NLP常用模型概览

在NLP当中，语言模型更加重要一些.不同语言模型的区别，也就是对文本提取特征的不同。常用的模型有一下几种。

BoW，最原始的特征集，一个单词/分词就是一个特征。往往一个数据集就会有上万个特征；有一些简单的指标可以帮助筛选掉一些对分类没帮助的词语，例如去停词，计算互信息熵等等，但不管怎么训练，特征维度都很大，每个特征的信息量太小；

Statistics，包括Term frequency(TF) , Inverse document frequency(IDF), 以及合并起来的TF-IDF。这种语言模型主要是用词汇的统计特征来作为特征集，每个特征都能够说得出物理意义，看起来会比bag-of-words效果好，但实际效果也差不多；

NGram，一种考虑了词汇顺序的模型，就是N阶Markov链，每个样本转移成了转移概率矩阵。也能取得不错的效果；

然而，这些特征的一个关键问题，在于并没有对样本特征进行很大程度的压缩，没有提取出关键的信息。这样的问题就是在分类问题中会出现overfit，也就是训练好的分类器只在训练它的数据集（例如用Facebook训练）中有效，换一个数据集（例如换成了Twitter）效果就会很差。

近期，机器学习界的一个研究热点，叫做稀疏表示（Sparse Representation）。也就是认为不管维度多么高的数据集，其实其关键特征就那么几个。这几个关键特征之间刻画了初噪声之外的全部关键信息，并且特征之间没有太多的相关性。

PSI、LDA就是文本的稀疏表示，代表的这一类语言模型叫做Topic Model。认为单词量再大的文本，其文章主题就那么几个。一个K个主题的LDA模型，可以把一个文本压缩成K维的向量：每一个维度就是该文本属于该主题的概率，这个向量也叫做Topic Proportion（注意和Topic Distribution区别）。然后得到压缩后的K维数据集后，再使用任何的分类器，甚至最简单的余弦相似性指标，都可以得到非常漂亮的分类效果。

NLP模型详情

1. Bag-of-Words
   * 统计模型，抛开了文本数据中的上下文信息以及语法语义，单纯的从词频的分布来描述一个文本数据，简单，但不能解决复杂问题。
2. TF-IDF
   * TF指词频，IDF指单词作为特征的文本区分度，用于给词频特征（BoW）加权，能够让更具有代表性的词特征作用更显著。
3. N-Gram
   * 统计语言模型的一种，本质上是一种N-1阶的Markov chain。对于二元模型，p(wi|w1…wi-1) = p(wi|wi-1)=p(wi,wi-1)/p(wi-1)，通过概率公式的推导，可以发现最终p(wi|wi-1)=p(wi,wi-1)/p(wi-1)=Count(wi,wi-1)/Count(wi-1)，N元模型以此类推。这里说的N-Gram其实就是指长度为N的词序列可能出现的所有情况，在现实的语料中，所有可能出现的N-Gram包括两部分：语料中可观测到的，观测不到的，因为现实N-Gram的数量是V的N次方，V是指词典大小，因此现实语料中很可能会出现观测不到的情况，那么就会导致零概率的情况发生，通过平滑方法可以解决零概率问题。N越小，出现零概率的情况越少，N等于1的时候表示这是一个上下文无关的统计语言模型。
4. LDA
5. 对数线性模型
6. 最大熵模型
7. HMM
   * 应用：词类标注、语音识别、局部句法剖析、语块分析、命名实体识别、信息抽取等。应用于自然科学、工程技术、生物科技、公用事业、信道编码等多个领域。
   * HMM原理是通信模型和马尔科夫理论的结合。通信模型是指通过以观察到的一组消息序列o10203….，求出是的P(s1s2s3…|o1o2o3…)最大的s1s2s3…输入序列利用贝叶斯公式，可以推导出P(s1s2s3…|o1o2o3…)=P(o1o2o3…|s1s2s3…)\*P(s1s2s3…)，因为序列o1o2o3…是确定事件，因此P(o1o2o3…)=1.

HMM又叫做隐马尔科夫模型，是基于马尔科夫链的扩展，马尔科夫模型的主要参数是节点间状态的转移概率，有了转移概率和随机变量的域，通过任意一个随机的初始状态都能得到一组随机变量序列，而我们则可以通过转移概率计算出每个序列出现的概率。而这样的转移概率，我们可以通过大量的观测数据，利用统计的方法计算出各种状态转移条件的频率，根据大数定律得出转移概率，这就是马尔科夫模型训练的过程。隐马尔科夫模型中的‘隐’字的意思是我们无法直接观测到每个节点的状态，也就是s1s2s3…是观察不到的，因此也无法去通过统计的方法推算转移概率了，然而HMM每个时刻t都会输出一个符号ot，而且ot只和st相关，这个叫做独立输出假设，于是s1s2s3…变成了隐含状态，而o1o2o3…则变成了观测值，于是我们就把HMM和通信模型对应在了一起。

HMM有三个基本问题：

* 给定一个模型，如何计算出特定输出序列的概率（Forward-Backward算法）
* 给定一个模型和某个特定的输出序列，如何找到最可能产生这个输出的状态序列（维特比算法）
* 给定足够量的观测数据，如何估计隐马尔科夫模型的参数即转移概率矩阵（最大期望算法，鲍姆-韦尔奇算法，EM算法）

HMM中的两个概率：转移概率，生成概率（发射概率）

1. CRF
   * 条件随机场用于序列标注，中文分词、中文人名识别和歧义消解等自然语言处理中，表现出很好的效果。原理是：对给定的观察序列和标注序列，建立条件概率模型。条件随机场可用于不同预测问题，其学习方法通常是极大似然估计。
   * CRF是HMM的对数概率形式，实际上是一种对数线性模型
   * 不同于HMM，CRF引入了特征函数的概念。在HMM中一般情况下只会使用到当前词和相邻的前一到两个词之间的特征，并且一般只使用词性作为特征（方便计算？）。而CRF则可以使用多种类型的特征，如词的位置，词的形状，词根等。有一种说法是，把CRF中每个词对应的各种类型的特征通过统计求得其概率分布（HMM的训练过程）后对它们求平均，作为新的转移概率，其实这样跟HMM没有任何本质上的区别。所以这里还是以特征函数的概念为准，最基本的特征函数，满足该特征，加一分，不满足零分，对一个序列中所有词的所有特征函数的结果求和，分数最高的则为我们想要的标注序列。
   * 由于存在特征函数的概念，CRF算法中又多了一步如何选取特征函数的步骤。
   * CRF（同HMM）模型的训练方法：极大似然估计和正则化的极大似然估计，具体的优化实现算法有IIS负责均衡算法、梯度下降算法和你牛顿法BFGS（一种优化效果最好的拟牛顿法）
   * CRF（同HMM）模型的预测算法：CRF的预测原理是根据训练好的CRF模型和一组可观测的输出序列，求最可能产生这组输出序列的输入序列，即标注序列。其中著名的预测算法是维特比算法
2. LSTM