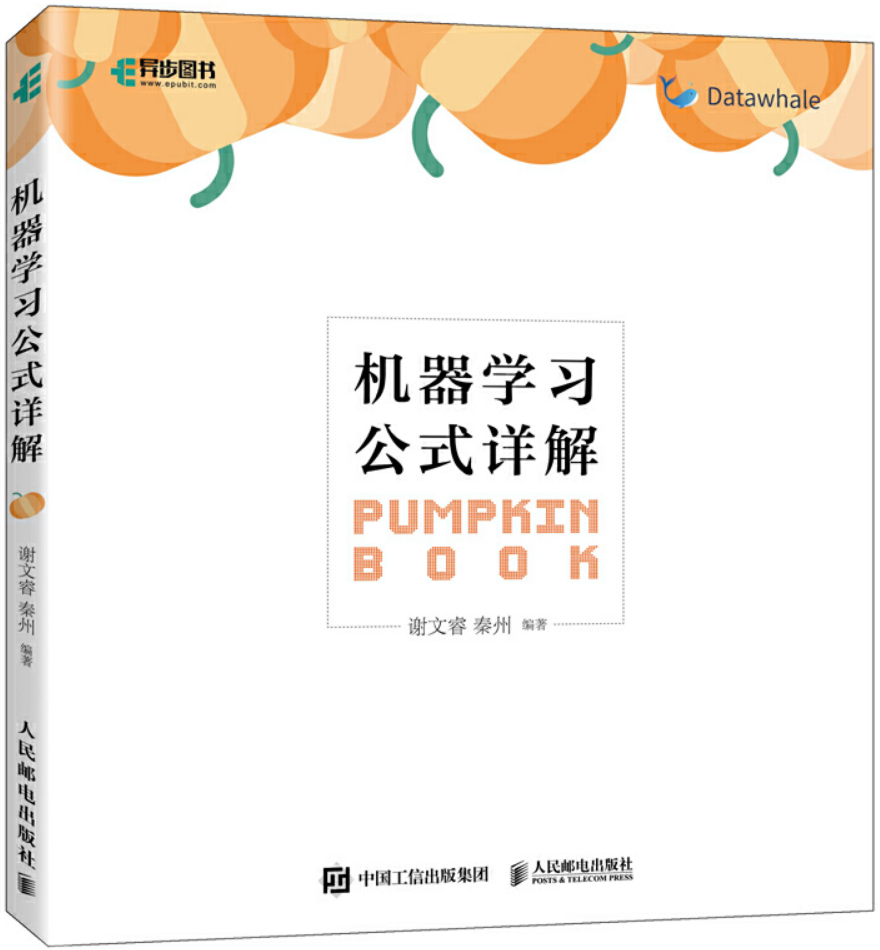

《机器学习公式详解》（南瓜书）

第7章 贝叶斯分类器

本节主讲：谢文睿

本节大纲

西瓜书对应章节：7.1、7.2、7.3

1. 贝叶斯决策论
2. 生成式模型和判别式模型
3. 朴素贝叶斯分类器
4. 半朴素贝叶斯分类器

贝叶斯决策论

贝叶斯决策论是概率框架下实施决策的基本方法，对分类任务来说，在所有相关概率都已知的理想情形下，贝叶斯决策论考虑如何基于这些概率和误判损失来选择最优的类别标记。

贝叶斯决策论

以一个多分类任务为例：假设当前有一个N分类问题，即$\mathcal{Y}=\{c_1,c_2,\ldots,c_N\}$

【定义】：λ_{ij}是将一个真实标记为c_j的样本误分类为c_i所产生的损失。

【定义】：单个样本$\boldsymbol{x}$的期望损失（条件风险）为

$$R\left(c_i|\boldsymbol{x}\right)=\sum_{j=1}^{N}\lambda_{ij}P\left(c_j|\boldsymbol{x}\right)$$

其中，$P\left(c_j|\boldsymbol{x}\right)$为后验概率

【定义】：全部样本构成的总体风险为

$$R(h)=\mathbb{E}_{\boldsymbol{x}}\left[R(h(\boldsymbol{x})|\boldsymbol{x})\right]$$

其中，h为分类器（模型）。显然，分类效果越准确的h，其条件风险和总体风险也越小

贝叶斯决策论

贝叶斯判定准则：为最小化总体风险$R(h)$，只需在每个样本上选择那个能使条件风险$R(c|\boldsymbol{x})$最小的类别标记，即

$$h^*(\boldsymbol{x}) = \arg\min_{c\in\mathcal{Y}} R(c|\boldsymbol{x})$$

此时，h^*称为贝叶斯最优分类器

贝叶斯决策论

具体地，若目标是最小化分类错误率，则误判损失λ_{ij}可写为

$$\lambda_{ij} = \begin{cases} 0, & \text{if } i = j \\ 1, & \text{otherwise} \end{cases}$$

此时单个样本$\boldsymbol{x}$的期望损失（条件风险）为

$$R\left(c_i|\boldsymbol{x}\right) = \sum_{j=1}^{N} \lambda_{ij} P\left(c_j|\boldsymbol{x}\right)$$

$$R\left(c_i|\boldsymbol{x}\right) = 1 * P(c_1|\boldsymbol{x}) + ... + 1 * P(c_{i-1}|\boldsymbol{x}) + 0 * P(c_i|\boldsymbol{x}) + 1 * P(c_{i+1}|\boldsymbol{x}) + ... + 1 * P(c_N|\boldsymbol{x})$$

又$\sum_{j=1}^{N} P(c_j|\boldsymbol{x}) = 1$，则

$$R(c_i|\boldsymbol{x}) = 1 - P(c_i|\boldsymbol{x})$$

贝叶斯决策论

于是，按照贝叶斯判定准则，最小化分类错误率的贝叶斯最优分类器为

$$h^*(\boldsymbol{x}) = \arg\min_{c \in \mathcal{Y}} R(c|\boldsymbol{x})$$

$$h^*(\boldsymbol{x}) = \arg\min_{i \in \{1,2,\ldots,N\}} R(c_i|\boldsymbol{x})$$

$$h^*(\boldsymbol{x}) = \arg\min_{i \in \{1,2,\ldots,N\}} 1 - P(c_i|\boldsymbol{x})$$

$$h^*(\boldsymbol{x}) = \arg\max_{i \in \{1,2,\ldots,N\}} P(c_i|\boldsymbol{x})$$

$$h^*(\boldsymbol{x}) = \arg\max_{c \in \mathcal{Y}} P(c|\boldsymbol{x})$$

即对每个样本$\boldsymbol{x}$，选择后验概率$P(c_i|\boldsymbol{x})$最大的类别c_i作为标记

生成式模型和判别式模型

从贝叶斯决策论（概率框架）的角度：机器学习所要做的就是基于有限的训练样本集尽可能准确地估计出后验概率$P(c|\boldsymbol{x})$

从机器学习自己的角度：给定一个样本$\boldsymbol{x}$，求一个能准确分类$\boldsymbol{x}$的$f(\boldsymbol{x})$，其有些算法可以看作是对后验概率建模$P(c|\boldsymbol{x})$（例如对数几率回归），而有些算法则是纯粹完成样本分类（例如SVM）

生成式模型和判别式模型

判别式模型：给定$\boldsymbol{x}$，直接建模$P(c|\boldsymbol{x})$来预测c

生成式模型：先对联合概率$P(\boldsymbol{x}, c)$建模，然后再由此推导得出$P(c|\boldsymbol{x})$

生成式模型和判别式模型

对于生成式模型，其建模思路为

$$P(c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}, c)}{P(\boldsymbol{x})}$$

再根据贝叶斯定理，上式可恒等变形为

$$P(c|\boldsymbol{x}) = \frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})}$$

其中，$P(c)$是类“先验”概率，$P(\boldsymbol{x}|c)$是样本$\boldsymbol{x}$相对于类别标记c的类条件概率，$P(\boldsymbol{x})$是用于归一化的“证据”因子。

朴素贝叶斯分类器

属性条件独立性假设：对已知类别，假设所有属性相互独立

$$P(c|\boldsymbol{x}) = \frac{P(c)P(\boldsymbol{x}|c)}{P(\boldsymbol{x})} = \frac{P(c)}{P(\boldsymbol{x})}\prod_{i=1}^{d} P\left(x_i|c\right)$$

其中，d为属性数目，x_i为$\boldsymbol{x}$在第i个属性上的取值。基于贝叶斯判定准则

$$h^*(\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} P(c|\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} \frac{P(c)}{P(\boldsymbol{x})}\prod_{i=1}^{d} P\left(x_i|c\right)$$

由于对所有类别来说$P(\boldsymbol{x})$都相同，所以$P(\boldsymbol{x})$视作常量可以略去

$$h_{nb}(\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} P(c)\prod_{i=1}^{d} P\left(x_i|c\right)$$

朴素贝叶斯分类器

估计$P(c)$:

$$P(c) = \frac{|D_c|}{|D|}$$

其中，D_c表示训练集D中类别标记为c的样本集合，$|D_c|$表示集合D_c的样本总数

朴素贝叶斯分类器

估计$P(x_i|c)$：

【第i个属性为离散属性】：

$$P\left(x_i|c\right)=\frac{|D_{c,x_i}|}{|D_c|}$$

其中，D_{c,x_i}表示D_c中在第i个属性上取值为x_i的样本组成的集合

【第i个属性为连续属性】（以正态分布假设为例）：

$$p\left(x_i|c\right)=\frac{1}{\sqrt{2\pi}\sigma_{c,i}}\exp\left(-\frac{\left(x_i-\mu_{c,i}\right)^2}{2\sigma_{c,i}^2}\right)$$

其中，$\mu_{c,i}$和$\sigma_{c,i}^2$分别是第c类样本在第i个属性上取值的均值和方差。

半朴素贝叶斯分类器

半朴素贝叶斯分类器：适当考虑一部分属性间的相互依赖信息，从而既不需进行完全联合概率计算，又不至于彻底忽略了比较强的属性依赖关系。

【独依赖估计（ODE）】：假设每个属性在类别之外最多依赖于一个其他属性，即

$$P(c|\boldsymbol{x}) \propto P(c) \prod_{i=1}^{d} P\left(x_i|c, pa_i\right)$$

其中，pa_i为属性x_i所依赖的属性，称为x_i的父属性。

半朴素贝叶斯分类器

【超父独依赖估计（SPODE）】：假设所有属性都依赖于同一个“超父”属性

$$
\begin{aligned}
P(c \mid \boldsymbol{x}) &= \frac{P(\boldsymbol{x}, c)}{P(\boldsymbol{x})} = \frac{P\left(c, x_i\right) P\left(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d \mid c, x_i\right)}{P(\boldsymbol{x})} \\
&\propto P\left(c, x_i\right) P\left(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d \mid c, x_i\right) \\
&= P\left(c, x_i\right) \prod_{j=1}^{d} P\left(x_j \mid c, x_i\right)
\end{aligned}
$$

其中，x_i是“超父”属性

预告

下一节：集成学习

西瓜书对应章节：第8章

结束语

欢迎加入【南瓜书读者交流群】，我们将在群里进行答疑、勘误、本次直播回放、本次直播PPT发放、下次直播通知等最新资源发放和活动通知。

加入步骤：

1. 关注公众号【Datawhale】，发送【南瓜书】三个字获取机器人“小豚”的微信二维码
2. 添加“小豚”为微信好友，然后对“小豚”发送【南瓜书】三个字即可自动邀请进群