机器学习

Homework1

杨亿酬 3230105697

2025-10-19 - 2025-10-20

1. 数据集包含 100 个样本，其中正、反例各一半，假定学习算法所产生的模型是将新样本预测为训练样本数较多的类别（训练样本数相同时进行随机猜测），试给出用 10 折交叉验证法和留一法分别对错误率进行评估的结果。

10折交叉验证法：每折10个样本，90个样本用于训练，10个样本用于测试
假定样本正反例平均分布，则训练样本中正反例各45个，测试样本正反例各5个，根据模型算法，错误率为50%
如果训练样本中正例较多，则全部预测为正，此时测试样本中反例较多，错误率大于50%，反之亦然
留一法：99个样本用作训练，1个样本用作测试
如果测试样本为正例，则训练样本中有49个正例，50个反例，模型会预测为反，结果错误
总体错误率100%
这一结果说明：留一法对数据分布很敏感，不同的测试方法会给出完全不同的评估结果

2. 令码长为 9，类别数为 4，试给出海明距离意义下理论最优的 一种ECOC 二元码。

海明距离是两个等长字符串对应位置不同字符的个数
ECOC将多分类问题编码为多个二分类问题
海明距离意义下理论最优的二元码是让任意两个类别的二元码之间海明距离都尽可能大，从而降低分类错误的可能性
码长为9，类别数为4，则总码数为9*4=36
首先证明任意两个类别之间海明距离≥7是达不到的：
不失一般性，假设类A的编码为000000000，类B与A的海明距离为7，编码为111111100，则对类C的任意编码$x_1x_2...x_9$，若C与A的海明距离为7，则C的编码包含7个1，那么C与B的海明距离至多为4
下面给出一种海明距离为6的构造方法:
A={000000000},B={000111111},C={111000111},D={111111000}

3. 假设某机器学习模型的原始类别和预测类别如下表所示，求它的混淆矩阵、准确率、精确率、召回率、F1 score。

样本序号	1	2	3	4	5	6	7	8	9	10
原始类别	1	1	1	-1	-1	-1	1	1	-1	1
预测类别	1	1	-1	-1	-1	1	-1	1	-1	1

$TP=4, FN=2, FP:1, TN:3$

混淆矩阵: $\begin{bmatrix} TP & FN \\ FP & TN \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix}$

准确率 $= \frac{TP+TN}{TP+FN+FP+TN} = \frac{7}{10} = 70\%$

召回率 $= \frac{TP}{TP+FN} = \frac{4}{6} = 66.7\%$

$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall} = \frac{28}{41} = 68.29\%$

4.对以下数据集，构造 ID3 决策树，判断是否买房：

用户 ID	年龄	性别	收入	是否买房
1	27	男	15W	否
2	47	女	30W	是
3	32	男	12W	否
4	24	男	45W	是
5	45	男	30W	否
6	56	男	32W	是
7	31	男	15W	否
8	23	女	30W	是

注：年龄分为 20-30，30-40，40+三个阶段，收入分为 10-20，20-40，40+三个级别。

设买房为+，不买为-，4+4-,$Entropy(S) = 1.0$

attribute:$a \in \{$年龄，性别，收入$\}$

年龄:

20-30:2+，1-,$E = 0.918$

30-40:0+,2-,$E = 0$

40+:2+,1-,$E = 0.918$

$Gain(S, age) = 1.0 - \frac{3}{8} * 0.918 - \frac{3}{8} * 0.918 = 0.312$

性别:

男：2+，4-，$E = 0.918$

女：2+，0-，$E = 0$

$Gain(S, sex) = 1.0 - \frac{6}{8} * 0.918 = 0.312$

收入:

10-20:0+,3-,$E=0$

20-40:3+,1-,$E=0.811$

40+:1+,0-,$E=0$

$Gain(S,income)=1.0-\frac{4}{8}*0.811=0.595$

选择$Gain$最高的属性收入作为第一层节点

10-20与40+级别$Entropy=0$已经达到根节点

对20-40级别：3+,1-,$E=0.811$

年龄:

20-30:1+，0-,$E=0$

30-40:0+,0-

40+:2+,1-,$E=0.918$

$Gain(S,age)=0.811-\frac{3}{4}*0.918=0.123$

性别:

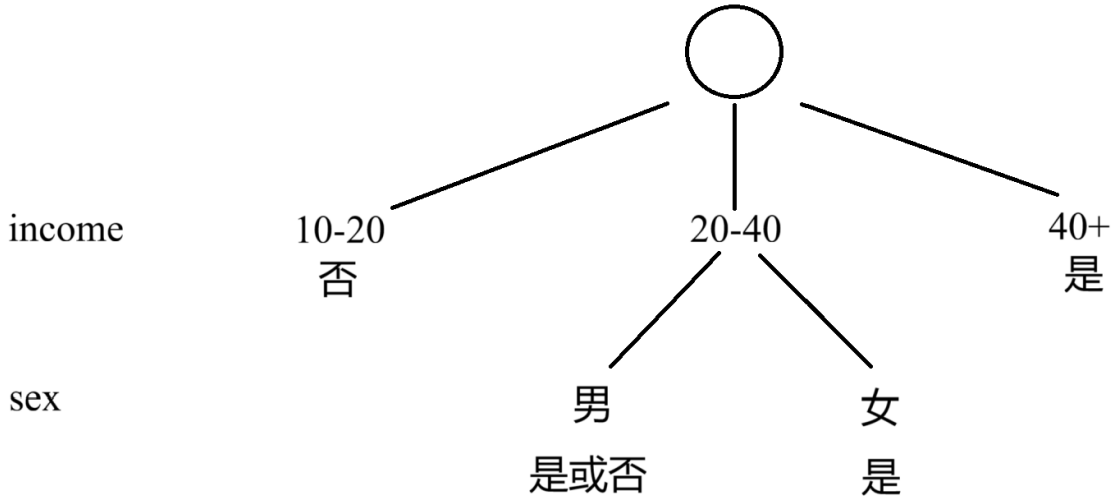
男：1+，1-，$E=1$

女：2+，0-，$E=0$

$Gain(S,sex)=0.811-\frac{2}{4}*1.0=0.311$

选择$Gain$最高的性别作为第二层节点

对20-40级别收入的男性，年龄都在40+阶段，此时无法继续划分



5、 判断下面说法是否正确：

(i) If a learning algorithm is suffering from high bias, adding more training examples will improve the test error significantly.

[Answer] False
The resulting model after adding more training examples will still in high bias.
High bias means that the model is underfitting mainly because the model is too easy.
The word `bias` should be changed to `variance` because under that case,the model is overfitting.Adding more training examples can help the model learn more diverse data,reducing its dependency on a specific sample distribution.

(ii) We always prefer models with high variance (over those with high bias) as they will able to better fit the training set.

[Answer] False
High variance means the model is likely to be overfitting,resulting in poor generalization ability of the model.
We should try to achieve the balance of variance and bias.

(iii) A model with more parameters is more prone to overfitting and typically has higher variance.

[Answer] True
More parameters means the model is more complicated and depending more on the specific data pattern.This means is model is more likely to be overfitting and has a high variance when we test it.

(iv) Introducing regularization to the model always results in equal or better performance on the training set.

[Answer] False
Regularization is used for avoiding overfitting.However,if a model is currently underfitting,introducing regularization may cause the model to perform even worse.

(v) Using a very large value of regularization parameter λ cannot hurt the performance of your hypothesis.

[Answer] False
A very large value of regularization parameter will greatly 'compress' the weights,resulting in the degradation of the decision boundary.This means that the model is underfitting