最大熵模型

「Max Entropy Model」

对某一训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$

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

$$\tilde{P}(X=x)=\frac{v(X=x)}{N}$$ v 为频数

feature function: $f(x,y)=\begin{cases}1, & x\ 与\ y\ 满足某事实\\ 0, & else\end{cases}$

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

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

if $E_{\tilde{P}}(f)=E_{P}(f)$ => 约束条件

训练模型的目的是获取训练数据中的信息？

$$C\equiv\{P\in\mathcal{P}\mid E_P(f_i)=E_{\tilde{P}}(f_i),\ i=1,2,\cdots,n\}$$

Cond_Ent

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

$$H(Y|X)=\sum_{x\in X}p(x)H(Y|X=x)$$

$$= -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log p(y|x)$$

$$= -\sum_{x,y} p(x)\, p(y|x) \log p(y|x)$$

2. 最大熵模型的学习

Input: $T = \{(x_1, y_1), \cdots, (x_N, y_N)\}$,

n 个特征函数 $f_i(x, y)$, $i = 1, \cdots n$,

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

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

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

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

$$\text{s.t. } E_p(f_i) - E_{\tilde{p}}(f_i) = 0, \quad i = 1, 2, \cdots n$$

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

对偶化处理，引进拉格朗日乘因子 $w_0, w_1, \cdots w_n$

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

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

$$+ \sum_{i=1}^{n} w_i \left(E_{\tilde{p}}(f_i) - E_p(f_i) \right)$$

$$p_w = \arg\min_{p \in \mathcal{C}} L(p, w) = p_w(y|x) \Rightarrow \text{最大熵}$$

$$\frac{\partial L(p,w)}{\partial p(y|x)} = \sum_{x,y} \tilde{p}(x) \left(1 + \log p(y|x) \right) - \sum_y w_0$$

$$- \sum_{x,y} \left(\tilde{p}(x) \sum_{i=1}^{n} w_i f_i(x,y) \right) \qquad - \sum_y w_0$$

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

$$\because \sum_y w_0 = \sum_y w_0 \left(\sum_x \tilde{p}(x) \right)$$

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

令 $\frac{\partial L(p,w)}{\partial p(y|x)} = 0 \Rightarrow p(y|x) = \frac{\exp\left(\sum_{i=1}^{n} w_i f_i(x,y) \right)}{\exp(1 - w_0)}$

$\because \sum_y p(y|x) = 1$

$$\Rightarrow \sum_y \frac{\exp\left(\sum_{i=1}^{n} w_i f_i(x,y) \right)}{\exp(1 - w_0)} = 1$$

$$\Rightarrow \exp(1 - w_0) = \sum_y \exp\left(\sum_{i=1}^{n} w_i f_i(x,y) \right)$$

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

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

即为最大熵模型

⇒ 等价于极大似然估计

3. 最优化算法

常用方法：改进的迭代算法，SGD，牛顿法，或拟牛顿法

- 牛顿法 = $\min_{x \in R^n} f(x)$ ⇒ x^*（极小值解）

迭代过程中，第 k 次迭代值为 $x^{(k)}$，则

$$f(x) = f(x^{(k)}) + g_k^T (x - x^{(k)}) + \frac{1}{2}(x - x^{(k)})^T H(x^{(k)})(x - x^{(k)})$$

Hesse matrix $H(x^{(k)}) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right]_{n \times n}$, $x \in R^n$

若 $f(x)$ 有极小值 ⇒ $\nabla f(x^{(k+1)}) = 0$

⇒ $g_k^T + H(x^{(k)})(x^{(k+1)} - x^{(k)}) = 0$

⇒ $x^{(k+1)} = x^{(k)} - H_k^{-1} g_k$

OR $x^{(k+1)} = x^{(k)} + p_k$, $H_k p_k = -g_k$

具体算法为：

Input: $f(x)$, $x \in R^n$, $g(x)=\nabla f(x)$, $H(x)$, ε;

Output: x^*;

① 取初始点 $x^{(0)}$, $k=0$

② 计算 $g_k = g(x^{(k)})$

③ if $\|g_k\| < \varepsilon$, then $x^* = x^{(k)}$, return

④ 计算 $H_k = H(x^{(k)})$, $p_k = -H_k^{-1} g_k$

⑤ $x^{(k+1)} = x^{(k)} + p_k$, $k = k+1$, 转②

(2) 拟牛顿法

用 $n \times n$ $G_k = G(x^{(k)})$ 来近似代替 $H_k^{-1} = H^{-1}(x^{(k)})$

其中 H_k 应满足以下性质：

$$\nabla f(x^{(k+1)}) = g_k + H_k(x^{(k+1)} - x^{(k)})$$

$$\Rightarrow \underbrace{g_{k+1} - g_k}_{y_k} = H_k(\underbrace{x^{(k+1)} - x^{(k)}}_{\delta_k})$$

$$\Rightarrow H_k^{-1} y_k = \delta_k$$

同样地 $G_k y_k = \delta_k$

DFGS 算法：$B_{k+1}\delta_k = y_k$

令 $B_{k+1} = B_k + P_k + Q_k$

$\Rightarrow B_{k+1}\delta_k = B_k\delta_k + P_k\delta_k + Q_k\delta_k$

$$\begin{cases} P_k\delta_k = y_k \\ Q_k\delta_k = -B_k\delta_k \end{cases}$$

$$\Rightarrow B_{k+1} = B_k + \frac{y_k y_k^T}{y_k^T\delta_k} - \frac{B_k\delta_k\delta_k^T B_k}{\delta_k^T B_k\delta_k} \quad (*)$$

Input: $f(x)$, $g(x) = \nabla f(x)$, ε;
Output: 极小值 x^*;

① 选取初始点 $x^{(0)}$, B_0 为 positive-definite Matrix (Identity)
$k=0$.

② 计算 $g_k = g(x^{(k)})$, if $\|g_k\| < \varepsilon$.
then $x^* = x^{(k)}$, return

③ 由 $B_k P_k = -g_k \Rightarrow P_k$

④ 一维搜索：求 λ_k，使得

$$f(x^{(k)} + \lambda_k P_k) = \min_{\lambda>0} f(x^{(k)} + \lambda P_k)$$

⑤ $x^{(k+1)} = x^{(k)} + \lambda_k P_k$

⑥ $g_{k+1} = g(x^{(k+1)})$, if $\|g_{k+1}\| < \varepsilon$.
then $x^* = x^{(k+1)}$
else 按 (*) 求出 B_{k+1}

⑤ $k=k+1$，转至③

对于最大熵模型而言：

$$\min_{w\in R^n} f(w)=\sum_x \tilde{P}(x)\log\sum_y \exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)$$
$$-\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n} w_i f_i(x,y)$$

$$g(w)=\left(\frac{\partial f(w)}{\partial w_1},\frac{\partial f(w)}{\partial w_2},\cdots,\frac{\partial f(w)}{\partial w_n}\right)$$

with $$\frac{\partial f(w)}{\partial w_i}=\sum_{x,y}\tilde{P}(x)P_w(y|x)f_i(x,y)-E_{\tilde{P}}(f_i)$$

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

上次修改：23:06