机器学习第一章课后习题

162050127 颜劭铭

2022 年 5 月 4 日

1 习题

1.1 使用最多包含 k 个合取式的析合范式来表达表 1.1 西瓜分类问题的假设空间，试估算共有多少种可能的假设.

在不包括通配符 * 的情况下，西瓜的色泽有两种可能，根蒂有三种可能，敲声有三种可能，因此由排列组合可知每个合取式可能表示的假设有 $2 \times 3 \times 3 = 18$ 种。

而假设空间，即好瓜的假设可能的情况有 $C_{18}^0 + C_{18}^1 + C_{18}^2 + \cdots + C_{18}^{18} = 2^{18}$ 种可能，其中 C_{18}^0 指的是 18 种合取式中没有好瓜，以此类推。

在引入通配符 * 的情况下，西瓜的色泽有三种可能，根蒂有四种可能，敲声有四种可能，因此由排列组合可知每个合取式可能表示的假设有 $3 \times 4 \times 4 = 48$ 种。

但是，我们可以假设西瓜的属性可以由 $(A_1, A_2, A_3)(B_1, B_2, B_3, B_4)(C_1, C_2, C_3, C_4)$ 表示，其中 A_1 代表西瓜颜色为青绿，A_2 代表西瓜颜色为乌黑，A_3 代表通配符 *，我们在这里用二值编码的方法去表示假设中的合取式。

例如西瓜的色泽用 $(1,1,0)(1,0,1)(0,1,1)(0,0,1)(1,1,1)$ 表示都是好瓜，这些假设可以进行合并去重得到都可以用 $(1,1,1)$ 进行表示，都是好瓜。

因此 48 种合取式可能表示的假设可以去重后表示为 18 种合取式，所以当使用 k 个合取式的析合范式来表达西瓜分类问题的假设空间时 (这边对于这个 k 个不太理解，所以我选择了最多的那种情况进行计算)，总共应该有 2^{18} 种假设，考虑空集的话应为 $2^{18} - 1$ 种假设。

1.2 若数据包含噪声，则假设空间中有可能不存在与所有训练样本都一致的假设，在此情形下，试估计一种归纳偏好用于假设选择.

对于假设空间有可能不存在与所有训练样本都一致的假设，我的理解是可能假设与训练样本特征属性相同但是标签不同引起的噪声。

① 第一种想法是去除噪声，忽略那些特征属性相同但标签不同的数据，但是感觉这样不仅会忽略掉噪声，也会忽略掉有效信息。

② 第二种想法 (因为第一章讲的是二分类问题) 是将那些特征向量相同，但是标签既有正例又有反例的样本都认为是正例，虽然样本中有一半是噪声，但是保证了 50% 的正确率。

③ 第三种想法是引入一个准确率的概念，准确率等于假设和样本相同的特征属性数/假设总的特征属性数，先保证假设和样本的标签是正确的，这个时候归纳偏好尽可能选择那些准确率高的假设。

1.3 若换用其他性能度量 ℓ，则式 (1.1) 将改为

$$E_{\mathbf{ote}}\left(\mathfrak{L}_a \mid X, f\right)=\sum_h \sum_{x \in \chi-X} P(x) \ell(h(x), f(x)) P\left(h \mid X, \mathfrak{L}_a\right)$$

试证明“没有免费的午餐定理”仍成立.

已知原式子为：

$$E_{\mathrm{ote}}\left(\mathfrak{L}_a \mid X, f\right)=\sum_h \sum_{x \in \chi-X} P(x) \ell(h(x), f(x)) P\left(h \mid X, \mathfrak{L}_a\right)$$

借助和西瓜书上式子 1.1 相同的证明方法，上式可以转化为：

$$\begin{aligned}
\sum_f E_{\mathrm{ote}}\left(\mathfrak{L}_a \mid X, f\right) &= \sum_f \sum_h \sum_{x \in \chi-X} P(x) \cdot \ell(h(x), f(x)) \cdot P\left(h \mid X, \mathfrak{L}_a\right) \\
&= \sum_{x \in \chi-X} P(x) \sum_h P\left(h \mid X, \mathfrak{L}_a\right) \sum_f \ell(h(x), f(x)) \\
&= \sum_{x \in \chi-X} P(x) \sum_h P\left(h \mid X, \mathfrak{L}_a\right) \cdot \frac{1}{2} 2^{|x|}(\ell(h(x)=f(x))+\ell(h(x) \neq f(x))) \\
&= 2^{|x|-1} \cdot A \sum_{x \in \chi-X} P(x) \cdot 1
\end{aligned}$$

要求性能度量 ℓ 满足 $\ell(h(x), f(x))=\ell(h(x)=f(x))+\ell(h(x) \neq f(x))=A$ 为常数。

对于本章研究的二分类问题，即要求 $\ell(h(x), f(x))=\ell(0,0)+\ell(0,1)=\ell(1,1)+\ell(1,0)=A$.

1.4 试述机器学习能在互联网搜索的哪些环节起什么作用.

① 首先搜索引擎中当输入搜索内容时，要进行语义分割，对于识别出的内容返回最恰当的搜索结果。

② 搜索方法不仅仅有文字搜索，还有图片、视频搜索，从图片视频中提取信息进行搜索，以及这些年来大火的语音搜索，有人机交互的功能。

③ 返回搜索结果的时候有推荐系统，比如京东淘宝根据用户平时购买的东西找到相通性和特点，对于返回的搜索结果进行排序。