第三讲 基于后验概率最大化准则的分类模型

2025年3月

决策函数与条件概率

- 回顾支持向量机模型中，从训练样本集 $\{(x_i, y_i)\}_{i=1}^N$ 直接学得决策函数 $f(x) = \mathrm{sign}(\sum_{i=1}^N \alpha_i y_i K(x_i, x) + b)$, 对新数据实例$x$的预测$\hat{y}$直接由决策函数给出，即 $\hat{y} = f(x)$.
- 本讲介绍以条件概率分布$P(Y|X)$而非决策函数$f(x)$为模型的分类方法:
  - 通常先从训练样本集$T = \{(x_i, y_i)\}_{i=1}^N$学得条件概率分布$P(Y|X)$.
  - 再对新数据实例x 按照后验概率最大化原则确定预测$\hat{y}$，即

$$\hat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}} P(y|x).$$

概要

1. 后验概率最大化分类准则

训练样本集$D=\{(x_i,y_i)\}_{i=1}^N$, 其中 $x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)})^T\in\mathcal{X}$, $y_i\in\mathcal{Y}=\{c_1,c_2,\cdots,c_K\}$, $i=1,2,\cdots,N$.

将x的类别预测为c_i所产生的风险(期望损失)为

$$R(Y=c_i|x)=\sum_{j=1}^{K}\lambda_{ij}P(Y=c_j|x),$$

其中λ_{ij}是将属于c_j的样本判定为c_i类的损失.

最优预测

依据贝叶斯决策论，对输入实例x的最优预测$\hat{y}$应该满足

$$\hat{y}=\operatorname*{argmin}_{c_i} R(Y=c_i|x)$$

如果我们采用0-1损失函数，即 $\lambda_{ij} = \begin{cases} 1, & i \neq j \\ 0, & i = j \end{cases}$ 则

$$\begin{aligned} R(Y = c_i|x) &= \sum_{j \neq i} 1 \times P(Y = c_j|x) + 0 \times P(Y = c_i|x) \\ &= \sum_{j \neq i} P(Y = c_j|x) = 1 - P(Y = c_i|x). \end{aligned}$$

相应地，对输入实例x的最优预测$\hat{y}$应该满足

$$\begin{aligned} \hat{y} &= \underset{c_i}{\operatorname{argmin}}[1 - P(Y =_i |x)] \\ &= \underset{c_i}{\operatorname{argmax}} P(Y = c_i|x), \end{aligned}$$

即依据贝叶斯决策论，输入实例x的最优预测$\hat{y}$为使得后验概率$P(y|x)$最大的类标记。

如何计算相应的后验概率$P(Y=c_i|x)$?

- 对于判别式模型，直接从训练样本集$D=\{(x_i,y_i)\}_{i=1}^N$学习出后验概率$P(Y=c_i|x)$.
- 而对于生成式模型，我们需要学习出联合概率分布$P(Y=c_i,x)$，然后依

$$P(Y=c_i|x)=\frac{P(Y=c_i,x)}{P(x)}$$

计算出$P(Y=c_i|x)$.

概要

1. 后验概率最大化分类准则
2. 逻辑斯谛回归模型

逻辑斯谛回归模型直接以参数形式给出条件概率分布$P(Y|X)$.

二项逻辑斯谛回归模型

设$\mathcal{X} = \mathbf{R}^n$，$\mathcal{Y} = \{c_1, c_2\}$(或者$\{1, 0\}$)，则二项逻辑斯谛回归模型是如下的后验概率分布：

$$
\begin{aligned}
P(Y = c_1|x) &= \frac{\exp(w \cdot x + b)}{1 + \exp(w \cdot x + b)}, \\
P(Y = c_2|x) &= \frac{1}{1 + \exp(w \cdot x + b)},
\end{aligned}
$$

其中$x \in R^n$是输入实例，Y是输出对应的随机变量，参数权值向量$w \in \mathbf{R}^n$，偏置$b \in \mathbf{R}$。

二项逻辑斯谛回归模型的分布具有逻辑斯蒂函数的形式:

- 当$w \cdot x + b \to +\infty$， $P(Y = c_1|x) \to 1$而 $P(Y = c_2|x) \to 0$;
- 当$w \cdot x + b \to -\infty$， $P(Y = c_1|x) \to 0$ 而 $P(Y = c_2|x) \to 1$.

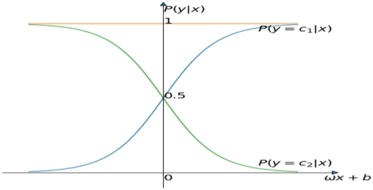

Figure: 二项逻辑斯谛模型

- 对于输入实例x，二项逻辑斯谛回归模型按照后验概率最大化原则对x进行分类，即

$$y = \begin{cases} c_1, & \text{if } P(Y = c_1|x) > P(Y = c_2|x) \\ c_2, & \text{otherwise} \end{cases}$$

- 如果引进$Y = c_1$的几率$\frac{P(Y=c_1|x)}{P(Y=c_2|x)}$，上面的分类准则也可以表示为

$$y = \begin{cases} c_1, & \text{if } \frac{P(Y=c_1|x)}{P(Y=c_2|x)} > 1 \\ c_2, & \text{otherwise} \end{cases}$$

- 进一步我们考虑$Y = c_1$的对数几率

$$\begin{aligned} \operatorname{logit}(Y = c_1) &= \log \frac{P(Y = c_1|x)}{P(Y = c_2|x)} \\ &= w \cdot x + b, \end{aligned}$$

  则上面的分类准则也可以表示为

$$y = \begin{cases} c_1, & \text{if } w \cdot x + b > 0 \\ c_2, & \text{otherwise} \end{cases}$$

- 注意$\operatorname{logit}(Y = c_1)$是输入$x$的线性函数，因此二项逻辑斯谛回归模型属于**对数线性模型**.

对于多类分类任务，可以考虑多项逻辑斯谛回归模型：
不妨设$\mathcal{Y}=\{c_1,c_2,\cdots,c_K\}$，则多项逻辑斯谛回归模型是如下的后验概率分布：

$$
\begin{aligned}
P(Y=c_k|x) &= \frac{\exp(w_k\cdot x+b_k)}{1+\sum_{l=1}^{K-1}\exp(w_l\cdot x+b_l)},\ k=1,2,\ldots,K-1 \\
P(Y=c_K|x) &= \frac{1}{1+\sum_{l=1}^{K-1}\exp(w_l\cdot x+b_l)},
\end{aligned}
$$

其中$x\in\mathbf{R}^n$是输入实例，Y是输出随机变量，参数 $w_k\in\mathbf{R}^n$，$b_k\in\mathbf{R}$，$k=1,2,\ldots,K-1$。

对多项逻辑斯谛回归模型来说，下面的对数几率仍然是输入x的线性函数：

$$
\begin{aligned}
\log \frac{P(Y=c_1|x)}{P(Y=c_K|x)} &= w_1 \cdot x + b_1, \\
\log \frac{P(Y=c_2|x)}{P(Y=c_K|x)} &= w_2 \cdot x + b_2, \\
&\vdots \\
\log \frac{P(Y=c_{K-1}|x)}{P(Y=c_K|x)} &= w_{K-1} \cdot x + b_{K-1}.
\end{aligned}
$$

参数的极大似然估计

给定$D=\{(x_i,y_i)\}_{i=1}^N$，其中 $x_i\in\mathbf{R}^n, y_i\in\mathcal{Y}=\{0,1\}, 1\leq i\leq N.$

- 我们以$\theta=(w,b)$表示二项逻辑斯谛回归模型的参数，令

$$p(x;\theta)=P(Y=1|x),$$

则似然函数为

$$L(\theta)=\prod_{i=1}^{N}p(x_i;\theta)^{y_i}(1-p(x_i;\theta))^{1-y_i}.$$

对数似然函数为

$$\log L(\theta)=\sum_{i=1}^{N}y_i\log p(x_i;\theta)+(1-y_i)\log(1-p(x_i;\theta))$$

$$
\begin{aligned}
\log L(\theta) &= \sum_{i=1}^{N} y_i \log p(x_i;\theta) + (1-y_i)\log(1-p(x_i;\theta)) \\
&= \sum_{i=1}^{N} y_i \log \frac{p(x_i;\theta)}{1-p(x_i;\theta)} + \log(1-p(x_i;\theta)) \\
&= \sum_{i=1}^{N} y_i(w\cdot x_i + b) - \log(1+\exp(w\cdot x_i + b))
\end{aligned}
$$

则参数θ的极大似然估计为

$$
\hat{\theta} = (\hat{w}, \hat{b}) = \operatorname*{argmax}_{\theta=(w,b)} \log L(\theta)
$$

通常采用梯度下降法、牛顿法或拟牛顿法来求解.

为了最大化对数似然函数$\log L(\theta)$，我们令$\log L(\theta)$对w和b的偏导数为0可得到

$$\frac{\partial \log L(\theta)}{\partial w} = \sum_{i=1}^{N} x_i(y_i - p(x_i;\theta)) = 0 \tag{1}$$

$$\frac{\partial \log L(\theta)}{\partial b} = \sum_{i=1}^{N} (y_i - p(x_i;\theta)) = 0 \tag{2}$$

特别地，由$\frac{\partial \log L(\theta)}{\partial b} = 0$可以得到

$$\sum_{i=1}^{N} y_i = \sum_{i=1}^{N} p(x_i;\theta) \tag{3}$$

这意味着对两类的输出来说，观测次数和期望次数均一致。

概要

- 利用贝叶斯公式来计算后验概率也是常见的学习方法之一.
- 给定 $D = \{(x_i, y_i)\}_{i=1}^N$，其中 $x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})^T \in \mathcal{X}$, $y_i \in \mathcal{Y} = \{c_1, c_2, \cdots, c_K\}$，$i = 1, 2, \cdots, N$.
- 由条件概率公式可得

$$P(Y = c_i | x) = \frac{P(Y = c_i, x)}{P(x)}.$$

注意到

$$P(Y = c_i, x) = P(x | Y = c_i) P(Y = c_i),$$

且由全概率公式可得

$$P(x) = \sum_{k=1}^{K} P(x | Y = c_k) P(Y = c_k),$$

由此得到

贝叶斯公式

$$P(Y=c_i|x)=\frac{P(x|Y=c_i)P(Y=c_i)}{\sum_{k=1}^{K}P(x|Y=c_k)P(Y=c_k)}$$

- 先由训练样本集$D=\{(x_i,y_i)\}_{i=1}^N$学得先验概率分布$P(Y)$和条件概率分布$P(X|Y)$,
- 对于给定的实例x，由贝叶斯公式得到关于每一c_i的后验概率$P(Y=c_i|x)$.

因此按照贝叶斯公式来计算后验概率的分类模型是生成式模型.

概要

1. 后验概率最大化分类准则
2. 逻辑斯谛回归模型
3. 贝叶斯公式
4. 朴素贝叶斯分类器

待估参数

- 需要学习的先验概率分布共有K个参数

$$\{P(Y=c_k)\}_{k=1}^{K}.$$

- 需要学习的条件概率分布参数：
  - 不妨假设数据每维特征都是离散的，且第i维特征$X^{(i)}$的可能取值的个数为m_i，则对每个类别c_k，我们需要估计的条件概率分布$P(X|Y=c_k)$的参数

    $$P(X^{(1)}=x^{(1)},X^{(2)}=x^{(2)},\cdots,X^{(n)}=x^{(n)}|Y=c_k)$$

    共有$\prod_{i=1}^{n} m_i$个.
  - 要估计如此多的参数在很多实际学习任务中是不现实的.

- 朴素贝叶斯方法假定在类已确定的条件下各维数据特征都是条件独立的，即

$$P(X^{(1)}=x^{(1)},X^{(2)}=x^{(2)},\cdots,X^{(n)}=x^{(n)}|Y=c_k)$$
$$=\prod_{i=1}^{m}P(X^{(i)}=x^{(i)}|Y=c_k).$$

- 基于条件独立性假设，只需要对每个类c_k，对每维特征估计条件概率分布$P(X^{(i)}=x^{(i)}|Y=c_k)$即可，这将需要估计的参数由$\prod_{i=1}^{n}m_i$个 消减到$\sum_{i=1}^{n}m_i$个.

- 给定训练样本集$D=\{(x_i,y_i)\}_{i=1}^N$，其中对每个$1\le i\le N$来说，$x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)})\in\mathcal{X}$，$y_i\in\{c_1,c_2,\cdots,c_K\}$.
- 设第j维特征$X^{(j)}$的可能取值为$a_1^{(j)},a_2^{(j)},\cdots,a_{m_j}^{(j)}$,则
  - 先验概率分布的极大似然估计为:

$$\hat{P}(Y=c_k)=\frac{\sum\limits_{j=1}^{N}I(y_j=c_k)}{N},\quad k=1,2,\cdots,K.$$

  - 对每个类c_k，第i维特征的条件概率分布的极大似然估计为:

$$\hat{P}(X^{(i)}=a_l^{(i)}|Y=c_k)=\frac{\sum\limits_{j=1}^{N}I(x_j^{(i)}=a_l^{(i)},y_j=c_k)}{\sum\limits_{j=1}^{N}I(y_j=c_k)},$$

其中$l=1,2,\ldots,m_i$，$i=1,2,\ldots,n$.

- 对于新的输入实例$x=(x^{(1)},x^{(2)},\cdots,x^{(n)})$，则后验概率

$$\hat{P}(Y=c_k|x)=\frac{\left(\prod_{i=1}^{n}\hat{P}(X^{(i)}=x^{(i)}|Y=c_k)\right)\hat{P}(Y=c_k)}{\sum_{l=1}^{K}\left(\prod_{i=1}^{n}\hat{P}(X^{(i)}=x^{(i)}|Y=c_l)\right)\hat{P}(Y=c_l)}.$$

- 依据后验概率最大化分类规则，则x的类别预测为

$$y=\operatorname*{argmax}_{c_k\in\mathcal{Y}}\hat{P}(Y=c_k|x).$$

- 注意到

$$\hat{P}(Y=c_k|x)\propto\left(\prod_{i=1}^{N}\hat{P}(X^{(i)}=x^{(i)}|Y=c_k)\right)\hat{P}(Y=c_k),$$

则x的类别预测也可表示为

$$y=\operatorname*{argmax}_{c_k\in\mathcal{Y}}\left(\prod_{i=1}^{n}\hat{P}(X^{(i)}=x^{(i)}|Y=c_k)\right)\hat{P}(Y=c_k).$$

- 如果x的某维特征$X^{(i)}$的取值$x^{(i)}$和类c_k在训练样本中没有同时出现，则相应条件概率的极大似然估计

$$\hat{P}(x^{(i)}|Y = c_k) = 0.$$

- 从而后验概率

$$\hat{P}(Y = c_k|x) = 0.$$

- 这使得类c_k事实上被移出x的可能类标记的候选，从而可能导致分类错误.
- 采用贝叶斯估计可以避免这种情况.

- 与极大似然估计相比，贝叶斯估计
  - 对每个类出现的频数 $\sum_{j=1}^{N} I(y_j = c_k)$ 和每维特征和类一起出现的频数 $\sum_{j=1}^{N} I(x_j^{(i)} = a_l^{(i)}, y_j = c_k)$ 都加以$\lambda \geq 0$来对相关概率进行平滑.
  - $\lambda = 1$ 时称为Laplace平滑.
- 具体来说，
  - 先验概率分布的贝叶斯估计为

$$\hat{P}_\lambda(Y = c_k) = \frac{\sum_{j=1}^{N} I(y_j = c_k) + \lambda}{N + K\lambda}, \quad k = 1, 2, \cdots, K.$$

- 对每个类c_k，第i维特征的条件概率分布的贝叶斯估计为

$$\hat{P}_\lambda(X^{(i)} = a_l^{(i)} | Y = c_k) = \frac{\sum\limits_{j=1}^{N} I(x_j^{(i)} = a_l^{(i)}, y_j = c_k) + \lambda}{\sum\limits_{j=1}^{N} I(y_j = c_k) + m_i \lambda},$$

其中$l = 1, 2, \ldots, m_i$，$i = 1, 2, \ldots, n$.

- 如果$\lambda = 0$，此时贝叶斯估计就是极大似然估计.

小结

- 后验概率最大化原则

$$\hat{y} = \underset{c_i}{\operatorname{argmax}} P(Y = c_i|x)$$

- 给定训练样本集，如何学得后验概率分布通常有两种方式，即判别式和生成式:
  - 判别式模型直接学习$P(Y|X)$.
  - 生成式模型学习$P(X, Y)$.
- 判别式模型：逻辑斯谛回归模型
  - 二项逻辑斯谛回归模型.
  - 多项逻辑斯谛回归模型.
- 生成式模型:朴素贝叶斯分类器
  - 极大似然估计.
  - 贝叶斯估计.