最大熵模型

cloud

2017.1.29

1 最大熵原理

最大熵模型是由最大熵原理推导实现的，这里首先叙述一般的最大熵原理。最大熵原理是概率模型学习的一个准则，它认为学习概率模型时，在所有可能的概率模型或分布中，熵最大是最好的模型，通常用约束条件来确定概率模型的集合。最大熵原理也可以表述为在满足约束条件的模型集合中选取熵最大的模型。

假设离散随机变量 X 的概率分布是 $P(X)$，则其熵是:

$$H(P) = -\sum_{x} P(x) log P(x)$$

熵满足下列不等式:

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

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式才成立。也就是说当 X 服从均匀分布时熵最大。

最大熵原理认为要选择的模型首先是满足已有的事实即约束条件，那些不确定的部分都是“等可能的”。最大熵原理通过熵的最大化来表示等可能性。

2 最大熵模型

最大熵原理是统计学习的一般原理，将它应用到分类得到最大熵模型。

假设分类模型是一个条件概率分布 $P(Y|X)$，$X \in \mathcal{X} \subseteq R^n$ 表示输入，$Y \in \mathcal{Y}$ 表示输出，$\mathcal{X}$ 和 $\mathcal{Y}$ 分别表示输入和输出的集合。

给定一个训练数据集 $T = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(N)}, y^{(N)})\}$，学习的目标是用最大熵原理选择最好的分类模型。

首先考虑模型应该满足的条件，给定训练数据集可以确定联合分布 $P(X,Y)$ 的经验分布和边缘分布 $P(X)$ 的经验分布，分别用 $\hat{P}(X,Y)$ 和 $\hat{P}(X)$ 表示。

$$\hat{P}(X=x,Y=y)=\frac{v(X=x,Y=y)}{N}$$
$$\hat{P}(X=x)=\frac{v(X=x)}{N}$$

其中 $v(X=x,Y=y)$ 表示训练数据中样本 (x,y) 出现的频数，$v(X=x)$ 表示训练数据中输入 x 出现的频数，N 表示训练样本容量。

用特征函数 $f(x,y)$ 描述输入 x 和输出 y 之间的某一事实，其定义为：

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

特征函数 $f(x,y)$ 关于经验分布 $\hat{P}(X,Y)$ 的期望值，用 $E_{\hat{P}}(f)$ 表示如下：

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

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

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

如果模型能够获取训练数据中的信息，那么可以假设这两个期望值相等即 $E_P(f)=E_{\hat{P}}(f)$ 或 $\sum_{x,y}\hat{P}(x)P(y|x)f(x,y)=\sum_{x,y}\hat{P}(x,y)f(x,y)$，那么上述任一式子都可作为模型学习的约束条件。假设有 n 个特征函数 $f_i(x,y),i=1,2,...,n$，那么就有 n 个约束条件。

那么最大熵模型可以写成，假设满足所有约束条件的模型集合为：

$$\mathcal{C}\equiv\{P\in\mathcal{P}|E_P(f_i)=E_{\hat{P}}(f_i),i=1,2,...,n\}$$

定义在条件概率分布 $P(Y|X)$ 上的条件熵为：

$$H(P)=-\sum_{x,y}\hat{P}(x)P(y|x)logP(y|x)$$

则模型集合 $\mathcal{C}$ 中条件熵 $H(P)$ 最大的模型称为最大熵模型，上式中的对数为自然对数。

3 模型学习

最大熵模型的学习可以形式化为约束最优化问题。对于给定的训练数据集 $T = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(}$ 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：

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

$$s.t. \quad E_P(f) = E_{\hat{P}}(f), i = 1, 2, ..., n$$

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

求解上面的约束最优化问题，要将约束最优化的原始问题转换为无约束最优化的对偶问题。首先引进拉格朗日乘子 ω_i，定义拉格朗日函数 $L(P, \omega)$：

$$L(P, \omega) \equiv -H(P) + \omega_0 (1 - \sum_y P(y|x)) + \sum_{i=1}^{n} \omega_i (E_{\hat{P}}(f_i) - E_P(f_i))$$

$$= \sum_{x,y} \hat{P}(x) P(y|x) log P(y|x) + \omega_0 (1 - \sum_y P(y|x)) + \sum_{i=1}^{n} \omega_i (\sum_{x,y} \hat{P}(x, y) f_i(x, y) - \sum_{x,y} \hat{P}(x) P(y|x) f_i(x, y))$$

最优化的原始问题是：

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

对偶问题是：

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

由于拉格朗日函数 $L(P, \omega)$ 是 P 的凸函数，所有原始问题和对偶问题时等价的，那么求解对偶问题的解就是求解原始问题的解。

首先求解对偶问题内部的极小化问题 $\min_{P \in \mathcal{C}} L(P, w)$，因为它是 ω 的函数将其记为 $\Psi(\omega) = \min_{P \in \mathcal{C}} L(P, w) = L(P_\omega, \omega)$，$\Psi(\omega)$ 称为对偶函数，将其解记为：

$$P_\omega = arg \min_{P \in \mathcal{C}} L(P, w) = P_\omega(y|x)$$

具体地求 $L(P, \omega)$ 对 $P(y|x)$ 的偏导数为：

$$\frac{\partial L(P, \omega)}{\partial P(y|x)} = \sum_{x,y} \hat{P}(x) (log P(y|x) + 1) - \sum_y \omega_0 - \sum_{x,y} (\hat{P}(x) \sum_{i=1}^{n} \omega_i f_i(x, y))$$

$$= \sum_{x,y} \hat{P}(x) (log P(y|x) + 1 - \omega_0 - \sum_{i=1}^{n} \omega_i f_i(x, y))$$

令偏导数等于 0，在 $\hat{P}(x) > 0$ 的情况下，解得：

$$P(y|x) = exp(\sum_{i=1}^{n} \omega_i f_i(x,y) + \omega_0 - 1) = \frac{exp(\sum_{i=1}^{n} \omega_i f_i(x,y))}{exp(1-\omega_0)}$$

由于 $\sum_y P(y|x) = 1$，得：

$$P_\omega(y|x) = \frac{exp(\sum_{i=1}^{n} \omega_i f_i(x,y))}{Z_\omega(x)}$$

其中 $Z_\omega(x) = \sum_y exp(\sum_{i=1}^{n} \omega_i f_i(x,y))$ 是规范化因子，$f_i(x,y)$ 是特征函数，ω_i 是特征的权值，由 $P_\omega(y|x)$ 和 $Z_\omega(x)$ 表示的模型就是最大熵模型，这里 ω 是最大熵模型中的参数向量。

之后求解对偶问题外部的极大化问题：

$$\max_{\omega} \Psi(\omega)$$

将其解记为 ω^* 即：

$$\omega^* = arg \max_{\omega} \Psi(\omega)$$

这里得到 ω^* 用来表示 $P^* \in \mathcal{C}$，这里 $P^* = P_{\omega^*} = P_{\omega^*}(y|x)$ 是学习到的最优模型，那么最大熵模型的学习归结为对偶函数 $\Psi(\omega)$ 的极大化。

4 参考文献

1. 李航博士的统计学习方法