清晰易懂的条件随机场原理总结

原创 石头 机器学习算法那些事 1月1日

如果问机器学习初学者，《统计学习方法》中最难理解的章节是什么？我想大部分人的回答是条件随机场。小编前段时间看了很多条件随机场的学习资料，整理出了这篇文章，请大家耐心看，初学者应该也能看懂。

目录

1. 一例说明条件随机场是什么

假设你有贾斯丁·比伯一天生活的照片，你想要给每一张照片贴上一个标签，比如吃饭，舞蹈，睡觉，唱歌，驾驶等，该如何做？

一种方法是忽视照片的时间顺序特性，照片之间是相互独立的，训练数据有大量的照片和对应的标签，构建分类模型。比如训练数据包含近一个月的标签照片，你的分类模型可能会认为早上6点拍摄的黑色的照片是与睡眠相关的，有明亮色彩的照片往往与舞蹈相关，有汽车的照片与驾驶相关等等。

这种忽视时间顺序特性的方法会损失很多信息，比如，如果你看到一张嘴的特写照片，标签是唱歌还是吃饭？如果你考虑时间的顺序特性，假设前一张照片是吃饭或烹饪，那么这张照片的标签很可能是吃饭；若前一张照片是唱歌或舞蹈，那么这张照片的标签很可能是也唱歌。

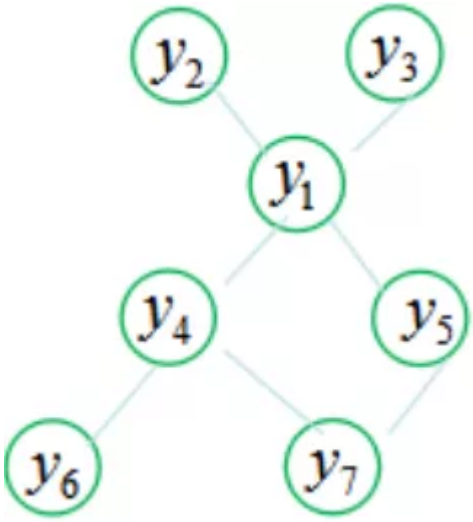
因此，为了提高标签的准确性，我们应该考虑邻近照片的标签，**这种方法就是条件随机场**。

2.条件随机场的定义以及应用场景

条件随机场的应用场景是给定输入的随机变量$X=\{x_1,x_2,\ldots,x_n\}$，预测随机变量$Y=\{y_1,y_2,\ldots,y_n\}$，当输出随机变量是离散值时，应用领域就是我们熟知的词性标注和语音识别，是不是和隐马尔科夫模型很相似，后续章节会分析两者的区别，本节先不介绍。

理解条件随机场的定义需要理解两个重要的知识点：（1）邻近，（2）马尔科夫随机场

1）如何理解“邻近”这一含义，请看如下的概率无向图模型：



变量y_1的邻近点是y_2，y_3，y_4和y_5，变量y_4的邻近点是y_1和y_6，变量y_6的邻近点是y_4。邻近点的含义是用无向边相连，存在相关的两个随机变量。

2）马尔科夫随机场的本质是概率无向图，之所以叫马尔科夫随机场的原因是随机变量间满足成对马尔科夫性、局部马尔科夫性和全局马尔科夫性，马尔科夫性是关于条件独立的一种方法。

成对马尔科夫性： u和v是无向图G中任意两个没有边连接的结点，对应的随机变量为Y_u和Y_v，其他所有结点为O，对应的随机变量组是Y_O。成对马尔科夫性是指给定随机变量组Y_O的条件下，随机变量组Y_u和Y_v是相互独立的，即：

$$P(Y_u,Y_v \mid Y_O)=P(Y_u \mid Y_O)P(Y_v \mid Y_O)$$

如上图，y_2和y_6是无向图G中任意两个没有边连接的结点，其他所有结点为O，对应的随机变量组是$y_O = \{y_1, y_3, y_4, y_5\}$，由成对马尔科夫原理，得：

$$P(y_2, y_6 \mid y_O) = P(y_2 \mid y_O)P(y_6 \mid y_O)$$

局部马尔科夫性：假设v是无向图G中任意一个结点，W是与v“邻近”的所有结点，O是v与W以外的所有结点。v表示的随机变量是Y_v，W与O表示的随机变量组分别是Y_W和Y_O。局部马尔科夫性是指给定随机变量Y_W的条件下，随机变量Y_v与随机变量组Y_O是相互独立的，即：

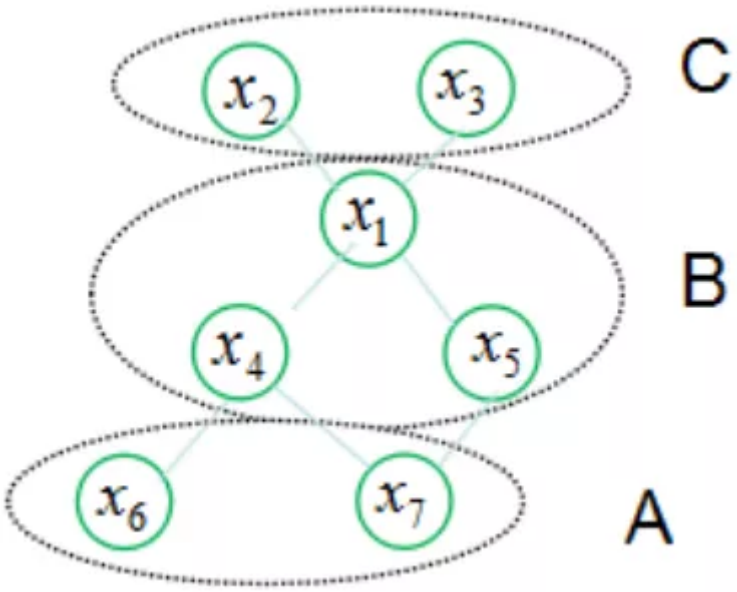
$$P(Y_v, Y_O \mid Y_W) = P(Y_v \mid Y_W)P(Y_O \mid Y_W)$$

如上图，y_1是无向图G中的一个结点，与结点y_1邻近的点为随机变量组$y_W = \{y_2, y_3, y_4, y_5\}$，无向图中$y_1$与$y_W$以外的结点是随机变量组$y_O = \{y_6, y_7\}$，由局部马尔科夫原理，得：

$$P(y_1, y_O \mid y_W) = P(y_1 \mid y_W)P(y_O \mid y_W)$$

全局马尔科夫性：结点集合A，B，C是无向图G中任意的结点集合，且各集合无交集，对应的随机变量组是Y_A，Y_B和Y_C。

如下图：



全局马尔科夫性是给定随机变量组Y_B的条件下，随机变量组Y_A和Y_C是相互独立的。

$$P(Y_A, Y_C \mid Y_B) = P(Y_A \mid Y_B)P(Y_C \mid Y_B)$$

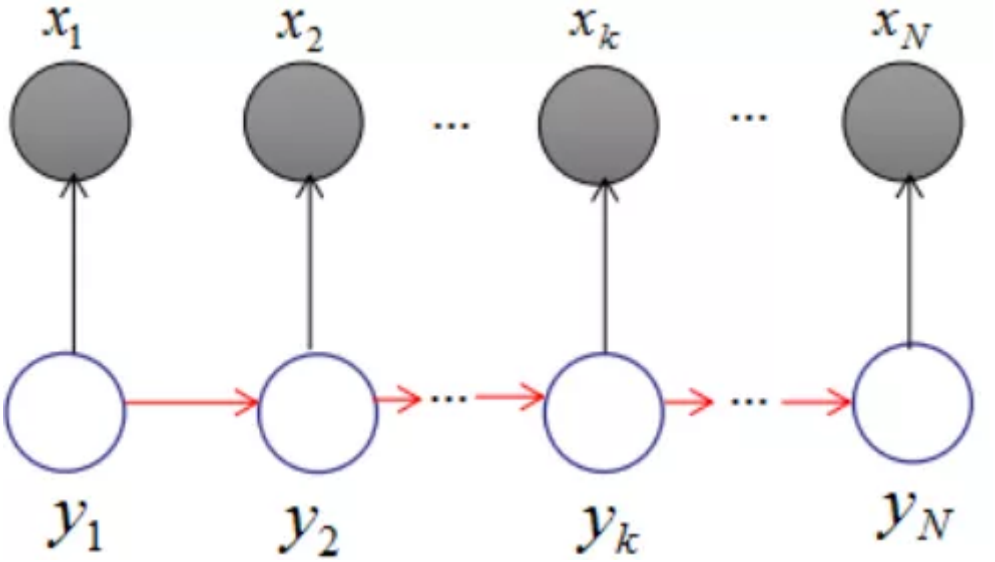
结合这两个知识点，我们给出条件随机过程的定义：给定随机变量组X，随机变量组Y符合马尔科夫随机场，则称条件概率分布$P(Y \mid X)$为条件随机场。

由马尔科夫随机场的局部马尔科夫性质可得：

$$P(Y_v \mid X, Y_w, w \neq v) = P(Y_v \mid X, Y_w, w \sim v)$$

其中$w \neq v$表示节点v以外的所有结点，$w \sim v$表示与节点v邻近的结点。

更进一步的介绍，若结点Y_i的邻近点只有Y_{i-1}和Y_{i+1}，且随机变量组X和随机变量组Y有相同的图结构，则称条件概率分布$P(Y \mid X)$为线性链条件随机场。如下图为线性链条件随机场：



由马尔科夫随机场的局部马尔科夫性质可得：

$$P(Y_i \mid X, Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n) = P(Y_i \mid X, Y_{i-1}, Y_{i+1})$$

小结：这节内容有点多，小小的总结一下，首先理解马尔科夫随机过程的两个概念：（1）邻近，（2）马尔科夫性质。然后条件随机场是给定随机变量组X的条件下，随机变量组Y构成一个马尔科夫随机场，最后我们用马尔科夫随机场的马尔科夫性质去推导条件概率分布$P(Y \mid X)$。

3.条件随机场举例——词性标注

本节我们通过词性标注（part-of-speech）来说明条件随机场的算法过程。

词性标注的目标是给一个句子（包含一系列单词和符号）加上词性标注如形容词（adjective），名词（noun），介词（preposition），动词（verb），副词（adverb），冠词（article）。

举个例子，“Bob drank coffee at Starbucks”，句子标注可能是“Bob（名词）drank（动词）coffee（名词）at（介词）Starbucks（名词）“

所以让我们构建一个条件随机场来标记句子的词性，与其他分类器一样，我们首先需要确定一组特征函数f_i。

3.1 条件随机场的特征函数

条件随机场的一组特征函数类似于先验概率，根据自己的先验知识定制一套规则，每个特征函数的输入有：

1）一个句子s

2）句子第i个位置的单词

3）当前单词的标签l_i

4）前一个单词的标签l_{i-1}

特征函数的输出是实数值。由特征函数的输入可知，句子单词的标签只和邻近的前一个标签相关，因此这是一种特殊的线性链条件随机场，常规的线性链条件随机场的标签和前后两个标签相关，若是更常规的条件随机场，那么标签和邻近的标签相关，邻近的含义可参考第2节的介绍。

3.2 特征函数与概率的转化

接下来，给每个特征函数分配一个权重（我将在后面讨论如何从训练数据学习权重），给定一个句子s，我们通过把句子中所有单词的特征加权求和，得到句子s的标签（l）分数。有：

$$score(l \mid s) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1})$$

上式第一个积分表示每个单词在所有特征函数的求和，共有m个特征函数；第二个积分表示对句子的每个单词求和，共有n个单词。

词性标注本质上是多分类，我们把每个标签序列的分数转化为概率进行分类，实现方法是对分数求指数后再进行归一化转化为概率，范围为0~1，即：

$$p(l \mid s) = \frac{\exp[score(l \mid s)]}{\sum_{l'} \exp[score(l \mid s)]} = \frac{\exp[\sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1})]}{\sum_{l'} \exp[\sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i', l_{i-1}')]}$$

由上式可知，给定一个句子s，词性标注l的分数越高，则相应的条件概率也越大，选择条件概率最大的词性标注l为句子s的标注序列。

3.3 词性标注的特征函数举例

如何根据自己的先验知识去设置词性标注的特征函数呢？下面例举几个词性标注的特征函数：

$$f_1(s,i,l_i,l_{i-1})=1$$

当l_i是副词且第i个单词以”-ly”结尾时，我们就让特征函数$f_1=1$，反之为0；如果该特征函数的权重λ_1是正数且数值越大时，我们认为以“-ly”结尾的单词是副词的可能性越大。

$$f_2(s,i,l_i,l_{i-1})=1$$

当第一个单词是动词且这个句子以问号结尾时，我们就让特征函数$f_2=1$，反之为0；如果该特征函数的权重λ_2是正数且数值越大时，我们认为句子以分号结尾时，第一个单词是动词的可能性越大。

$$f_3(s,i,l_i,l_{i-1})=1$$

当l_{i-1}是形容词且l_i是名词时，我们就让$f_3=1$，反之为0；如果该特征函数的权重λ_3是正数且数值越大时，我们认为形容词后面是一个名词的可能性越大。

$$f_4(s,i,l_i,l_{i-1})=1$$

当l_{i-1}是介词且l_i是介词时，我们就让$f_4=1$，反之为0；如果该特征函数的权重是λ_4负数且绝对值越大时，我们倾向于认为介词后面是一个介词的可能性越小。

这些特征函数的设置符合我们对语句的先验常识，若特征函数与权重的乘积能够增加$score(l/s)$，则该特征函数的词性标注越可能实现；反之可能性越小。

小结：为了构建条件随机场，你只需要定义一串特征函数（特征函数的输入为整个句子，当前字段以及相近的词性标注），分配特征函数相应的权重，并将它们乘积求和，然后转化为概率的形式求最优标注序列。

若我们确定了特征函数，还需要知道相应的权重，当特征函数和权重参数确定时，模型也就确定下来了，如何通过训练数据学习模型的权重参数，这一过程为条件随机场的学习算法。

3.4 条件随机场的学习算法

回顾之前的内容，给定句子s，标注序列l的条件概率为：

$$p(l|s)=\frac{\exp[score(l|s)]}{\sum_{l'}\exp[score(l|s)]}=\frac{\exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i,l_{i-1})]}{\sum_{l'}\exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i^{'},l_{i-1}^{'})]}$$

假设我们有大量训练数据（每个训练数据包含语句和相应的词性标注），随机初始化条件随机场模型的权重参数，我们通过训练数据逐渐优化权重参数，使权重参数往最优方向移动，这一算法实现是我们常用的梯度上升算法：

1）取条件概率的对数，即 $\log p(l|s)$

$$\log p(l|s)=\log\left(\frac{\exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i,l_{i-1})]}{\sum_{l'}\exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i^{'},l_{i-1}^{'})]}\right)$$
$$=\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i,l_{i-1})-\log\left(\sum_{l'}\exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i^{'},l_{i-1}^{'})]\right)$$

2）令 $\log p(l|s)$ 对每一个特征函数求梯度，得：

$$\frac{\partial\log p(l|s)}{\partial\lambda_j}=\frac{\partial\left(\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i,l_{i-1})\right)}{\partial\lambda_j}-\frac{\partial\left(\log\left(\sum_{l'}\exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i^{'},l_{i-1}^{'})]\right)\right)}{\partial\lambda_j}$$
$$=\sum_{i=1}^{n}f_j(s,i,l_i,l_{i-1})-\sum_{l'}\frac{\exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i^{'},l_{i-1}^{'})]}{\sum_{l'}\exp[\sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j(s,i,l_i^{'},l_{i-1}^{'})]}*\sum_{i=1}^{n}f_j(s,i,l_i^{'},l_{i-1}^{'})$$
$$=\sum_{i=1}^{n}f_j(s,i,l_i,l_{i-1})-\sum_{l'}p(l^{'}|s)*\sum_{i=1}^{n}f_j(s,i,l_i^{'},l_{i-1}^{'})$$

注意到等式第一项是训练数据真实标注下特征函数对梯度的贡献，等式第二项是当前模型下特征函数对梯度贡献值的期望，对数条件概率的梯度是真实标注下特征函数与当前标注下特征函数期望的差值。现在你应该知道该梯度的含义了吧，即梯度上升是当前标注下特征函数值增加最快的方向。

3）梯度上升的方向迭代权重参数 λ_i：

$$\lambda_j = \lambda_j + \alpha \frac{\partial \log p(l \mid s)}{\partial \lambda_j}$$

其中α是学习率。

$$\lambda_j = \lambda_j + \alpha \left[\sum_{i=1}^{n} f_j(s, i, l_i, l_{i-1}) - \sum_{l'} p(l' \mid s) * \sum_{i=1}^{n} f_j(s, i, l'_i, l'_{i-1}) \right]$$

4）当对数条件概率$\log p(l \mid s)$不再增加时或增加的值小于给定的阈值时，迭代结束；反之重复之前的步骤。

现在我们知道了特征函数和权重参数，模型也相应的确定。如何对输入的语句进行词性标注，这一过程称为条件随机场的预测算法。

3.5 条件随机场的预测算法

条件随机场的预测问题是给定条件随机场$p(l \mid s)$和输入语句s'，求条件概率最大的词性标注序列l'。

最简单的方法是用直接法预测最优标注序列，算法思想是列出所有可能的标注序列，然后选择概率最大的标注序列l'。这种方法虽然简单，但是计算量非常复杂，如一个语句有n个单词，每个单词的可能词性标注有k种，那么所有可能的标注序列就有k^n中。

与隐马尔可夫模型的预测算法一样，条件随机场的预测算法是著名的维特比算法，《统计学习方法》一书关于维特比算法的例子讲的挺详细的，这里就不再详细介绍了。

至此，条件随机场算法已基本介绍完了，下面开始比较条件随机场与其他常见的机器学习算法。

4. 与逻辑斯蒂克回归算法比较

K分类的多项逻辑斯蒂回归算法：

$$P(Y=k|x)=\frac{\exp(w_k\cdot x)}{1+\sum_{k=1}^{K-1}\exp(w_k\cdot x)},k=1,2,...,K-1$$

结合分母，得：

$$P(Y=k|x)=\frac{\exp(w_k\cdot x)}{\sum_{k=0}^{K-1}\exp(w_k\cdot x)},k=1,2,...,K-1$$

是不是和条件随机场$p(l|s)$的定义有点像？这是因为条件随机场是多个顺序随机变量的逻辑斯蒂回归。逻辑斯蒂回归是用于分类的对数线性模型，而条件随机场是用于标签序列的对数线性模型。

5. 与隐马尔可夫模型的比较

隐马尔科夫模型也可用于词性标注，条件随机场通过特征函数得到标签序列的分数，分数越高表示标签序列出现的可能性越大；隐马尔科夫模型采用生成式的标签方法，定义：

$$p(l,s)=p(l_1)\prod_i p(l_i|l_{i-1})p(s_i|l_i)$$

其中$p(l_i|l_{i-1})$是转移概率（比如介词转移为名词的概率），$p(s_i|l_i)$是观测概率（比如名词为“Dad”的概率）。

条件随机场不仅包含了隐马尔科夫模型，且能处理隐马尔科夫模型不能处理的情况。下面给出解释：

取隐马尔科夫联合概率的对数，得：

$$\begin{aligned}\log p(l,s)&=\log p(l_1)\prod_i p(l_i|l_{i-1})p(s_i|l_i)\\&=\log p(l_1)+\log\prod_i p(l_i|l_{i-1})+\log\prod_i p(s_i|l_i)\end{aligned}$$

如果我们把上式的对数条件概率看作是条件随机场的权重系数，那么上式等价于条件随机场，只要特征函数满足以下条件：

1）对于每个隐马尔科夫模型的转移概率$p(l_i = \mathrm{y} \mid l_{i-1} = x)$，如果$l_{i-1} = x$且$l_i = y$，那么条件随机场的转移特征函数为$\mathrm{f}_{\mathrm{x,y}}(s,i,l_{i-1},l_i) = 1$，每个特征函数的权重系数$\mathrm{w}_{x,y} = \log p(l_i = \mathrm{y} \mid l_{i-1} = x)$

2）对于每个隐马尔科夫模型的观测概率$p(s_i = \mathrm{z} \mid l_i = x)$，如果$l_i = x$且$s_i = \mathrm{z}$，那么条件随机场的观测特征函数为$\mathrm{g}_{\mathrm{x,y}}(s,i,l_{i-1},l_i) = 1$，每个特征函数的权重系数：

$$\mathrm{w}_{x,z} = \log p(s_i = \mathrm{z} \mid l_i = x)$$

因此，条件随机场计算的分数$p(l \mid \mathrm{s})$与隐马尔科夫模型等价，每个隐马尔科夫模型都可以用某些条件随机场表示。

然而，条件随机场可以预测更丰富的标签序列，有下面两个原因：

1）条件随机场的标签分布是马尔科夫随机场，因此可以构建更复杂的特征函数，根本原因在于马尔科夫随机场的邻近节点有多个，而隐马尔科夫模型标签的邻近点只有两个。

2）隐马尔科夫模型的观测概率必须满足下面条件：

$$0 \le p(s_i \mid l_i) \le 1$$

$$\sum_{\mathrm{w}} p(s_i = w \mid l_1) = 1$$

而条件随机场的权重系数是无限制的。

基于上面两个原因，马尔科夫随机场可以预测更丰富的标签序列。

6.小结

条件随机场包含三个重要知识点：邻近节点的含义，马尔科夫随机场（概率无向图），特征函数（与先验知识相关）。理解了这三个重要知识点，就不难理解条件随机场的定义了，条件随机场是在给定输入序列X的条件下，标签序列Y符合马尔科夫随机场。