最大熵模型(Maximum Entropy Model)

- 模型
  - 在没有更多信息情况下，对未知情况不做任何主观假设，不确定部分是等可能的
  - 在分类时，对于一系列可能的条件概率分布模型，在满足已知约束的条件下，我们选择熵最大的模型作为我们的分类模型
  - P(Y|X)*=argmax(H(P(Y|X)))，P(Y|X)为条件概率分布，X为特征变量，Y为类别变量
- 策略
  - 相关计算
    - 根据所获得的数据，计算先验信息
    - 两类分布
      - 经验联合分布　P'(X=x, Y=y)=count(X=x, Y=y)/N
      - 经验边缘分布　P'(X=x)=count(X=x)/N
    - 贝叶斯　P(X=x, Y=y)=P'(X=x)*P(Y=y|X=x)
  - 特征函数
    - 用来表示元素是否属于某个子集
    - 用多个特征(指示)函数来描述数据样本中的规律, 我们以二值函数为例
      - $f_i(x,y)=1$　x和y满足条件
      - $f_i(x,y)=0$　x和y不满足条件
  - 约束条件
    - 特征函数fi(x,y)在经验联合分布上的期望=在联合分布上的期望
      - $E_{p'}(f_i)=\sum_{x,y}P'(x,y)f_i(x,y)$　关于经验分布期望
      - $E_p(f_i)=\sum_{x,y}P'(x)P(y|x)f_i(x,y)$　关于条件分布期望
      - $E_p(f_i)=E_{p'}(f_i)$　两者相等
    - 同一个特征条件下，概率累计为1　$\sum_y P(y|x)=1$
  - 目标函数　$H(P)=-\sum_{x,y}P'(x)P(y|x)logP(y|x)$　取熵最大下的P(y|x)
- 算法
  - 最优化问题　$P=argmax(-\sum_{x,y}P'(x)P(y|x)logP(y|x))$
    - s.t.
      - $E_p(f_i)=E_{p'}(f_i)$
      - $\sum_y P(y|x)=1$
  - 两种求解过程
    - 拉格朗日对偶性求解
      - 将有约束目标函数转为无约束求最小值，对于每个约束添加一个权值w　$L(P,w)=\sum_{x,y}P'(x)P(y|x)logP(y|x)+w_0(1-\sum_y P(y|x))+\sum_{i=1}^{n}w_i(E_{p'}(f_i)-E_p(f_i))$
      - 原始问题　$\min_P\max_w L(P,w)$
      - 对偶问题　$\max_w\min_P L(P,w)$
        - 先固定w，求内部极小下的P　$\psi(w)=\min_P L(P,w)$
          - $P_w(y|x)=\frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right)$
          - $Z_w(x)=\sum_y\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right)$
        - 根据第一步的结果，将P带入求解外部极大值下的w
          - $\psi(w)=\sum_{x,y}P'(x,y)\sum_{i=1}^{n}w_i f_i(x,y)-P'(x,y)\sum_{i=1}^{n}logZ_w(x)$
          - $w^*=arg\max_w\psi(w)$
    - 最大似然估计求解
      - $L(P_w)=\frac{1}{N}log\prod_{x,y}P(y|x)^{count(x,y)}=log\prod_{x,y}P(y|x)^{P'(x,y)}=\sum_{x,y}P'(x,y)logP(y|x)$
      - 带入在拉格朗日对偶求解第一步解P(y|x)
      - $L(w)=\sum_{x,y}P'(x,y)\sum_{i=1}^{n}w_i f_i(x,y)-P'(x,y)\sum_{i=1}^{n}logZ_w(x)$　此表达式与φ(w) 一致，求该L(w)最大下的w
  - 求解算法
    - 改进迭代尺度(IIS)--专用于求解最大熵模型
      - 目标函数：L(w)
      - 基本思想　变尺度迭代wi->wi+δi，直到找到L(w)最大值
      - 核心推导　在w的迭代更新中，找到一个δ，确保每次L(w+δ)-L(w)的值最大
        - $L(w+\delta)-L(w)=\sum_{x,y}P'(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)-\sum_{x,y}P'(x)log\frac{Z_{w+\delta}(x)}{Z_w(x)}$
        - $L(w+\delta)-L(w)>=\sum_{x,y}P'(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)-\sum_{x,y}P'(x)\frac{Z_{w+\delta}(x)}{Z_w(x)}+1$　利用$-log\alpha>=1-\alpha$
        - 上式右边设为：A(δ|w)　$A(\delta|w)=\sum_{x,y}P'(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)-\sum_x P'(x)\sum_y P_w(y|x)\sum_{i=1}^{n}\delta_i f_i(x,y)$
        - 利用jensen不等式 $g(\sum_{i=1}^{n}\lambda_i x_i)<=\sum_{i=1}^{n}\lambda_i g(x_i),\sum_{i=1}^{n}\lambda_i=1$ 改写上式右侧　$B(\delta|w)=\sum_{x,y}P'(x,y)\sum_{i=1}^{n}\delta_i f_i(x,y)+1-\sum_x P'(x)\sum_y P_w(y|x)\sum_{i=1}^{n}(\frac{f_i(x,y)}{f^*(x,y)})exp(\delta_i f^*(x,y))$　其中f*(x,y)=∑f(x,y)
        - 对上式对δi求导并设为0，得到　$\sum_{x,y}P'(x)P_w(y|x)f_i(x,y)exp(\delta_i f^*(x,y))=E_{P'}(f_i)$　$\delta_i=\frac{1}{f^*(x,y)}log\frac{E_{P'}(f_i)}{E_P(f_i)}$
      - 基本流程
        - 1 输入：根据所获样本求经验分布P'(X=x,Y=y)、特征函数fi和模型Pw(y|x)
        - 2 初始化所有权值为0
        - 3 权值更新wi->wi+δi　δi计算　$\delta_i=\frac{1}{f^*(x,y)}log\frac{E_{P'}(f_i)}{E_P(f_i)}$
        - 4 不断重复3直到w收敛
    - 拟牛顿
    - 梯度下降
- API
  - 目前sklearn并没有最大熵模型
- 主要应用
  - 分类相关，如文本分类等
- 优劣分析
  - 优点
    - 可以灵活选择特征来调节模型
    - 原理简单，可解释性强
    - 模型结构简单
  - 缺点
    - 随着样本量、约束函数的增多，计算量较大，效率低下