机器学习基础

最大熵模型

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 最大熵原理
- 最大熵模型
- 最大熵模型与逻辑斯蒂回归
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 最大熵原理
- 最大熵模型
- 最大熵模型与逻辑斯蒂回归
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

最大熵原理

- 熵：$0 \leq H(X) \leq \log n$，当 X 是等概分布时取到最大值
- 学习概率模型时，在所有可能的概率模型（即概率分布）中，熵最大的模型是最好的模型
  - 在满足已知条件的情况下，选取熵最大的模型
  - 在满足已知条件前提下，如果没有更多的信息，则那些不确定部分都是“等可能的”

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

最大熵原理

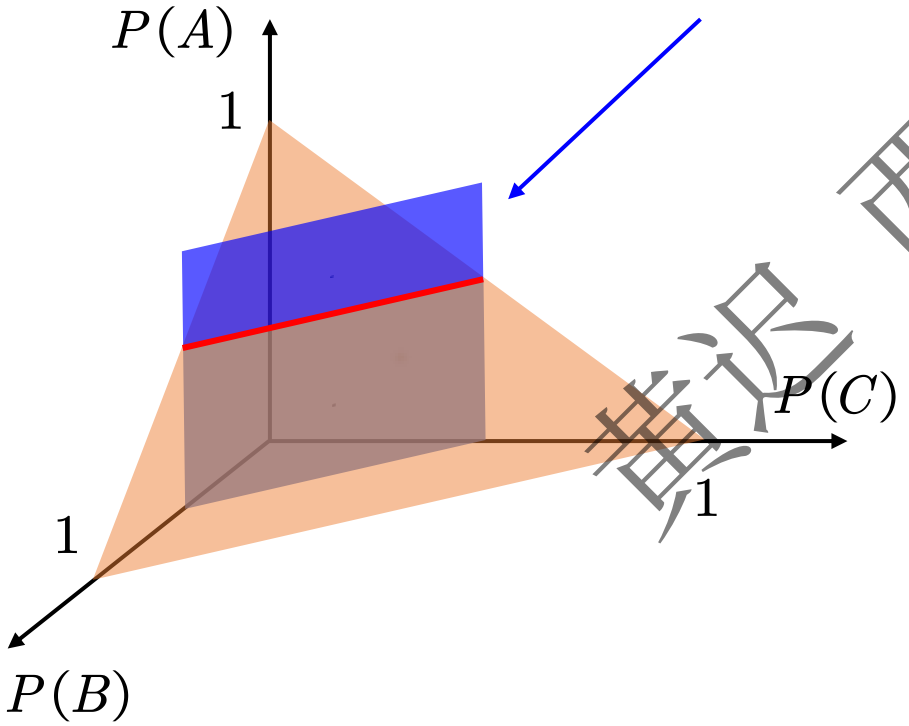
- … in making inferences on the basis of partial information we must use that probability distribution which has maximum-entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have.
- Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review*, 106: 620-630.

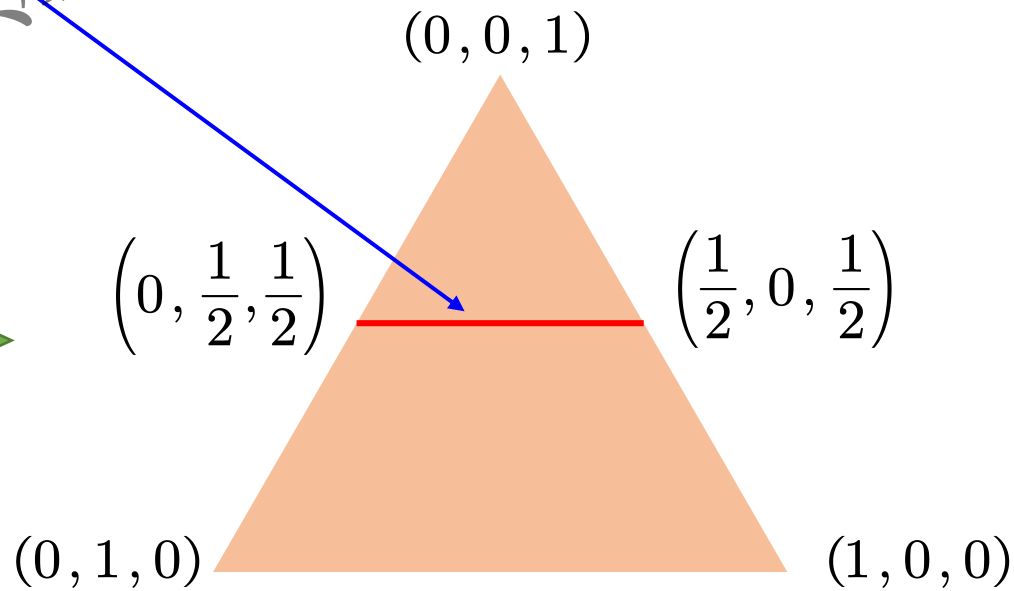
黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

几何解释

$P(A)+P(B)+P(C)=1$

约束：$P(B)+P(C)=\frac{1}{2}$





黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 最大熵原理
- 最大熵模型
- 最大熵模型与逻辑斯蒂回归
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

最大熵模型的定义

- 定义特征函数 $f(x,y)=\begin{cases}1, & x与y满足某一事实\\ 0, & 否则\end{cases}$
- 该函数在经验分布 $\tilde{P}(X,Y)$（计算训练样本的频率得到）下的期望是 $\mathbb{E}_{\tilde{P}}(f)=\sum_{x,y}\tilde{P}(x,y)f(x,y)$
- 函数在真实分布 $P(X,Y)$ 下的期望是

$$\mathbb{E}_P(f)=\sum_{x,y}P(x,y)f(x,y)=\sum_{x,y}P(y|x)P(x)f(x,y)\doteq\sum_{x,y}P(y|x)\tilde{P}(x)f(x,y)$$

- 如果模型训练准确，有 $\mathbb{E}_{\tilde{P}}(f)=\mathbb{E}_P(f)$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

最大熵模型的定义

- 对训练样本集 T，以及特征函数 $f_i(x,y), i=1,\cdots,n$，最大熵模型等价于约束优化问题

$$\begin{aligned} \max_{P\in\mathcal{P}} \quad & H(P)=-\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x) \\ \text{s.t.} \quad & \mathbb{E}_P(f_i)=\mathbb{E}_{\tilde{P}}(f_i),\ i=1,\cdots,n \\ & \sum_y P(y|x)=1,\ \forall x \end{aligned}$$

- 可以证明（Ref. [1]）满足约束且达到最大熵的分布是唯一的

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

参数学习

- 定义拉格朗日函数

$$L(P,w) = -H(P) + \sum_x w_x \left(1 - \sum_y P(y|x)\right) + \sum_{i=1}^{n} w_i \left(\mathbb{E}_{\tilde{P}}(f_i) - \mathbb{E}_P(f_i)\right)$$

- 原始问题是 $\min\limits_{P\in\mathcal{P}} \max\limits_{w} L(P,w)$；对偶问题是 $\max\limits_{w} \min\limits_{P\in\mathcal{P}} L(P,w)$
- 满足Slater条件，是强对偶
- 先求极小化问题 $\min\limits_{P\in\mathcal{P}} L(P,w)$，令 $\dfrac{\partial L(P,w)}{\partial P(y|x)} = 0$，有

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

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

参数学习

- 令 $\Psi(w)=L(P_w,w)$，则有

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

- 求解外部的极大化问题 $\max_w \Psi(w)$ 的解为 w^*（改进的迭代尺度法IIS）
- 则满足最大熵原理的条件概率分布为 $P^*=P_{w^*}(y|x)$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

例 6.2

- 将英语中的 in 翻译成法语，根据专家翻译总结有5种可能翻译A, B, C, D, E。已知$\tilde{P}(A)+\tilde{P}(B)=0.3$。
- 令$P(y_i)=P(y_i|X=\text{in})\quad \tilde{P}(y_i)=\tilde{P}(y_i|X=\text{in})\quad f(x,y)=\begin{cases}1, & y=y_1 \text{ or } y_2\\ 0, & \text{otherwise}\end{cases}$
- 则$\mathbb{E}_P(f)=\sum_{x,y}\tilde{P}(x)P(y_i|x)f(x,y)=\sum_y P(y_i)f(x,y)=P(y_1)+P(y_2)$

$$\mathbb{E}_{\tilde{P}}(f)=\sum_{x,y}\tilde{P}(x)\tilde{P}(y_i|x)f(x,y)=\sum_y \tilde{P}(y_i)f(x,y)=\tilde{P}(y_1)+\tilde{P}(y_2)=0.3$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

极大似然估计

- 对偶函数的极大化等价于最大熵模型的极大似然估计
- 已知经验分布$\tilde{P}(X,Y)$，则真实分布$P(X,Y)$的最大对数似然函数为

$$\max_{P\in\mathcal{P}} \log \prod_{x,y} P(x,y)^{\tilde{P}(x,y)} \Leftrightarrow \max_{P\in\mathcal{P}} \log \prod_{x,y} \tilde{P}(x)^{\tilde{P}(x,y)} P(y|x)^{\tilde{P}(x,y)}$$

$$\Leftrightarrow \max_{P\in\mathcal{P}} \left\{ \sum_{x,y} \tilde{P}(x,y) \log \tilde{P}(x) + \sum_{x,y} \tilde{P}(x,y) \log P(y|x) \right\}$$

$$\Leftrightarrow \max_{P\in\mathcal{P}} \sum_{x,y} \tilde{P}(x,y) \log P(y|x)$$

$$\Leftrightarrow \max_{P\in\mathcal{P}} \left\{ \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{n} w_i f_i(x,y) - \sum_{x} \tilde{P}(x) \log Z_w(x) \right\}$$

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 最大熵原理
- 最大熵模型
- 最大熵模型与逻辑斯蒂回归
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

两者都是对数线性模型

- 对数线性模型是取对数后模型是参数的线性组合
- 逻辑斯蒂回归是用线性模型的预测结果去逼近真实的对数几率

$$\log \frac{P(Y=1|x)}{1-P(Y=1|x)} = w \cdot x$$

- 最大熵模型的条件概率取对数有

$$\log P_w(y|x) = \sum_{i=1}^{n} w_i f_i(x,y) - \log Z_w(x)$$

关于 w 的线性函数

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

两者等价

- 最大熵模型和逻辑斯蒂回归是等价的
- 对于二分类问题，定义特征函数 $f_i(x,y)=\begin{cases} x^{(i)}, & y=1 \\ 0, & \text{otherwise} \end{cases}$，则满足最大熵原理的分布

$$P_w(Y=1|x)=\frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{n}w_i f_i(x,1)\right)=\frac{\exp\left(\sum_{i=1}^{n}w_i f_i(x,1)\right)}{\exp\left(\sum_{i=1}^{n}w_i f_i(x,0)\right)+\exp\left(\sum_{i=1}^{n}w_i f_i(x,1)\right)}$$

$$=\frac{\exp\left(\sum_{i=1}^{n}w_i x^{(i)}\right)}{1+\exp\left(\sum_{i=1}^{n}w_i x^{(i)}\right)}=\boxed{\frac{\exp(w\cdot x)}{1+\exp(w\cdot x)}}$$

逻辑斯蒂回归的条件概率分布

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

- 最大熵原理
- 最大熵模型
- 最大熵模型与逻辑斯蒂回归
- 总结

黄迟 计算机与人工智能学院
huangchi@swufe.edu.cn

算法优点

- 最大熵模型获得的是所有满足约束条件的模型中信息熵极大的模型,作为经典的分类模型时准确率较高
- 可以灵活地设置约束条件，通过约束条件的多少可以调节模型对未知数据的适应度和对已知数据的拟合程度

算法缺点

- 由于约束函数数量和样本数目有关系，导致迭代过程计算量巨大，实际应用比较难