第二次作业（6-10讲）

1.向量机相关的基本概念

VC维（Vapnik–Chervonenkis Dimension）

- **定义**：VC维是衡量一个模型（或函数集）“复杂度”的指标，即模型能**完全正确分类的最大样本数**。
- **意义**：VC维越大，模型的表示能力越强，但也更容易**过拟合**。
- **在SVM中**：SVM通过**最大化间隔（Margin）来控制模型复杂度，从而间接降低VC维，实现较好的泛化能力**。

核函数（Kernel Function）

- **定义**：核函数是一种数学工具，用于将原始输入空间中的样本隐式地映射到**高维特征空间**，使得线性不可分问题在高维空间中变得线性可分。
- **作用**：核函数让SVM能在**高维空间中构造超平面**，而无需显式计算映射。

支持向量（Support Vectors）

- **定义**：在SVM的训练结果中，**距离分离超平面最近的样本点**称为支持向量。
- **性质**：
  - 它们位于或接近分类间隔边界上；
  - 只有这些点对最终的超平面位置有影响；
  - 其他样本点（远离间隔边界的）对结果无影响。
- **意义**：支持向量是模型“记住”的关键信息点，体现了SVM的**稀疏性**。

SVM的最佳准则

- **核心思想**：在满足分类正确（或尽量正确）的前提下，使分类间隔最大。
- **线性不可分情形**：引入**松弛变量**与**惩罚因子 C**，形成“软间隔SVM”，在“间隔最大化”与“分类错误最小化”之间折中。

2.决策树学习相关问题

共有 10 个样本，正类 + 有 4 个，负类 − 有 6 个。

- 总体熵（父节点）：

$$H(P) = -\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10} \approx 0.97095$$

样本按属性分布：

- **属性 A**：

    - A=TA=TA=T：7 个样本，其中正 4，负 3。
    - A=FA=FA=F：3 个样本，其中正 0，负 3（纯负类）。
- **属性 B**：
    - B=TB=TB=T：4 个样本，其中正 3，负 1。
    - B=FB=FB=F：6 个样本，其中正 1，负 5。

（1）按信息增益（ID3 的准则）

信息增益定义：$IG(A) = H(P) - \sum_v \frac{|P_v|}{|P|} H(P_v)$。

属性 A 的信息增益

- $H(A=T) = -\frac{4}{7}\log_2\frac{4}{7} - \frac{3}{7}\log_2\frac{3}{7} \approx 0.86313$
- $H(A=F) = 0$（纯节点）
- 加权熵：$H_{split(A)} = \frac{7}{10}\cdot 0.86313 + \frac{3}{10}\cdot 0 = 0.68966$
- 信息增益：$IG(A) = 0.97095 - 0.68966 \approx 0.28129$

属性 B 的信息增益

- $H(B=T) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} \approx 0.81128$
- $H(B=F) = -\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6} \approx 0.63704$
- 加权熵：$H_{split(B)} = \frac{4}{10}\cdot 0.81128 + \frac{6}{10}\cdot 0.63704 \approx 0.71452$
- 信息增益：$IG(B) = 0.97095 - 0.71452 \approx 0.25643$

比较：$IG(A) \approx 0.2813 > IG(B) \approx 0.2564$。

结论（ID3）：应选择 **属性 A** 作为分裂特征。

（2）按基尼指数（CART 的准则）

CART 通常选择使分裂后 **加权基尼不纯度最小** 的特征。基尼不纯度：$G = 1 - \sum p_i^2$。

属性 A 的基尼

- $G(A=T) = 1 - (4/7)^2 - (3/7)^2 \approx 0.48980$
- $G(A=F) = 0$（纯节点）
- 加权基尼：$G_{split(A)} = \frac{7}{10}\cdot 0.48980 + \frac{3}{10}\cdot 0 = 0.34286$

属性 B 的基尼

- $G(B=T) = 1 - (3/4)^2 - (1/4)^2 = 0.375$
- $G(B=F) = 1 - (1/6)^2 - (5/6)^2 \approx 0.27778$
- 加权基尼：$G_{split(B)} = \frac{4}{10}\cdot 0.375 + \frac{6}{10}\cdot 0.27778 \approx 0.31667$

比较：$G_{split(A)} \approx 0.34286 > G_{split(B)} \approx 0.31667$。

结论（CART）：应选择 **属性 B** 作为分裂特征（因为基尼更小）。

3.集成学习的两大类型

迭代提升法（Boosting）

核心思想：
通过**串行（序列）训练**多个弱分类器，每一轮都**关注前一轮分错的样本**，逐步提升整体性能。

关键特征：

- **训练方式：** 顺序迭代，每一轮的学习器依赖前一轮结果。
- **样本权重调整：** 被错误分类的样本权重在下一轮中增加。
- **组合方式：** 各弱分类器加权投票或加权求和（如 AdaBoost 的权重与分类器准确率相关）。
- **代表算法：** AdaBoost、Gradient Boosting、XGBoost、LightGBM 等。

优点：

- 能显著降低偏差（bias）。
- 对难分类样本敏感，提升准确率。

缺点：

- 对噪声和异常值较敏感。
- 计算开销较大，训练难以并行化。

自助聚合法（Bagging）

核心思想：
通过**并行训练**多个学习器，每个学习器在**有放回的随机抽样数据集**上训练，然后**投票或平均结果**。

关键特征：

- **训练方式：** 并行，每个学习器独立。
- **数据子集生成：** 从原训练集通过 **Bootstrap抽样** 得到多个子集（样本有重复）。
- **组合方式：** 分类任务多数投票，回归任务取平均。
- **代表算法：** Bagging、随机森林（Random Forest）。

优点：

- 减少方差（variance），降低过拟合风险。
- 可以并行计算，效率高。

缺点：

- 不能显著降低偏差。
- 若基学习器过于简单，性能提升有限。

随机森林（Random Forest）与 Bagging 的区别：

比较项	Bagging	随机森林（Random Forest）
基学习器	任意模型（常用决策树）	决策树
数据随机性	仅样本随机（Bootstrap抽样）	样本随机 + 特征随机
特征选择	所有特征都参与划分	每次划分节点时**随机选择部分特征**参与
偏差与方差	主要降低方差	进一步降低方差，防止树间高度相关
性能	一般好	通常更稳健、准确

4.基本概念解释与各项指标计算

符号	英文全称	中文含义	说明
TP	True Positive	真阳性	实际为垃圾邮件（Spam），预测也为垃圾邮件
FN	False Negative	假阴性	实际为垃圾邮件，但预测为非垃圾邮件
FP	False Positive	假阳性	实际为非垃圾邮件，但预测为垃圾邮件
TN	True Negative	真阴性	实际为非垃圾邮件，预测也为非垃圾邮件

准确率（Accuracy）

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Accuracy = \frac{600 + 9000}{600 + 9000 + 100 + 300} = \frac{9600}{10000} = 0.96$$

错误率（Error Rate）

$$ErrorRate = 1 - Accuracy = \frac{FP + FN}{\text{总样本数}} = \frac{100 + 300}{10000} = 0.04$$

精确率（Precision，又称查准率）

$$Precision = \frac{TP}{TP + FP} = \frac{600}{600 + 100} = \frac{600}{700} \approx 0.8571$$

召回率（Recall，又称查全率）

$$Recall = \frac{TP}{TP + FN} = \frac{600}{600 + 300} = \frac{600}{900} = 0.6667$$

F1 度量（F1-score）

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$F1 = \frac{2 \times 0.8571 \times 0.6667}{0.8571 + 0.6667} \approx 0.75$$