三、7 最大熵模型

2016年3月13日 星期日　　上午10:58

一、简介

最大熵模型（maximum entropy model）。

是一种分类模型，思想是：在满足现有观测练数据情况下，让未知的变量熵最大。现有的观测数据就是训练数据，未知变量一般是模型参数等。

二、思想

熵

是一种不确定性的度量、混乱的程序。信息论中的熵，衡量每条消息中包含的信息量，越随机的信息源熵越大。又被称为信息熵、信源熵、平均自信息量。

熵，来源于热力学，在热力学中，代表热力学系统无序的程度。

信息熵 $H(P) = -\Sigma P(x) \log P(x)$

最大熵原理

简介

最大熵原理是概率学习的一个准则，认为在学习概率模型的时候，所有可能的概率分布中，熵最大的模型是最好的模型。

解释

可能的概率分布（利用观测数据）

首先“所有可能的概率分布”是我们通过观测数据解出来的，可以看做一个模型簇，这个簇中每个模型都满足观测数据的要求，但由于仍有一部分变量未确定，所有是一个模型簇（而不是一个确定的模型）。换句话说，

这些未确定的变量，可能通过观测数据无法求解。那么，我们就得利用最大熵原理对其进行估计。

未知的参数估计（利用最大熵原理）

最大熵原理其实很简单，就是对于未知变量，在没有任何额外信息的情况下，他们应该是“等可能的”，即服从均匀分布的。但是“等可能”无法很好的用数学的形式表达，所以求助于“熵”，我们让熵最大，其实就是让未知变量取均匀分布。

我们可以推出“均匀分布”和“熵最大”是等价的：信息熵 H（P）= -Σ P（x）logP（x），容易看出熵的上界：H（P）<= log |X|，|X|是X取值的个数，可以推出当X均匀分布的时候（P（x）=C），熵取到上界，即熵最大。

应用

把最大熵原理应用到分类问题，得到最大熵模型。

二、原理

问题定义

这是一个分类任务，目标是估计条件概率分布P（Y|X），X是输入，Y是输出。给定数据集（xi，yi），用最大熵原理，选择出最好的分类模型。

相关变量

给定数据集，我们可以求出训练集上的联合分布P（X,Y），P（X），训练集上的这些分布是经验分布，记为P~（X,Y），P~（X），可以直接求得

$$\tilde{P}(X=x,Y=y)=\frac{\nu(X=x,Y=y)}{N}$$

$$\tilde{P}(X=x)=\frac{\nu(X=x)}{N}$$

$$P(X=x)=\frac{\quad}{N}$$

定义特征函数f（x，y），这个函数是二值函数（值域0、1），用于描述x与y之间的事实。

$$f(x,y)=\begin{cases}1, & x与y满足某一事实\\ 0, & 否则\end{cases}$$

模型思想的表达式

根据最大熵原理，我们需要：A、符合训练集上的分布 B、满足A基础上，熵最大

符合训练集

这里也不是要让训练数据都分对，而是让期望相同：对于二值函数f（x，y），在经验分布P（x，y）熵的期望与在模型P（Y|X）上的期望相同，经验分布也就是实际训练集的分布。

经验分布上的期望

$$E_{\tilde{p}}(f)=\sum_{x,y}\tilde{P}(x,y)f(x,y)$$

模型上的期望

$$E_{P}(f)=\sum_{x,y}\tilde{P}(x)P(y|x)f(x,y)$$

熵最大

条件概率P（Y|X）上的条件熵

$$H(P)=-\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)$$

最终表达式

满足

$$\sum_{x,y}\tilde{P}(x)P(y|x)f(x,y)=\sum_{x,y}\tilde{P}(x,y)f(x,y)$$

$$H(P)=-\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)$$

让H（P）最大

三、求解思路

很明显变成了一个约束最优化问题，可以仿照SVM的解法。（二、中最终表达式也可以写错这样）

$$\max_{P\in C} \quad H(P)=-\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)$$

$$\text{s.t.} \quad E_P(f_i)=E_{\tilde{p}}(f_i),\quad i=1,2,\cdots,n$$

$$\sum_y P(y|x)=1$$

注意，这里P指的是P（Y|X），这里引入n，n是指特征函数f（x，y）的个数，相当于有多少个特征函数就有多少个约束条件。

分为以下几步

1、按照习惯，改为等价求最小值问题

（这仅仅是为了后面推导方便，没什么特殊的）

$$\min_{P\in C} \quad -H(P)=\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)$$

2、通过拉格朗日乘子法，转化为无约束问题

$$L(P,w)\equiv -H(P)+w_0\left(1-\sum_y P(y|x)\right)+\sum_{i=1}^{n}w_i(E_{\tilde{p}}(f_i)-E_P(f_i))$$

$$=\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)+w_0\left(1-\sum_y P(y|x)\right)$$

$$+\sum_{i=1}^{n}w_i\left(\sum_{x,y}\tilde{P}(x,y)f_i(x,y)-\sum_{x,y}\tilde{P}(x)P(y|x)f_i(x,y)\right)$$

（这里引入参数w）

原始问题转化为

$$\min_{P\in C}\max_{w} L(P,w)$$

新引入的w里，我们需要遍历w取max。

3、表示出对偶问题，缩放，先求解对偶问题

为什么要转化为对偶问题？因为对偶问题好求解，最后我们能证明出对偶问题其实就是原问题的解。

为什么可以转化为对偶问题？因为对偶问题是原始问题“缩放”所

得，相当于条件更宽松了。

因为一般地，有max min f（x）<= min max f（x）理解为：所有班第一名中最菜的，比所有班倒数第一中最牛逼的要强。

原问题

$$\min_{P\in\mathbf{C}} \max_{w} L(P,w)$$

对偶问题

$$\max_{w} \min_{P\in\mathbf{C}} L(P,w)$$

这里由于拉格朗日函数L（P，w）是P的凸函数，所以对偶问题的解与原问题的解是等价的。

4、求解出内层min L（P，w）

单纯求解内层的min L（P，w），可以通过求解对P的偏导（对P（Y|X）的偏导），导数=0得最值。

$$\frac{\partial L(P,w)}{\partial P(y|x)} = \sum_{x,y}\tilde{P}(x)(\log P(y|x)+1) - \sum_{y} w_0 - \sum_{x,y}\left(\tilde{P}(x)\sum_{i=1}^{n} w_i f_i(x,y)\right)$$

$$= \sum_{x,y}\tilde{P}(x)\left(\log P(y|x)+1-w_0-\sum_{i=1}^{n} w_i f_i(x,y)\right)$$

我们能找到，去最小值时的P（Y|X）,我们表示出一个输入（x，y）的形式。

$$P_w(y|x) = \frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)$$

$$Z_w(x) = \sum_{y}\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)$$

这里由于w仍未知，依然带着它。Z称为规范化因子。

至此，问题定义中的P（Y|X）我们表示出来了，虽然表达式中有未知参数，不过这些参数可以通过数值优化的方法求得。

5、求解外层max，最终求解出整个 max min L（P，w）

其实到4、已经能看出最大熵模型的一般形式，见四、最大熵模型的一般形式

的一般形式。

现在Ψ(w)=min L（P，w）已经求得，对应Ψ(w)，求解max Ψ(w)就是求解最大熵模型的解（详见统计学习方法P87推导）。所以我们继续求解max min L（P，w），从而求出最终的问题，详见五、优化算法。

四、最大熵模型的一般形式

可以写成更一般的形式：

$$P_w(y|x)=\frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right)$$

$$Z_w(x)=\sum_{y}\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right)$$

其中n是特征维度，x∈R^n是输入向量，y∈{1,2..,K}是输出，w∈R^n是权重向量，f_i（x，y）是任意实值特征函数。这里特征函数怎么理解？？？！！！

五、优化算法

目标

对三、5中的Ψ(w)进行max（或者说对最大熵模型极大似然，因为这俩是等价的），Ψ(w)可以化简为（详见P87 6.27）

$$L(w)=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_i f_i(x,y)-\sum_{x}\tilde{P}(x)\log Z_w(x)$$

方法

数值优化的方法有很多，可选的方法有：改进的迭代尺度法（IIS）、拟牛顿法、随机梯度下降？。

下面只介绍IIS，拟牛顿法见统计学习方法P91（讲的不是很细）

改进的迭代尺度法思想

improved iterative scaling，IIS

基本想法：

在当前参数w=（w1,w2...wn）下找到一组新的参数w + δ =（w1 +δ1,w2+δ2..., wn+δn），使得模型逼近优化目标，并且不断重复此方法，知道达到要求。在最大熵模型中，优化目标是loglikelihood。

基本步骤：

假设参数从w变为w+δ，找到损失函数增加量的下界F = L（w+δ）- L（w）>= μ；找到让下界μ最大的δ，一般通过让"F对δ的偏导=0”的方式找到。

步骤

1、在参数w上加δ，找到改变量的下界。

通过各种推导得出

$$L(w+\delta)-L(w)\geqslant A(\delta\mid w)$$

$$A(\delta|w)=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_{x}\tilde{P}(x)\sum_{y}P_w(y\mid x)\exp\sum_{i=1}^{n}\delta_i f_i(x,y)$$

找到适当的δ就可以，但是变量δ是个多维向量，不是单变量形式，我们需要把他继续化为单变量的形式。尽量一次只有一个变量δi，其他的δj（i!=j）都是常量。

2、发现下界是多维变量δ表示的，找到只有一个变量δi情况下的下界。

引入变量f^#

$$f^{\#}(x,y)=\sum_{i}f_i(x,y)$$

因为f是0/1的，所以f^#表示所有（x，y）的特征出现的次数

通过各种推导得到一个新的下界

$$L(w+\delta)-L(w)\geqslant B(\delta\mid w)$$

$$B(\delta\mid w)=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_{x}\tilde{P}(x)\sum_{y}P_w(y\mid x)\sum_{i=1}^{n}\left(\frac{f_i(x,y)}{f^{\#}(x,y)}\right)\exp(\delta_i f^{\#}(x$$

)

$x, y))$

3、求下界对于δi的偏导，δi是唯一一个变量，其偏导=0就是，此时下界的最大值。

令下界对于变量δi的偏导=0，得到

$$\sum_{x,y} \tilde{P}(x) P_w(y|x) f_i(x,y) \exp(\delta_i f^{\#}(x,y)) = E_{\tilde{p}}(f_i)$$

此时的δi是当前参数情况下，应该在w增加上的参数。可以使得F = L（w+δ）- L（w）的下界最大的。

4、消除可能的变量f^#（x，y），计算δi

请看3、中的式子，其他的都是此时的参数/常量，只有f^#（x，y）是不知道的。

如果f^#（x，y）是恒定的，即∀（x，y），有f^#（x，y）= M，那么能此时已经可以直接写出δi的表达式；但是常常f^#（x，y）不是恒定的，那么得用数值优化的方式求解δi，如牛顿迭代。

六、与其他算法对比

1、与SVM

从模型思路角度，

1、两者都把满足现有训练数据作为基本条件。

基本SVM要求所有数据线性可分，并且确实把所有数据分开；soft-margin的SVM允许不把部分训练数据完全正确的分类。

最大熵基本的条件是，期望相同：对于二值函数f（x，y），（概率上）经验分布的期望与模型P（Y|X）的期望相同，经验分布也就是实际训练集的分布。

2、在满足训练数据情况下，额外的条件不同

SVM 要求把最难分的尽量分开。

最大熵要求未知的随机变量熵最大。

2、与crf

crf是在最大熵模型基础上发展来的，可以理解为crf是用最大熵做状态特征、用tag之间转移概率做转移特征。优化算法等也有很多相似之初，crf的优化算法可以看做在最大熵的优化算法上加了转移特征这一小改动。

3、与logistic

两者有类似的形式（通过化简可以看出，可以在二分类和多分类下对比，logistic的多分类版在统计学习方法P80）。他们又成为对数线性模型（log linear model）

七、参考资料

1、主要是统计学习方法，机器学习-周志华、模式分类、CS229中没有相关部分。

2、最大熵论文，还没看 http://www.isi.edu/natural-language/people/ravichan/papers/bergeretal96.pdf