第一次作业（1–5 讲）

1. 数据模型泛化能力、过拟合与欠拟合，以及降低风险的方法

（1）泛化能力（generalization）
泛化能力指模型在**未见过的、来自同一分布的数据**上的表现能力。一个泛化能力强的模型在训练集上学到的规律应能推广到新的样本，而不是仅记住训练样本的噪声或细节。

（2）过拟合（overfitting）

- 定义：模型在训练集上表现很好（训练误差很小），但在验证/测试集上表现差（验证误差/测试误差高）。通常是因为模型过于复杂或者训练数据不足或噪声过多，模型“记住”了训练数据的细节（噪声），从而不能泛化。
- 现象：训练误差持续下降，验证误差在一定点后开始上升。

（3）欠拟合（underfitting）

- 定义：模型对训练数据拟合不足，训练误差较高，说明模型容量或表达能力不足，不能捕捉数据中的真实规律。
- 现象：训练误差和验证误差都比较高，并且二者相差不大。

（4）降低过拟合的常用方法

- 增加训练数据（更多样本，数据增强）。
- 正则化（L1、L2、权重衰减）。
- 简化模型（减少参数 / 降低网络层数或宽度）。
- 交叉验证（例如 k-fold CV）用于模型选择与超参调优。
- 提前停止（early stopping）：在验证误差开始上升时停止训练。
- Dropout（对神经网络），Random Forest 等集成方法通过随机性减少过拟合。
- 使用更合适的特征选择或降维（PCA 等）。
- 集成方法（bagging、boosting）在某些情况下也能提高泛化。

（5）降低欠拟合的常用方法

- 增加模型复杂度（更深/更宽的网络，更高阶模型）。
- 增加特征或用更有表达力的特征（特征工程）。
- 降低正则化强度（减小正则化系数）。
- 更长时间训练或更合适的优化算法与学习率调度。
- 改进模型结构（引入非线性、交互项等）。

---

2. 三类问题的判别函数（绘界面与分类判断）

给定判别函数：

$$d_1(X) = x_1 + 2x_2 - 4,$$
$$d_2(X) = x_1 - 4x_2 + 4,$$
$$d_3(X) = -x_1 + 3$$

先写出三条判别曲线（即对应$d_i(X) = 0$的直线）：

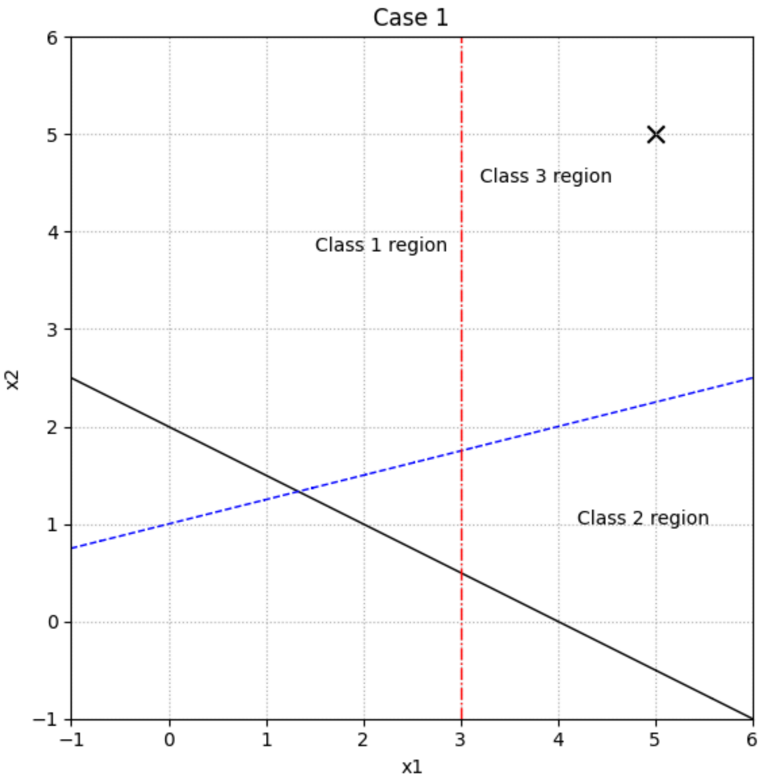
- $d_1(X) = 0$： $x_1 + 2x_2 - 4 = 0 \Rightarrow x_2 = \dfrac{4 - x_1}{2}$.（斜率 $-\frac{1}{2}$，截距 2）

- $d_2(X)=0$： $x_1-4x_2+4=0 \Rightarrow x_2=\dfrac{x_1+4}{4}$.（斜率 $\frac{1}{4}$，截距 1）
- $d_3(X)=0$： $-x_1+3=0 \Rightarrow x_1=3$.（垂直线）

下面分别按题中三种情形说明判别界面与所属区域并做给定点分类。

（1）多类情况 1

绘图为：



说明：通常理解为"每个$d_i(X)$表示类i的判别量，若$d_i(X)>0$则认为属于类i（同时要求其它$d_j(X)<0$）"。

区域描述（用不等式描述）：

- 类 1 的区域：$\{X \mid d_1(X)>0,\ d_2(X)<0,\ d_3(X)<0\}$。
- 类 2 的区域：$\{X \mid d_2(X)>0,\ d_1(X)<0,\ d_3(X)<0\}$。
- 类 3 的区域：$\{X \mid d_3(X)>0,\ d_1(X)<0,\ d_2(X)<0\}$。

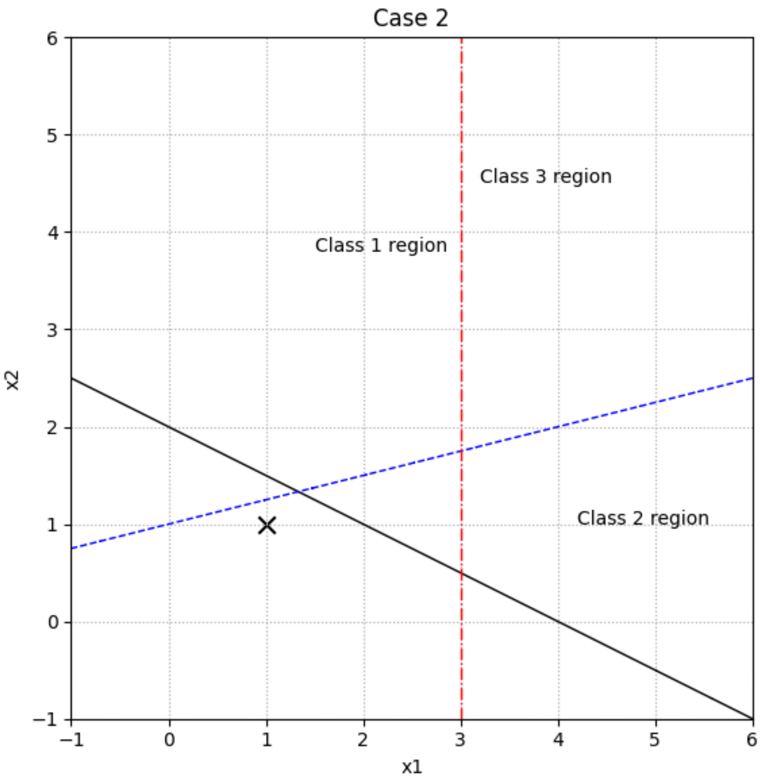
判断 $X=(5,5)^T$：
计算判别函数值：

$$d_1(5,5)=5+2\cdot 5-4=11$$
$$d_2(5,5)=5-4\cdot 5+4=5-20+4=-11$$
$$d_3(5,5)=-5+3=-2$$

结果$d_1>0,\ d_2<0,\ d_3<0$，因此按照情况 1 判为 **类 1**。

（2）多类情况 2（给定：$d_{12}=d_1$，$d_{13}=d_2$，$d_{23}=d_3$）

绘图为：



说明：多类情况 2 通常是"成对判别"（one-vs-one）：对于每一对i, j有一个判别函数$d_{ij}(X)$，若$d_{ij}(X) > 0$判为i，否则判为j。题中给的映射为：

$$d_{12}(X) = d_1(X) = x_1 + 2x_2 - 4$$
$$d_{13}(X) = d_2(X) = x_1 - 4x_2 + 4,$$
$$d_{23}(X) = d_3(X) = -x_1 + 3$$

判断$X = (1,1)^T$：

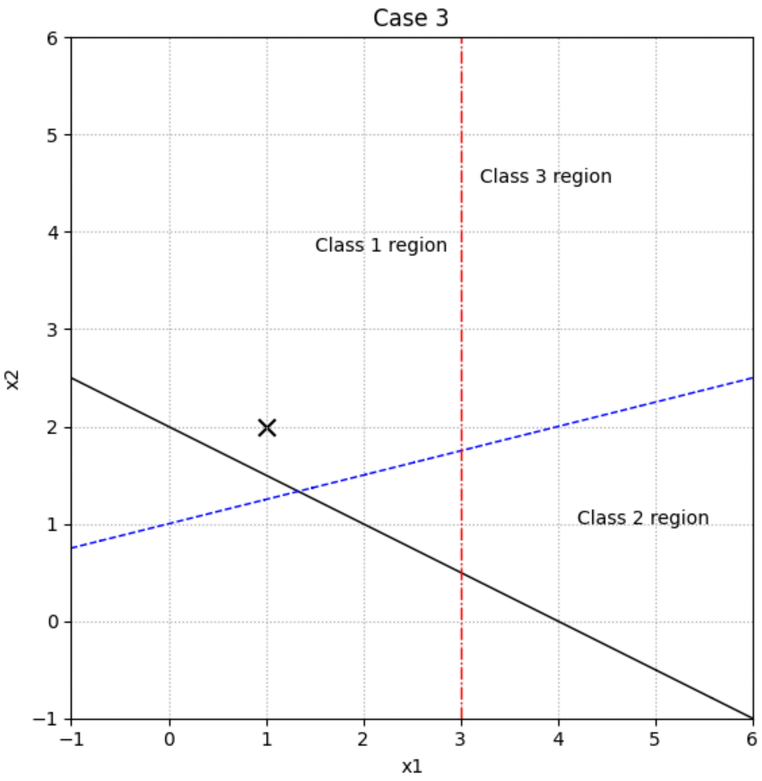
逐项计算如下：

$$d_{12}(1,1) = d_1(1,1) = 1 + 2 \cdot 1 - 4 = -1 \quad (\text{负，} 1 \text{ vs } 2 \text{ 判为 } 2),$$
$$d_{13}(1,1) = d_2(1,1) = 1 - 4 \cdot 1 + 4 = 1 \quad (\text{正，} 1 \text{ vs } 3 \text{ 判为 } 1),$$
$$d_{23}(1,1) = d_3(1,1) = -1 + 3 = 2 \quad (\text{正，} 2 \text{ vs } 3 \text{ 判为 } 2)$$

三个二分类投票结果：$\{2, 1, 2\}$，多数票为 **类 2**，因此$X = (1,1)^T$判为 **类 2**。

（3）多类情况 3（直接比大小）

绘图为:



规则：多类情况 3 常理解为“取最大判别量”：将模式分到使$d_i(X)$最大的类。
即 $\hat{\omega} = \arg\max_i d_i(X)$。

判断$X = (1,2)^T$：计算：

$$d_1(1,2) = 1 + 2 \cdot 2 - 4 = 1$$
$$d_2(1,2) = 1 - 4 \cdot 2 + 4 = 1 - 8 + 4 = -3$$
$$d_3(1,2) = -1 + 3 = 2$$

三者中最大的是$d_3 = 2$，因此$X = (1,2)^T$判为 **类 3**。

---

3. 贝叶斯判别（最小错误率与最小风险）

已知：

$P(\omega_1) = 0.2, P(\omega_2) = 0.8$，观测点$X$满足：

$$p(X \mid \omega_1) = 0.5, \quad p(X \mid \omega_2) = 0.2$$

（1）最小错误率（最小误分类概率）贝叶斯决策

比较后验（或等价比较$p(X \mid \omega_i)P(\omega_i)$）：

$$p(X \mid \omega_1)P(\omega_1) = 0.5 \times 0.2 = 0.10$$
$$p(X \mid \omega_2)P(\omega_2) = 0.2 \times 0.8 = 0.16$$

因为0.16>0.10，后验概率$P(\omega_2 \mid X)$更大，按最小错误率（即选择后验概率大的类）应判为 **正常细胞**ω_2。

（2）考虑损失函数（最小风险决策）

给定损失矩阵：

$$\begin{pmatrix} L_{11} & L_{12} \\ L_{12} & L_{22} \end{pmatrix} = \begin{pmatrix} 0 & 5 \\ 1 & 0 \end{pmatrix}$$

其中L_{ij}表示实际为j而判为i的损失（或按题意的行列约定）。

对于给定X，决策为α_1（判为ω_1）的条件风险：

$$R(\alpha_1 \mid X) = L_{11}P(\omega_1 \mid X) + L_{12}P(\omega_2 \mid X) = 0 \cdot P(\omega_1 \mid X) + 1 \cdot P(\omega_2 \mid X) = P(\omega_2 \mid X)$$

决策为α_2（判为ω_2）的条件风险：

$$R(\alpha_2 \mid X) = L_{21}P(\omega_1 \mid X) + L_{22}P(\omega_2 \mid X) = 5P(\omega_1 \mid X) + 0 \cdot P(\omega_2 \mid X) = 5P(\omega_1 \mid X)$$

比较大小：若$R(\alpha_1 \mid X) < R(\alpha_2 \mid X)$则选$\omega_1$，即当

$$P(\omega_2 \mid X) < 5P(\omega_1 \mid X)$$

用似然与先验代替后验比值：

$$\frac{P(\omega_1 \mid X)}{P(\omega_2 \mid X)} = \frac{p(X \mid \omega_1)P(\omega_1)}{p(X \mid \omega_2)P(\omega_2)} = \frac{0.5 \times 0.2}{0.2 \times 0.8} = \frac{0.10}{0.16} = 0.625$$

于是

$$\frac{P(\omega_1 \mid X)}{P(\omega_2 \mid X)} = 0.625 > \frac{1}{5} = 0.2$$

即 $\frac{P(\omega_1 \mid X)}{P(\omega_2 \mid X)} > \frac{1}{5}$ 价于 $P(\omega_2 \mid X) < 5P(\omega_1 \mid X)$，因此不等式成立，按最小风险准则应选择 ω_1（**异常细胞**）。

结论：在对错误代价不对称（将正常判为异常的损失=1，将异常判为正常的损失=5）的情况下，尽管后验上ω_2更大，但由于将异常漏判的代价很高，最小风险策略选择 **判为异常ω_1**。

4.朴素贝叶斯与条件概率计算

给定数据集（样本 1–10），观察后得到：共有 10 个样本，类“+”的样本数为 5（样本 2,5,6,9,10），类“-”的样本数为 5（样本 1,3,4,7,8）。

（1）估计条件概率（使用频率估计）

按定义：

$$P(A=1 \mid +) = \frac{n\{A=1 \text{ 且类为 } +\}}{n\{\text{类 } +\}}$$

计算结果：

- $P(A=1 \mid +) = \frac{3}{5} = 0.6$（类 + 中 A=1 出现在样本 2、5、10，共 3 个）
- $P(B=1 \mid +) = \frac{2}{5} = 0.4$（类 + 中 B=1 出现在样本 9、10，共 2 个）
- $P(C=1 \mid +) = \frac{4}{5} = 0.8$（类 + 中 C=1 出现在样本 2、5、6、10，共 4 个）
- $P(A=1 \mid -) = \frac{2}{5} = 0.4$（类 - 中 A=1 出现在样本 4、7，共 2 个）
- $P(B=1 \mid -) = \frac{2}{5} = 0.4$（类 - 中 B=1 出现在样本 3、7，共 2 个）
- $P(C=1 \mid -) = \frac{1}{5} = 0.2$（类 - 中 C=1 出现在样本 1，共 1 个）

（2）用朴素贝叶斯预测样本$(A=1, B=1, C=1)$

朴素贝叶斯假设在给定类条件下特征相互条件独立，于是后验概率与类先验乘以条件概率乘积成正比：

先验：

$$P(+)=\frac{5}{10}=0.5 \quad P(-)=0.5$$

似然（在类 + 下）：

$$P(A=1,B=1,C=1\mid +)=P(A=1\mid +)P(B=1\mid +)P(C=1\mid +)=0.6\times 0.4\times 0.8=0.192$$

后验（未归一化）：

$$P(+)\cdot P(\text{features}\mid +)=0.5\times 0.192=0.096$$

在类 − 下：

$$P(A=1,B=1,C=1\mid -)=0.4\times 0.4\times 0.2=0.032$$

后验（未归一化）：

$$0.5\times 0.032=0.016$$

比较未归一化后验：$0.096>0.016$，因此朴素贝叶斯预测**为类 +**。

（3）比较$P(A=1)$，$P(B=1)$，$P(A=1,B=1)$，陈述 A 与 B 的统计关系

先计算边缘概率（在全部 10 个样本上）：

- $P(A=1)=\frac{5}{10}=0.5$（A=1 在样本 2,4,5,7,10，共 5 个）
- $P(B=1)=\frac{4}{10}=0.4$（B=1 在样本 3,7,9,10，共 4 个）
- $P(A=1,B=1)=\frac{2}{10}=0.2$（A=1 且 B=1 在样本 7、10，共 2 个）

检验边缘独立性：若A与B边缘独立应满足

$$P(A=1,B=1)=P(A=1)P(B=1)=0.5\times 0.4=0.20$$

右侧等于观测到的0.2，因此 **在边缘上 A 与 B 是独立的**（用频率估计来看，两者的联合频率等于边缘概率乘积）。

（4）比较 $P(A=1\mid +)$，$P(B=1\mid +)$，$P(A=1,B=1\mid +)$，判断在给定类 + 的条件下 A、B 是否独立

已知：

- $P(A=1\mid +)=0.6$，$P(B=1\mid +)=0.4$
- 计算联合条件概率：在类 +（5 个样本）下，$A=1,B=1$出现在样本 10，仅 1 个，因此

$$P(A=1,B=1\mid +)=\frac{1}{5}=0.2$$

若A与B在给定类 + 时条件独立，应有

$$P(A=1,B=1\mid +)=P(A=1\mid +)\cdot P(B=1\mid +)=0.6\times 0.4=0.24$$

但观测到的联合概率$0.2\neq 0.24$，因此 **在给定类 + 的条件下，A 与 B 并非条件独立**（朴素贝叶斯的条件独立假设在该类下不完全成立）。