权重进行更新。见算法1。

6.1 Averaged Perceptron

使用 **w** 作为特征权重向量，称为 “perceptron”；如果使用 **v** 作为特征权重向量，称为 “averaged perceptron”。一般来说，averaged perceptron 会比 perceptron 的性能更好，更稳定。大家可以比较一下。算法第10行需要对特征空间中的所有特征进行累加，效率很低。实际上，每次可以将绝大部分累加操作（权重无变化的特征）延迟进行，直到需要的时候再进行（如一次迭代结束需要用 **v** 处理开发数据集，或者一个特征的权重开始有了变化）。[1]

6.2 特征抽取优化

从表1可以看到，对于不同的词性，所有特征模板的后缀都是相同的。而在算法第7行，需要对每一个词性，都产生一个特征向量。这一步非常耗时。可以优化一下。基本想法是：

1. 特征模板中都不考虑词性，得到一个部分特征（partial feature）的空间（词典，或者集合，或者 map），记为 $\mathcal{G}$。每一个部分特征对应到一个唯一的数字（$[0, |\mathcal{G}| - 1]$）。

2. 构建一个词性的 map，每一个词性对应一个唯一的数字（$[0, |\mathcal{T}| - 1]$），如 NN 对应 0, VV 对应 1，等等。

3. 这样，我们可以认为特征空间的维度为：$|\mathcal{G}| \times |\mathcal{T}|$，每一个词性占用的维度大小为 $|\mathcal{G}|$。如词性 i 对应的维度为 $[\boldsymbol{offset}_i, \boldsymbol{offset}_i + |\mathcal{G}| - 1]$，其中 $\boldsymbol{offset}_i = i \times |\mathcal{G}|$。

4. 当处理一个实例时，如：“我喜欢我的手机/?”，我们先构造一个部分特征的稀疏向量，如 $[3, 101, 237, ..., 1722]$（已经由字符串映射为数字）。这样，对于不同的词性，我们在计算其对应分值时（点积），只需要对每一个元素加上一个词性对应的 offset 即可得到真正的特征编号（可以共用这个向量，不需要对每个词性创建一个向量）。

7 编程作业

编写一个 Linear Model，包括训练、测试、评价等程序。数据和 HMM 模型使用的数据（训练数据 train.conll、开发数据 dev.conll）相同。满分 10 分，单独写评价程序（即没有写 HMM 模型的代码）额外给 3 分。

[1] 提示：每一个特征可以对应一个时间戳信息，即这个特征最后一次更新发生在哪一个 k。

Algorithm 1 Online training.

Input: Labeled data $\mathcal{D}$, and Feature space $\mathcal{E}$
Output: Feature weights **w**
Initialization: $\mathbf{v} = \mathbf{0}, \mathbf{w}^0 = \mathbf{0}$, $k = 0$;
for $m = 1$ **to** M **do** {*iterations*}
  for $j = 1$ **to** N **do** {*for sentence* S^j}
    for $i = 1$ **to** n_j **do** {*for word* w_i^j}
      $t^* = \arg\max_{t \in \mathcal{T}} \mathbf{w} \cdot \mathbf{f}(S^j, i, t)$
      if $t^* \neq y_i^j$ **then** {y_i^j is the correct tag}
        $\mathbf{w}^{k+1} = \mathbf{w}^k + \mathbf{f}(S^j, i, y_i^j) - \mathbf{f}(S^j, i, t^*)$
        $\mathbf{v} = \mathbf{v} + \mathbf{w}^{k+1}$
        $k = k + 1$
      end if
    end for
  end for
  Here, after each iteration, we can evaluate the current model **w** on some validation (development) dataset.
  我们可以使用 **v** 作为特征权重，对开发数据集进行处理，评价得到对应的准确率，并和使用 **w** 做比较。
end for
Output **w**