朴素贝叶斯

推导朴素贝叶斯中的概率估计公式

贝叶斯估计

对于先验概率 假设进行N次实验，先验概率 $$ P(Y=c_k) = \frac{1}{K} \ PK - 1 = 0 $$ 由频率是概率的极大似然估计 $$ P(Y=c_k) = \frac{\sum_{i=1}^{N}{I(y_i = c_k)}}{N} $$ 可得 $$ P(Y=c_k)N - \sum_{i=1}^{N}{I(y_i = c_k)} = 0 $$ 即 $$ \lambda(P(Y=c_k)K - 1) + P(Y=c_k)N - \sum_{i=1}^{N}{I(y_i = c_k)} = 0 $$ 可得 $$ P(Y=c_k) = \frac{ \sum_{i=1}^{N}{I(y_i = c_k)} + \lambda}{K\lambda + N} $$ 对于似然概率

- $S_j$是第j个特征可能取值的个数，这里假设每个特征的值是等概率分布，因此概率为$\frac{1}{S_j}$ $$ P(X^{(j)} = a_{jl} | Y =c_k) = \frac{1}{S_j} \ P(X^{(j)} = a_{jl} | Y =c_k) = \frac{\sum_{i=1}^{N}{I(X^{(j)} = a_{jl} , Y =c_k)}}{\sum_{i=1}^{N}{I(Y=c_k)}} $$ 同上 $$ \lambda{PS_j - 1} + P\sum_{i=1}^{N}{I(Y=c_k)} - \sum_{i=1}^{N}{I(X^{(j)} = a_{jl} , Y =c_k)} = 0 $$ 可得 $$ P(X^{(j)} = a_{jl} | Y =c_k)= \frac{\lambda +\sum_{i=1}^{N}{I(X^{(j)} = a_{jl} , Y =c_k)}}{\sum_{i=1}^{N}{I(Y=c_k)} + S_j\cdot \lambda} $$

后验概率最大化的含义

这里来证明一下朴素贝叶斯是如何从损失函数最小化推出后验概率最大化的

首先我们写出期望风险的公式 $$ \begin{aligned} R_{exp} &= \int{\int{L(y,f(\vec{x}))P(\vec{x},y)d\vec{x}}dy}\ &= \int_x{\int_y{L(y,f(\vec{x}))P(y|\vec{x})dy}P(\vec{x})d\vec{x}}\ &= \mathbb{E}_x{[\int_y{L(y,f(\vec{x}))P(y|\vec{x})dy}]} \ &= \mathbb{E}x{[\sum{k=1}^{K}{L(c_k,f(\vec{x}))P(c_k|\vec{x})}]}

\end{aligned} $$ 也就是对于每个$x$，我们对$L(y,f(\vec{x}))P(y|\vec{x})$进行最小化 $$ \begin{aligned} f(x) &= \argmin_{y\in Y}{\sum_{k=1}^KL(c_k,y)P(c_k|X=x)}\ &= \argmin_{y\in Y}{\sum_{k=1}^KP(y\not = {c_k}|X= x)} \ &= \argmin_{y\in Y}({1 - P(y={c_k}|X= x)}) \quad \text{所有概率和为1，等价为1减去预测正确的概率}\ &= \argmax_{y\in Y}{P(y ={c_k}|X= x)} \ \end{aligned} $$

支持向量机

合页损失函数的推导

在软间隔支持向量机里面，引入松弛变量$\xi_i$

目标函数为 $$ L = \frac{1}{2}||w||^2 + C\sum_{i = 1}^N{\xi_i} $$ 约束条件 $$ \begin{aligned} y_i(w^Tx_i + b) &\ge 1 - \xi_i \ \xi_i \ge 0 \end{aligned} $$

- $y_i(w^Tx_i + b) \ge 1$时，说明样本被正确分类且间隔足够大
  - 此时$1 - y_i(w^Tx_i + b)\le 0$，而约束条件为$\xi_i \ge 1 - y_i(w^Tx_i + b)$，$\xi_i$需要大于一个负数且大于0
  - 为了让目标函数最小，令$\xi_i = 0$
- $y_i(w^Tx_i + b) < 1$时，说明样本间隔不足或者被误分类
  - 此时 $1 - y_i(w^Tx_i + b)> 0$，同上面的推导，约束条件为$\xi_i \ge 1 - y_i(w^Tx_i + b)$，$\xi_i$需要大于一个正数且大于0
  - 为满足约束且使目标函数最小，$\xi_i = 1 - y_i(w^Tx_i + b)$

综上所述 $$ \xi_i = \max{(0,1 - y_i(w^Tx_i + b))} $$

代入目标函数并令$\lambda = \frac{1}{2C}$可得合页损失函数 $$ \sum_{i=1}^N[1 - y_i(w^Tx_i + b)]+ + \lambda||w||^2 $$ 其中 $$ [z]+ = \begin{cases} z , \quad z \ge 0 \ 0, \quad z < 0 \end{cases} $$

EM算法

EM算法的推导

它的推导和VAE（变分自编码器）以及Diffusion Model（扩散模型）很像，都是由于原函数难以求解，去找到它的下界函数，对下界函数求得最优解，达到优化原函数的目的。感兴趣的可以看一下VAE和Diffusion Model的推导

首先要了解EM算法的核心：**EM**算法是计算每个潜在变量**Z**的后验概率，然后通过加权平均，也就是求期望的方式得到一个近似的估计，最后我们要找到一个能让这个近似估计最大的参数**θ**

先来认识一下**KL**散度 $$ KL(q||p) = \sum_{Z}q(Z)\log{\frac{q(Z)}{p(Z)}} $$ 描述的是q分布和p分布之间的距离，因此取值是大于等于0的

首先要明确，EM算法的核心是在每一步最大化**Q**函数 $$ \begin{aligned} Q(\theta,\theta^{(i)})&=\mathbb{E}*{Z}(\log{P(Y,Z|\theta)}|Y,\theta^{(i)}) \ &= \sum*{Z}\log{P(Y,Z|\theta)}P(Z|Y,\theta^{(i)}) \end{aligned} $$

现在来看如何推导出上面这个Q函数

对于EM算法，我们希望最大化观测数据Y关于参数θ的对数似然函数 $$ \begin{aligned} L(\theta) = \log{P(Y|\theta)} &= \log{\sum_{Z}{P(Y,Z|\theta)}} \ &=\log{\sum_{Z}{q(Z)\frac{P(Y,Z|\theta)}{q(Z)}}} \end{aligned} $$

使用**Jeason**不等式 $$ \begin{aligned} \log{\sum_{Z}{q(Z)\frac{P(Y,Z|\theta)}{q(Z)}}} &\ge \sum_{Z}q(Z)\log{\frac{P(Y,Z|\theta)}{q(Z)}} \ &= \sum_{Z}q(Z)\log{P(Y,Z|\theta)} - \sum_{Z}q(Z)\log{q(Z)} \end{aligned} $$

记上式为 $$ \begin{aligned} \mathcal{{L}}(q,\theta) = \sum_{Z}q(Z)\log{P(Y,Z|\theta)} - \sum_{Z}q(Z)\log{q(Z)} \end{aligned} $$

因为 $$ \log{P(Y,Z|\theta)} = \log{P(Z|Y,\theta)P(Y|\theta)} $$

因此有 $$ \begin{aligned} \mathcal{{L}}(q,\theta) &= \sum_{Z}q(Z)\log{P(Y,Z|\theta)} - \sum_{Z}q(Z)\log{q(Z)} \ &= \sum_{Z}q(Z)\log{P(Z|Y,\theta)P(Y|\theta)} - \sum_{Z}q(Z)\log{q(Z)} \ &= \sum_{Z}q(Z)(\log{P(Z|Y,\theta)}+\log{P(Y|\theta)}) - \log{q(Z)})\ &= \log{P(Y|\theta)}\sum_{Z}q(Z) + \sum_{Z}q(Z)\log{\frac{P(Z|Y,\theta)}{q(Z)}} \ &= \log{P(Y|\theta)} - KL(q(Z)||P(Z|Y,\theta)) \end{aligned} $$

最终等式 $$ \begin{aligned} \log{P(Y|\theta)} &= KL(q(Z)||P(Z|Y,\theta)) + \mathcal{{L}}(q,\theta)\ \end{aligned} $$ 且 $$ \log{P(Y|\theta)} \ge \mathcal{{L}}(q,\theta) $$

老师上课的时候没有说$q(Z)$为什么要取$P(Z|Y,\theta^{i})$，这里解释一下

> □***E***步的时候，理论上我们可以取任意的*$q(Z)$*来构造下界*$\mathcal{{L}}(q,\theta)$*，但是为了让当前参数的似然函数与下界相等，我们要让 *$$ q(Z) = P(Z|Y,\theta^{i}) $$* 为什么要这样？这样选择能让在参数*$\theta^{(i)}$*处，下界紧贴似然函数，即：*$$ \log{P(Y,Z|\theta^{(i)})} = \mathcal{{L}}(q,\theta^{(i)}) $$*（此时*KL*散度为*0*）
>
> 当下界紧贴似然函数时，提升下界会对似然函数产生最大程度的提升，收敛速度最快
>
> 如果下界很松（*KL*散度很大），即使提升了下界，似然函数可能只有很小的增长
>
> 注意：选择*$q(Z) = P(Z|Y,\theta^{(i)})$*只是为了在*$E$*步的当前点*$\theta^{(i)}$*使得*KL*散度为零、下界紧贴似然函数，从而让后续优化更高效，一旦进入*M*步，参数*θ*开始更新，真实后验*$P(Z|Y,\theta)$*随之改变，*KL* 散度立即变为正数，下界与似然函数不再紧贴。

因此 $$ \begin{aligned} \log{P(Y|\theta)} &\ge \mathcal{{L}}(q,\theta)\ &= \sum_{Z}P(Z|Y,\theta^{(i)})\log{P(Y,Z|\theta)} + const \ &= \underbrace{\mathbb{E}*{Z}*

(\log{P(Y,Z|\theta)}|Y,\theta^{(i)})}{\text{Q函数}} + \underbrace{\sum_{Z}{P(Z|Y,\theta^{(i)})\log{P(Z|Y,\theta^{(i)})}}}_{\text{常数}} \end{aligned} $$

单调性的证明

在E步的时候，取 $$ q(Z) = P(Z|Y,\theta) $$ 做完M步之后，新下界大于旧下界 $$ \mathcal{{L}}(q,\theta^{(i + 1)}) \ge \mathcal{{L}}(q,\theta^{(i)}) $$ 旧似然等于旧下界 $$ \log{P(Y,Z|\theta^{i})} = \mathcal{{L}}(q,\theta^{(i)}) $$ 新似然做完M步后KL散度大于0，因此 $$ \log{P(Y,Z|\theta^{i + 1})} \ge \mathcal{{L}}(q,\theta^{(i + 1)}) $$ 最后有 $$ \log{P(Y,Z|\theta^{i + 1})} \ge \mathcal{{L}}(q,\theta^{(i + 1)})\ge\mathcal{{L}}(q,\theta^{(i)})=\log{P(Y,Z|\theta^{i})} $$ 保证了**EM**算法的单调性

高斯混合模型

高斯分布 $$ \phi(y|\theta_{k}) = \frac{1}{\sqrt{2\pi}\sigma_{k}}exp(-\frac{(x-\mu_k)^2}{2\sigma_{k}^2}) $$

用多个高斯分布的加权组合描述复杂数据的分布 $$ \begin{aligned} &P(y|\theta) = \sum_{k=1}^K \alpha_k \phi(y|\theta_k) \ &\text{s.t.} \quad \alpha_k \geq 0 \ (k=1,\dots,K), \quad \sum_{k=1}^K \alpha_k = 1 \end{aligned} $$

Q函数的定义 $$ Q(\theta,\theta^{(i)}) = \mathbb{E}_{\gamma}[\log{P(y,\gamma|\theta)}|y,\theta^{(i)}] $$

观测数据是由高斯混合模型产生的，模型参数为 $$ \theta = (\alpha_{k},\mu_{k},\sigma_{k}^2) $$

隐变量定义为 $$ \gamma_{jk} = \begin{cases} 1 &\text{第j个观测样本来自第k个高斯模型} \ 0 \end{cases} $$

现在来推导完全数据的似然函数，因为后续会使用这个代入Q函数 $$ \begin{aligned} P(y,\gamma|\theta) &= \prod_{j = 1}^{N}{p(y_j,\gamma_{j1},\gamma_{j2},\dots,\gamma_{jk}|\theta)} \ &=\prod_{j = 1}^{N}{\prod_{k=1}^{K}{[\alpha_{k}\phi(y_j|\theta_k)]^{\gamma_{jk}}}} \end{aligned} $$

令 $$ n_k = \sum_{j=1}^{N}\gamma_{jk} $$

有 $$ \begin{aligned} \prod_{j = 1}^{N}{\prod_{k=1}^{K}{[\alpha_{k}\phi(y_i|\theta_k)]^{\gamma_{jk}}}} &= \prod_{k = 1}^K{\alpha_k^{n_k}}\prod_{j=1}^N[\phi(y_j|\theta_k)]^{\gamma_{jk}} \end{aligned} $$

代入到Q函数中 $$ \begin{aligned} Q(\theta,\theta^{(i)}) &= \mathbb{E}*{\gamma}{\sum*{k=1}^N{n_k\log{\alpha_{k} + \sum_{j=1}^N{\gamma_{jk}[\log{\phi(y_j|\theta_{k})}]}}}}|y,\theta^{(i)}} \ &= \mathbb{E}*{\gamma}{\sum*{k=1}^N{n_k\log{\alpha_{k} + \sum_{j=1}^N{\gamma_{jk}[(\log{\frac{1}{\sqrt{2\pi}}}-\log{\sigma_{k}} - \frac{(y_j - \mu_k)^2}{2\sigma_k^2})]}}}}|y,\theta^{(i)}} \

\end{aligned} $$

我们可以很轻松地发现，真正需要计算的只有$\mathbb{E}(\gamma_{jk}|y,\theta^{(i)})$，因为其它的都可以通过期望的线性性质直接拆开得到结果 $$ \begin{aligned} \mathbb{E}(\gamma_{jk}|y,\theta^{(i)}) &= P(\gamma_{jk} = 1|y,\theta^{(i)}) \ &= \frac{P(\gamma_{jk} = 1|\theta^{(i)})P(y_j|\gamma_{jk} = 1,\theta^{(i)})}{P(y|\theta^{(i)})}\ &= \frac{\alpha_{k}\phi(y_j|\theta_k)}{\sum_{k=1}^K{\alpha_{k}\phi(y_j|\theta_k)}} \ &= \hat{\gamma}_{jk} \end{aligned} $$

并且我们有 $$ \begin{aligned} n_k &= \sum_{j=1}^{N}{\gamma_{jk}} \ \mathbb{E}[n_k|y,\theta^{(i)}] &= \sum_{j=1}^N{\mathbb{E}[\gamma_{jk}|y,\theta^{(i)}]} = \sum_{j=1}^{N}{\hat{\gamma}_{jk}} \end{aligned} $$

最后得到**Q**函数的最终形式 $$ \begin{aligned} Q(\theta,\theta^{(i)}) &= \sum_{k = 1}^{N}{\sum_{j=1}^{N}{(\mathbb{E}{\gamma_{jk}})}\log{\alpha_k}+ \sum_{j=1}^N{(\mathbb{E}{\gamma_{jk}})}[(\log{\frac{1}{\sqrt{2\pi}}}- \log{\sigma_{k}} - \frac{(y_j - \mu_k)^2}{2\sigma_k^2})]|y,\theta^{(i)}}\ &= \sum_{k = 1}^{N}{\sum_{j=1}^{N}{\hat{\gamma}{jk}}\log{\alpha_k}+ \sum{j=1}^N{\hat{\gamma}{jk}}[(\log{\frac{1}{\sqrt{2\pi}}}- \log{\sigma{k}} - \frac{(y_j - \mu_k)^2}{2\sigma_k^2})]|y,\theta^{(i)}} \end{aligned} $$

> 解释一下上方公式难以理解的点
>
> - $\mathbb{E}(\gamma_{jk}|y,\theta^{(i)})$的值实就是$P(\gamma_{jk} = 1|y,\theta^{(i)})$，因为前面定义过隐变量的值不是$1$就是$0$，这里可以很容易看出
> - $\hat{\gamma}_{jk}$实就是样本点y_j对高斯分布k的隶属度
> - $P(y|\theta^{(i)})$不是条件概率，而是在参数为$\theta^{(i)}$的情况下的似然函数，而y可以来自不同的高斯分布，因此写成求和的形式
> - $P(\gamma_{jk} = 1|\theta^{(i)})$就是在参数为$\theta^{(i)}$的情况下数据点来自第$k$个分量的概率，也就是权重$\alpha_k$

最后对模型参数α,μ,σ进行更新 $$ \begin{aligned} \hat{\alpha}k &= \frac{n_k}{N} = \frac{\sum{j=1}^{N}{\hat{\gamma}{jk}}}{N}\ \hat{\mu}k &= \frac{\sum{j = 1}^N{\hat{\gamma}{jk}y_j}}{\sum_{j = 1}^N{\hat{\gamma}{jk}}}\ \hat{\sigma}k &= \frac{\sum{j = 1}^N{\hat{\gamma}{jk}(y_j-\mu_k)^2}}{\sum_{j = 1}^N{\hat{\gamma}_{jk}}} \end{aligned} $$