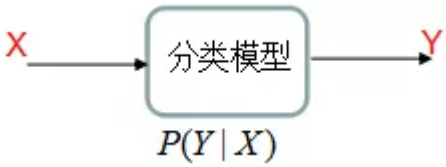
# 最大熵模型算法总结

原创 石头 机器学习算法那些事 2019-04-01

条件概率是机器学习模型的一种表现形式，应用这一模型，对于给定的输入X，得到各输出类的概率，选择最大概率的类为输出类，如下图：




本文介绍基于条件概率分类的两种模型算法：逻辑斯蒂（logistic）回归与最大熵模型，其中，logistic回归模型和最大熵模型分别是基于最大似然函数和熵来估计模型P(y|x)。公众号已有logistic回归模型的文章介绍，本文重点分析最大熵模型算法。

目录



## 1.最大熵模型算法

熵是衡量随机变量不确定性的指标，熵越大，随机变量的不确定性亦越大。假设X是一个离散型随机变量，其概率分布为：

$$P(X=x_i)=p_i\,,\quad i=1,2,...,n$$

随机变量X的熵定义为：

$$H(X)=-\sum_{i=1}^{n}p_i\log p_i$$

熵满足下列不等式：

$$0\le H(P)\le\log|X|$$

式中，|X|是x的取值个数，当且仅当X的分布是均匀分布时，右边的等号成立，也就是说，当X服从均匀分布时，熵最大。

### 1.1 最大熵模型的定义

最大熵原理是概率模型学习的一个准则，最大熵原理认为，学习概率模型时，在所有可能的概率模型（分布）中，熵最大的模型是最好的模型。条件概率是机器学习模型的一种表现形式，学习该模型的一种方法是最大化该条件概率的熵 ，即最大化下式：

$$H(P)=-\sum_{x,y}\overline{P}(x)P(y|x)\log P(y|x) \quad (1)$$

其中$\overline{P}(x)$表示变量X的经验分布：

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

其中v(X=x)表示训练数据中输入x出现的频数，N表示样本容量。

（1）式的未知变量$P(y|x)$就是需要学习的模型。

我们在构建分类模型$P(y|x)$的过程中假设训练数据集的联合概率分布与真实模型的联合概率分布相等，这一假设用特征函数f(x,y)的期望来描述，特征函数的定义：

$$f(x,y)=\begin{cases}1, & \text{训练数据集中包含的x，y}\\ 0, & \text{否则}\end{cases}$$

特征函数f(x,y)关于训练数据集的联合概率分布的期望值，用$E_{\overline{P}}(f)$表示：

$$E_{\overline{P}}(f)=\sum_{x,y}\overline{P}(x,y)f(x,y) \quad (2)$$

其中，$\overline{P}(x,y)=\frac{v(X=x,Y=y)}{N}$，v(X=x,Y=y)表示训练数据中样本(x,y)出现的频数。

特征函数f(x,y)关于模型$P(y|x)$与经验分布$\overline{P}(x)$的期望值，用$E_P(f)$表示：

$$E_P(f)=\sum_{x,y}\overline{P}(x)P(y|x)f(x,y) \quad (3)$$

假设两者期望相等，即：

$$E_{\overline{P}}(f)=E_P(f) \quad (4)$$

或

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

结合(1)(4)式，得到最大熵模型：

$$\max_{P} \quad H(P) = -\sum_{x,y} \overline{P}(x)P(y|x)\log P(y|x) \quad (5)$$

约束条件：

$$E_{\overline{P}}(f_i) = E_P(f_i), \quad i = 1,2,...,n$$

## 1.2 最大熵模型的学习

我们求解(5)式在约束条件下的最大值，其对应的模型P(Y|X)就是所学习的最优模型。

对于给定的训练数据集 $T = \{(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\}$ 以及特征函数 $f_i(x,y)$，i=1,2,...,n，

最大熵模型的学习等价于约束最优化问题：

$$\begin{aligned} \max_{P} \quad & H(P) = -\sum_{x,y} \overline{P}(x)P(y|x)\log P(y|x) \\ s.t \quad & E_{\overline{P}}(f_i) = E_P(f_i), \quad i = 1,2,...,n \\ & \sum_y P(y|x) = 1 \end{aligned}$$

将最大值问题转化为等价的求最小值问题：

$$\begin{aligned} \min_{P} \quad & -H(P) = \sum_{x,y} \overline{P}(x)P(y|x)\log P(y|x) & (2.1) \\ s.t \quad & E_{\overline{P}}(f_i) = E_P(f_i), \quad i = 1,2,...,n & (2.2) \\ & \sum_y P(y|x) = 1 & (2.3) \end{aligned}$$

引入拉格朗日乘子 $w_0, w_1, w_2, ..., w_n$ 将约束的最优化问题转换为无约束最优化的对偶问题，通过求解对偶问题求解原始问题。

定义拉格朗日函数L(P,w)：

$$\begin{aligned} L(P,w) &= -H(P) + w_0\left(1 - \sum_y P(y|x)\right) + \sum_{i=1}^{n} w_i\left(E_{\overline{P}}(f_i) - E_P(f_i)\right) \\ &= \sum_{x,y} \overline{P}(x)P(y|x)\log P(y|x) + w_0\left(1 - \sum_y P(y|x)\right) + \sum_{i=1}^{n} w_i\left(E_{\overline{P}}(f_i) - E_P(f_i)\right) \end{aligned}$$

最优化的原始问题：

对偶问题：

令



得:

$$P(y|x)=\frac{\exp(\sum_{i=1}^{n}w_i f_i(x,y))}{\exp(1-w_0)}$$

由于 $\sum_y P(y|x)=1$ ，对上式进行归一化得:

$$P_w(y|x)=\frac{1}{Z_w(x)}\exp(\sum_{i=1}^{n}w_i f_i(x,y)) \qquad (2.4)$$

其中,

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

令

$$\Psi(w)=\min_P L(P,w)$$

易知 $\Psi(w)$ 是关于w的函数，对偶问题外部的极大化问题:

$$\max_w \Psi(w)$$

根据上式求解的 $w^*$ 代入(2.4)式，得到最终的学习模型P(y|x)。

## 2. 最大熵模型例子

假设随机变量Y有5个取值 $\{y_1,y_2,y_3,y_4,y_5\}$ ，假设随机变量Y的条件概率分布满足如下条件:

$$P(y_1)+P(y_2)=P(\overline{y_1})+P(\overline{y_2})=\frac{3}{10} \qquad (2.1)$$

$$\sum_{i=1}^{5}y_i=\sum_{i=1}^{5}\overline{y_i}=1 \qquad (2.2)$$

求最大熵模型对应的概率分布P(Y)。

最大熵模型的目标函数:

$$\min_P \quad -H(P)=\sum_{i=1}^{5}P(y_i)\log P(y_i)$$

引进拉格朗日乘子 $w_0, w_1$，定义拉格朗日函数：

$$L(P,w)=\sum_{i=1}^{5}P(y_i)\log P(y_i)+w_1(P(y_1)+P(y_2)-\frac{3}{10})+w_0(\sum_{i=1}^{5}y_i-1)$$

令 $\frac{\partial L(P,w)}{\partial P(y_i)}=0$，得：

$$P(y_1)=P(y_2)=e^{-w_1-w_0-1}$$

$$P(y_3)=P(y_4)=P(y_4)=e^{-w_0-1}$$

将上式代入函数L(P,w)得 $\Psi(w)$，令 $\frac{\partial \Psi(w)}{\partial w}=0$，得：

$$e^{-w_1-w_0-1}=\frac{3}{20}$$

$$e^{-w_0-1}=\frac{7}{30}$$

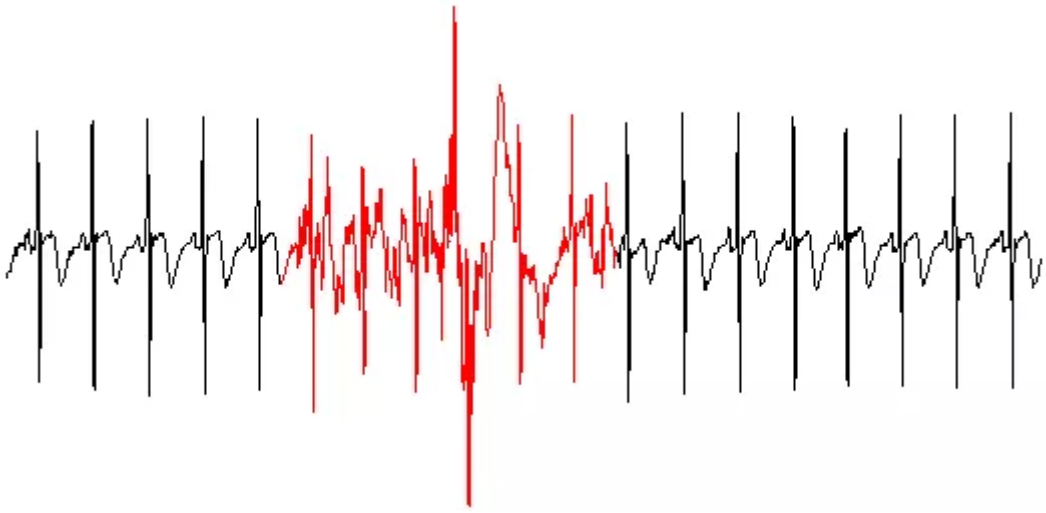
于是最大熵模型对应的概率分布：

$$P(y_1)=P(y_2)=\frac{3}{20}$$

$$P(y_3)=P(y_4)=P(y_4)=\frac{3}{20}$$

### 3. 熵模型在信号检测的应用

由第一节我们知道，熵是描述事物不确定性的指标。我们将熵的这一性质应用在信号检测领域，当信号包含了较强的随机噪声时或被噪声完全掩盖时，信号的随机性大大的增加了，其对应的熵也较大，根据这一原理对信号的质量进行检测，下图是用熵检测心电信号质量的效果图：



黑色表示较好的心电信号质量，红色表示较差的心电信号质量。

## 4. logistic回归算法

logistic回归是一种概率分类模型，对于二分类任务来说，其条件概率分布：

$$P(Y=1|x)=\frac{\exp(w\cdot x+b)}{1+\exp(w\cdot x+b)} \qquad (2.1)$$

$$P(Y=0|x)=\frac{1}{1+\exp(w\cdot x+b)} \qquad (2.2)$$

我们用最小化损失函数去估计上式的模型参数。对于给定的训练数据集 $T=\{(x_1,y_1),(x_2,y_2),...,(x_N,y_N)\}$，其中，$x_i\in R^n, y_i\in\{0,1\}$。

设：

$$P(Y=1|x)=\pi(x)$$
$$P(Y=0|x)=1-\pi(x)$$

似然函数为：

$$\prod_{i=1}^{N}[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i}$$

对数似然函数为：

$$L(w)=\sum_{i=1}^{N}[y_i\log\pi(x_i)+(1-y_i)\log(1-\pi(x_i))]$$

损失函数为：

$$J(w)=-L(w)$$
$$=-\prod_{i=1}^{N}[y_i\log(\pi(x_i))+(1-y_i)\log(1-\pi(x_i))] \qquad (2.3)$$

用梯度下降法求解w的估计值$\hat{w}$：

$$w=w-\alpha\frac{\partial J(w)}{\partial w} \quad ，\ 其中\ \alpha\ 为学习率$$

代入（2.1）（2.2）式，得到逻辑斯蒂回归模型P(y|x)，其中向量$\hat{w}$包含了b值 。

## 5. 小结

本文介绍基于条件概率分类的两种模型算法：logistic回归模型与最大熵模型，其中，logistic回归模型是基于最大似然函数估计模型P(y|x)，最大熵模型是基于熵这一指标估计模型P(y|x)。

参考

李航 《统计学习方法》

推荐阅读

logistic回归模型分析