机器学习

课程作业一

redacted

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-fold Cross-Validation:
     - 将 100 个样本分成 10 份，每份 10 个样本。假设划分是分层的，即每份包含 5 个正例和 5 个反例，以保持分布一致。训练集包含 9 份数据，共 90 个样本。由于每份都是平衡的，训练集中将包含 45 个正例和 45 个反例。因为训练集中正反例数量相等，模型会进行随机猜测。对于测试集中的每一个样本，模型都有 50% 的概率猜对，50% 的概率猜错。
     - 因此错误率为 50%。
   - 留一法:
     - 进行 100 次实验，每次取 1 个样本作为测试集，其余 99 个作为训练集。
     - 若测试样本为“正例”
       - 训练集为 49 个正例，50 个反例，其中反例多，故预测为“反例”。但是实际结果是正例，预测错误。
     - 若测试样本为“反例”
       - 训练集为 50 个正例，49 个反例，其中正例多，故预测为“正例”。但是实际结果是反例，预测错误。
     - 因此错误率为 100%。

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 ECOC 二元码并证明之。

   - 为了使海明距离最大化，应尽量让每一列（分类器）将类别均匀分开（即 2 个类标为 0，2 个类标为 1）。对于 4 个类别，将其分为两组（2 vs 2）只有以下 3 种不同的分法:
     - $\{1,2\}$ vs $\{3,4\} \rightarrow$ 列向量 $[0,0,1,1]^T$
     - $\{1,3\}$ vs $\{2,4\} \rightarrow$ 列向量 $[0,1,0,1]^T$
     - $\{1,4\}$ vs $\{2,3\} \rightarrow$ 列向量 $[0,1,1,0]^T$

     由于码长为 9，将这 3 种最优列向量各重复 3 次，从而构造出满秩且距离最大的矩阵。

     $$M = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$

   - 证明（计算海明距离）:
     - 任意两行之间的海明距离计算如下:
       - 行 1 与行 2：前 3 位相同，中 3 位不同，后 3 位不同 $\rightarrow$ 距离 = $0+3+3=6$
       - 行 1 与行 3：前 3 位不同，中 3 位相同，后 3 位不同 $\rightarrow$ 距离 = $3+0+3=6$
       - 行 1 与行 4：前 3 位不同，中 3 位不同，后 3 位相同 $\rightarrow$ 距离 = $3+3+0=6$
       - 行 2 与行 3：前 3 位不同，中 3 位不同，后 3 位相同 $\rightarrow$ 距离 = $3+3+0=6$
       - 其他任意组合的距离均为 6
     - 对于 4 个类别，单列最大总差异发生在 2 个 0 和 2 个 1 时，此时该列贡献的总配对距离为 $2 \times 2 = 4$。
     - 总共有 9 列，所有列的总配对距离之和上限为 $9 \times 4 = 36$。
     - 4 个类别共有 $C_4^2 = 6$ 对组合。
     - 平均距离上限 = $\frac{36}{6} = 6$。

- 构造的矩阵任意两行距离都达到了理论平均值的上限 6，因此它是理论最优的。

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。

样本序列	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, TN = 3, FP = 1, FN = 2
- 混淆矩阵：

	预测 +1	预测 −1
真实 +1	TP = 4	FN = 2
真实 −1	FP = 1	TN = 3

- 准确率 Accuracy = $\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} = 0.7$
- 精确率 Precision = $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = 0.8$
- 召回率 Recall = $\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \approx 0.6667$
- F1 score = $\frac{2\times \mathrm{Precision}\times \mathrm{Recall}}{\mathrm{Precision}+\mathrm{Recall}} \approx 0.7273$

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

- 对数据进行离散化处理
  - 年龄：
    - 20-30: 样本 1(否), 4(是), 8(是)
    - 30-40: 样本 3(否), 7(否)
    - 40+: 样本 2(是), 5(否), 6(是)
  - 收入：
    - 10-20: 样本 1(否), 3(否), 7(否)
    - 20-40: 样本 2(是), 5(否), 6(是), 8(是)
    - 40+: 样本 4(是)
- 计算根节点信息增益
  总样本 D (8 个)：4 正(是)，4 反(否)。 总熵 $H(D) = -0.5\log_2 0.5 - 0.5\log_2 0.5 = 1$。

1. 对于年龄
   - 20-30 (3 个: 2 正 1 反): $H = -\frac{2}{3}\log\frac{2}{3} - \frac{1}{3}\log\frac{1}{3} \approx 0.918$
   - 30-40 (2 个: 0 正 2 反): $H = 0$
   - 40+ (3 个: 2 正 1 反): $H \approx 0.918$
   - $\text{Gain}(\text{年龄}) = 1 - \left[\frac{3}{8} \times 0.918 + \frac{2}{8} \times 0 + \frac{3}{8}(0.918)\right] = 1 - 0.6885 = 0.3115$
2. 对于性别
   - 男 (6 个: 2 正 4 反): $H = -\frac{2}{6}\log\frac{2}{6} - \frac{4}{6}\log\frac{4}{6} \approx 0.918$
   - 女 (2 个: 2 正 0 反): $H = 0$
   - $\text{Gain}(\text{性别}) = 1 - \left[\frac{6}{8}(0.918) + 0\right] = 0.3115$
3. 对于收入
   - 10-20 (3 个: 0 正 3 反): $H = 0$
   - 20-40 (4 个: 3 正 1 反): $H = -\frac{3}{4}\log\frac{3}{4} - \frac{1}{4}\log\frac{1}{4} \approx 0.811$
   - 40+ (1 个: 1 正 0 反): $H = 0$
   - $\text{Gain}(\text{收入}) = 1 - \left[\frac{3}{8}(0) + \frac{4}{8}(0.811) + \frac{1}{8}(0)\right] = 1 - 0.4055 = 0.5945$

收入的增益最大，因此选择收入作为根节点。

- **构建分支**
  - **收入 = 10-20：样本 {1, 3, 7} 全为“否” → 叶节点：否**
  - **收入 = 40+：样本 {4} 全为“是” → 叶节点：是**
  - **收入 = 20-40：样本 {2, 5, 6, 8}。类别：{是, 否, 是, 是}。**
    - **子数据集 D' 熵 $H(D') = 0.811$。**
    - **剩余特征：年龄、性别。**

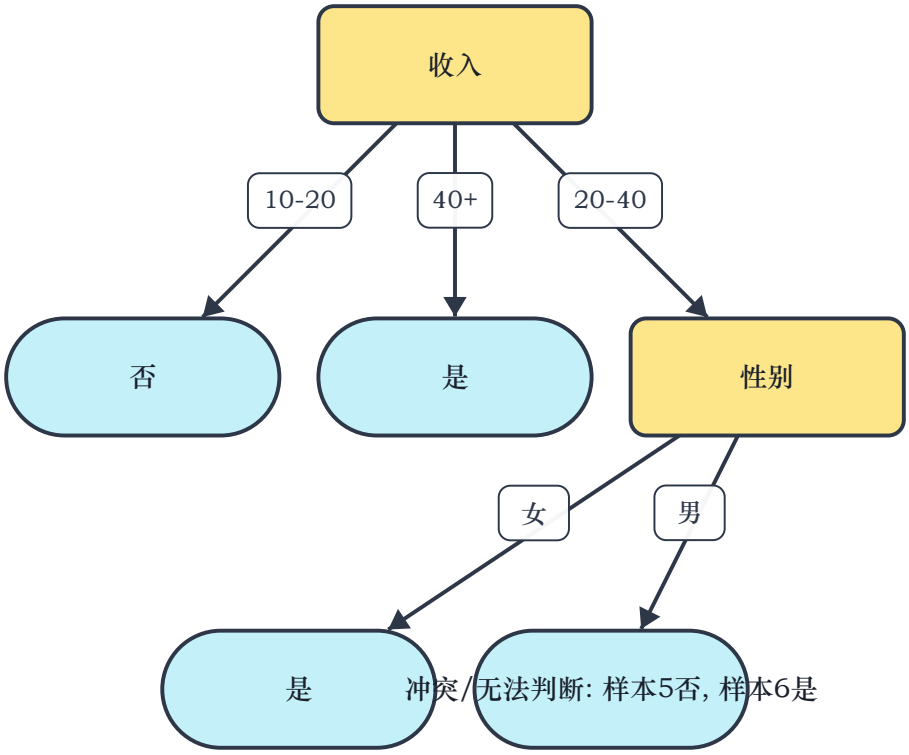
在 D' 上计算增益：

1. 对于年龄
   - 20-30 (样本 8: 是): $H = 0$
   - 40+ (样本 2, 5, 6: 是, 否, 是): $H = 0.918$
   - $\text{Gain}(\text{年龄}) = 0.811 - \left[\frac{1}{4}(0) + \frac{3}{4}(0.918)\right] = 0.811 - 0.6885 = 0.1225$
2. 对于性别
   - 女 (样本 2, 8: 是, 是): $H = 0$
   - 男 (样本 5, 6: 否, 是): $H = 1$
   - $\text{Gain}(\text{性别}) = 0.811 - \left[\frac{2}{4}(0) + \frac{2}{4}(1)\right] = 0.811 - 0.5 = 0.311$

性别的增益最大，因此选择性别作为第二个划分节点。

- **构建子分支（收入=20-40 下）**
  - **性别 = 女：样本 {2, 8} 全为“是” → 叶节点：是**
  - **性别 = 男：样本 {5, 6}。**
    - **样本 5：年龄 45 (40+)，收入 30 (20-40)，性别 男 → 否**
    - **样本 6：年龄 56 (40+)，收入 32 (20-40)，性别 男 → 是**

最终决策树结构：


收入
10-20
40+
20-40
否
是
性别
女
男
是
冲突/无法判断: 样本5否, 样本6是