

机器学习——第一次作业

Koorye

2024 年 3 月 2 日

1 过学习有什么表现？原因是什么？有哪些避免过学习的方法？

过学习表现为模型在训练集上的误差很低，但是在测试数据上误差很高。

过学习的原因是训练数据太少，或模型能力太强（参数过多、结构过于复杂）等。

避免过学习的方法有：

1. 扩大训练集：更多训练数据可以使模型拥有更为平滑的决策面。
2. 正则化：修改目标函数来惩罚模型复杂度，如 L1/L2 正则化等，目的是约束模型的学习能力。
3. 通过验证集来选择模型：通过验证集确定模型超参数，还可以通过 K 折交叉验证、留一法等数据不足的情况下更好地评估模型性能。

2 监督学习中，什么是推断，什么是决策，简述解决决策问题的三种方法及其复杂度。

推断指利用训练数据学习后验概率 $p(t|x)$ 或 $p(x, t)$ ，使用已知类别的样本，找到输入 x 和输出结果 t 之间的关系。

决策指使用后验概率进行最优分类，将输入 x 映射到最优分类结果 t 。

解决决策问题的三种方法有：

1. 生成式模型：通过贝叶斯定理求出后验概率 $p(t|x) = \frac{p(x|t)p(t)}{p(x)}$ ，再进行决策。复杂度高。
2. 判别式模型：直接得到条件概率 $p(t|x)$ 进行决策。复杂度中等。
3. 判别函数：直接得到判别函数 $f(x)$ ，将输入 x 映射到最优分类结果 t 。复杂度低。

3 （贝叶斯定理）一种癌症，得了这个癌症的人被检测出为阳性的几率为 80%，未得这种癌症的人被检测出阴性的几率为 90%，而人群中得这种癌症的几率为 1%，一个人被检测出阳性，问这个人得癌症的几率为多少？

设事件 A 为得癌症，事件 B 为检测出阳性。则有：

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)}. \quad (1)$$

由已知条件得

$$P(A) = 0.01, P(B|A) = 0.8, P(\sim B | \sim A) = 0.9. \quad (2)$$

则有

$$P(\sim A) = 0.99, P(B | \sim A) = 0.1. \quad (3)$$

代入得

$$P(A|B) = \frac{0.8 \cdot 0.01}{0.8 \cdot 0.01 + 0.1 \cdot 0.99} \approx 0.0748. \quad (4)$$

故约有 7.48% 的概率得癌症。

4 什么是极大似然参数估计和最大后验参数估计，简述它们的特点和联系，并说明什么是共轭先验。

极大似然估计 (MLE) 指在给定观测数据 D 的情况下，调整参数 θ 使得似然函数 $p(D|\theta)$ 最大。

$$\hat{\theta} = \arg \max_{\theta} P(D|\theta). \quad (5)$$

最大后验估计 (MAP) 指在给定观测数据 D 的情况下，调整参数 θ 使得后验概率 $p(\theta|D)$ 最大。

$$\hat{\theta} = \arg \max_{\theta} P(\theta|D) = \arg \max_{\theta} \frac{P(D|\theta)P(\theta)}{P(D)}. \quad (6)$$

极大似然估计的特点是只考虑样本数据，不考虑先验信息；最大后验估计的特点是考虑了先验信息。因此极大似然估计在数据量足够时才能比较准确，数据不足时容易产生过拟合；而最大后验估计可以通过先验信息来约束参数的取值，避免过拟合。

两者的联系是当先验分布为均匀分布时，最大后验估计退化为极大似然估计；当数据足够多时，两者的结果也会趋于一致。

共轭先验指在贝叶斯估计中，如果后验分布和先验分布属于同一分布族，则称先验分布是似然函数的共轭先验。例如，将二项分布的先验设为 Beta 分布，即

$$P(\theta) = \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1}, \quad (7)$$

有似然

$$P(x|n, \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x}, \quad (8)$$

则后验

$$P(\theta|x, n, \alpha, \beta) \propto \text{Beta}(\alpha + x, \beta + n - x). \quad (9)$$