机器学习——第三次作业

Koorye

2024 年 3 月 16 日

1 试写出偏置-方差分解中这两部分的公式并说明它们的含义，说明模型复杂度与偏置、方差以及过拟合、欠拟合间的关系。

$$\mathbb{E}_{\mathcal{D}}[\{y(x;\mathcal{D})-h(x)\}^2]=\underbrace{\{\mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})]-h(x)\}^2}_{bias^2}+\underbrace{\mathbb{E}_{\mathcal{D}}[\{y(x;\mathcal{D})-\mathbb{E}_{\mathcal{D}}[y(x;\mathcal{D})]\}^2]}_{variance}+\underbrace{\epsilon^2}_{noise}, \tag{1}$$

简单表示为

$$\text{expected loss}=\text{bias}^2+\text{variance}+\text{noise}. \tag{2}$$

如上式所示，偏置-方差分解是指模型的期望预测误差可以分解为偏置、方差和噪声三部分。其中：

1. **偏置**：所有数据集的平均预测与预期的回归函数之间的差异。度量了某种学习算法的平均估计结果所能逼近学习目标的程度，反映模型的准确程度。
2. **方差**：模型给出的解在平均值附近的波动。度量了在面对同样规模的不同训练集时，学习算法的估计结果发生变动的程度，反映模型的敏感程度。

模型复杂度与偏置、方差以及过拟合、欠拟合之间的关系是：

1. 偏置越**低**，模型复杂度越**高**。因为模型越复杂，越能逼近学习目标。
2. 偏置越**高**，模型复杂度越**低**。因为模型越简单，越难以学习目标。
3. 方差越**低**，模型复杂度越**低**。因为模型越简单，越不容易受到不同数据集的影响。
4. 方差越**高**，模型复杂度越**高**。因为模型越复杂，越容易受到不同数据集的影响。

2 介绍分类的三类方法及其特点，并列举每类的具体方法。

三类方法有：

1. **判别函数**：直接找到一个函数 $f(x)$，把每个输入 x 直接映射为类别标签。方法有 **Fisher 判别函数、感知机、SVM** 等。
2. **概率判别式模型**：直接对后验概率 $p(C_k|x)$ 建模，再使用决策论来确定每个输入 x 的类别标签。方法有 **Logistic 回归、神经网络**等。

3. **概率生成式模型**：先对类条件密度 $p(x|C_k)$ 和先验类概率分布 $p(C_k)$ 建模，再使用贝叶斯定理计算厚颜类概率分布 $p(C_k|x)$。

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)},$$

最后使用决策论确定每个输入 x 的类别。方法有**朴素贝叶斯分类器**等。

其中 1 和 2 的特点是计算量小，但是容易过拟合；3 的特点则是计算量大，但是不容易过拟合。

3 有以下 8 个关于客户的身高，体重，鞋码和性别的样本，现有一位身高“中”，体重“中”，鞋码“中”的客户，试用朴素贝叶斯方法估计该客户的性别。

编号	身高	体重	鞋码	性别
1	高	重	大	男
2	高	中	大	男
3	中	中	大	男
4	中	中	中	男
5	矮	轻	小	女
6	矮	轻	小	女
7	矮	中	中	女
8	中	中	中	女

首先求出先验概率 $p(C_k)$：

$$p(\text{性别} = \text{男}) = 4/8，p(\text{性别} = \text{女}) = 4/8。 \tag{3}$$

然后求出类条件概率 $p(x|C_k)$：

$$\begin{aligned} &p(\text{身高} = \text{中}|\text{性别} = \text{男}) = 2/4, p(\text{体重} = \text{中}|\text{性别} = \text{男}) = 3/4, p(\text{鞋码} = \text{中}|\text{性别} = \text{男}) = 1/4, \\ &p(\text{身高} = \text{中}|\text{性别} = \text{女}) = 1/4, p(\text{体重} = \text{中}|\text{性别} = \text{女}) = 2/4, p(\text{鞋码} = \text{中}|\text{性别} = \text{女}) = 2/4. \end{aligned} \tag{4}$$

最后得到后验概率 $p(C_k|x)$：

$$\begin{aligned} p(\text{性别} = \text{男}|\text{身高} = \text{中}, \text{体重} = \text{中}, \text{鞋码} = \text{中}) &= \frac{p(\text{身高} = \text{中}, \text{体重} = \text{中}, \text{鞋码} = \text{中})p(\text{性别} = \text{男})}{p(\text{身高} = \text{中}, \text{体重} = \text{中}, \text{鞋码} = \text{中})} \\ &\propto p(\text{身高} = \text{中}|\text{性别} = \text{男})p(\text{体重} = \text{中}|\text{性别} = \text{男})p(\text{鞋码} = \text{中}|\text{性别} = \text{男})p(\text{性别} = \text{男}) \\ &= \frac{2}{4} \cdot \frac{3}{4} \cdot \frac{1}{4} \cdot \frac{4}{8} = \frac{3}{64}, \end{aligned} \tag{5}$$

$$\begin{aligned} p(\text{性别} = \text{女}|\text{身高} = \text{中}, \text{体重} = \text{中}, \text{鞋码} = \text{中}) &= \frac{p(\text{身高} = \text{中}, \text{体重} = \text{中}, \text{鞋码} = \text{中})p(\text{性别} = \text{女})}{p(\text{身高} = \text{中}, \text{体重} = \text{中}, \text{鞋码} = \text{中})} \\ &\propto p(\text{身高} = \text{中}|\text{性别} = \text{女})p(\text{体重} = \text{中}|\text{性别} = \text{女})p(\text{鞋码} = \text{中}|\text{性别} = \text{女})p(\text{性别} = \text{女}) \\ &= \frac{1}{4} \cdot \frac{2}{4} \cdot \frac{2}{4} \cdot \frac{4}{8} = \frac{3}{32}. \end{aligned} \tag{6}$$

$\frac{3}{32} > \frac{3}{64}$，所以该客户的性别是**女**的概率更大。

4 考虑一个硬间隔（hard margin）支持向量机和下面来自两类的训练样本：

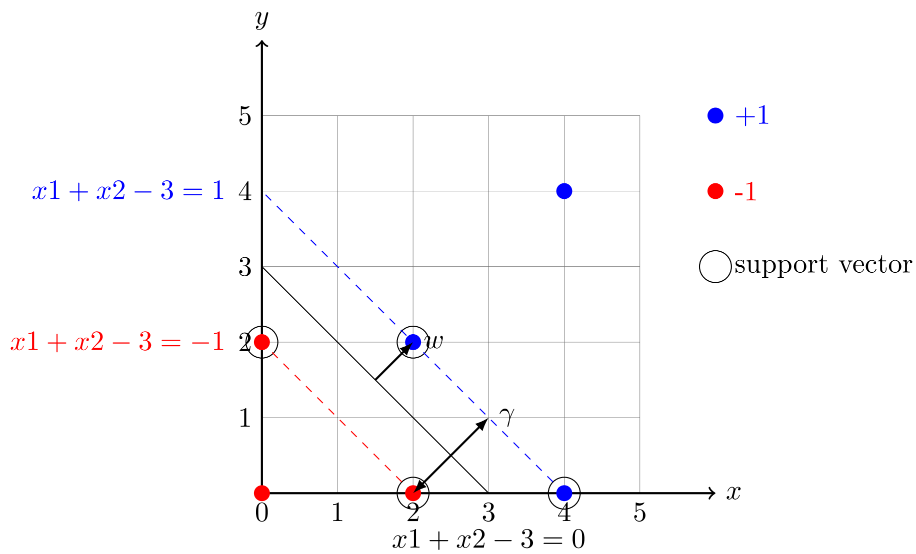
$$+1 : (2,2)\ (4,4)\ (4,0)$$

$$-1 : (0,0)\ (2,0)\ (0,2)$$

(a) 画出这 6 个点，通过观察画出最优分类面和权向量 w，给出分类面的方程 w1x1+w2x2+b=0。计算出它的间隔 margin（从分类面到最近数据点的距离）。

(b) 标出所有的支持向量，并说明原因？

画出这 6 个点如下图所示：



根据观察，最优分类面为 $x1+x2-3=0$，间隔为 $\gamma = \frac{2}{\|w\|} = \sqrt{2}$。从分类面到最近数据点的距离即为 $\frac{\sqrt{2}}{2}$。

支持向量有 $(2,2),(4,0),(0,2),(2,0)$，在图中被圈出。原因是这些点是离分类面最近的点，这些点决定了分类面的位置，而其他点对分类面的位置没有影响。