Assignment 4 详细解析

Support Vector Machine

第一题

1(a). 使用拉格朗日乘子法推导其对偶问题

假设我们有一个软边距支持向量机问题，其目标是最小化以下函数：

$$\min \frac{1}{2} w^\top w + C \sum_i \varepsilon_i$$

其中约束条件为：

$$\begin{gathered} \text{s.t. } y_i \left(w^\top x_i + b \right) \geq 1 - \varepsilon_i, \\ \varepsilon_i \geq 0 \end{gathered}$$

为了推导其对偶问题，我们首先引入拉格朗日乘子 α_i 和 μ_i，构造拉格朗日函数：

$$L(w, b, \varepsilon, \alpha, \mu) = \frac{1}{2} w^\top w + C \sum_i \varepsilon_i - \sum_i \alpha_i \left[y_i \left(w^\top x_i + b \right) - 1 + \varepsilon_i \right] - \sum_i \mu_i \varepsilon_i$$

对偶问题是关于拉格朗日函数的极大极小问题：

$$\max_{\alpha, \mu} \min_{w, b, \varepsilon} L(w, b, \varepsilon, \alpha, \mu)$$

其中，$\alpha_i \geq 0$ 和 $\mu_i \geq 0$。

1(b). 在鞍点进一步简化对偶问题并证明其等价于原始问题

在鞍点处，我们对 L 的偏导数求解，并使其等于零：

$$\frac{\partial L}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0$$

$$\frac{\partial L}{\partial b} = - \sum_i \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \varepsilon_i} = C - \alpha_i - \mu_i = 0$$

由此可得：

$$w = \sum_i \alpha_i y_i x_i$$

$$\sum_i \alpha_i y_i = 0$$

$$\alpha_i = C - \mu_i \leq C$$

将这些结果代入拉格朗日函数中，并消去 w 和 ε_i 后，我们得到对偶问题的形式：

$$\max \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^\top x_j$$

其中，约束条件为：

$$\text{s.t. } 0 \leq \alpha_i \leq C, \quad \sum_i \alpha_i y_i = 0$$

由于原始问题的强凸性，对偶问题等价于原始问题。

通过上述推导，我们证明了通过拉格朗日乘子法可以将软边距支持向量机问题的对偶问题简化为一个约束优化问题，其中对偶问题与原始问题等价。

2. 给定如下的 XOR 样本点，我们使用二次核训练一个 SVM，即我们的核函数是一个二次多项式核函数：$\kappa\left(x_i, x_j\right)=\left(x_i^T x_j\right)^d, d=2$。

2(a). 对应的映射函数 $\phi(x)$ 是什么？

推导多项式核对应的映射函数 $\phi(x)$

要推导得到多项式核函数 $K(x, z) = (x \cdot z)^d$ 对应的映射函数，可以通过以下步骤进行推导：

1. 首先，考虑输入空间中的两个样本点 $x = (x_1, x_2, ..., x_n)$ 和 $z = (z_1, z_2, ..., z_n)$。
2. 定义内积运算 $(x \cdot z) = x_1z_1 + x_2z_2 + ... + x_nz_n$。
3. 将内积运算的结果进行多项式展开，即 $(x \cdot z)^d = (x_1z_1 + x_2z_2 + ... + x_nz_n)^d$。
4. 根据二项式定理，展开 $(x \cdot z)^d$ 可以得到一系列项，每个项由输入样本点的各个维度的幂组成。
5. 将每个维度的幂项进行整理和组合，得到映射函数的形式。

对于 $n = 2, d = 2$ 的情况，推导过程如下：

$$(x \cdot z)^2 = (x_1z_1 + x_2z_2)^2 = x_1^2z_1^2 + 2x_1x_2z_1z_2 + x_2^2z_2^2$$

因此，根据展开的结果，得到映射函数为：

$$\phi(x) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$

这个映射函数将输入空间中的二维样本点映射到了三维特征空间，其中每个维度代表了对应的幂项。通过这样的映射，我们可以将原始输入空间中的数据转化为高维特征空间中的非线性特征表示，从而处理非线性问题。对于其他的 d 值，可以按照类似的步骤进行推导，得到对应的映射函数形式。

2(b). 使用下面的代码生成 XOR 数据，并根据 (a) 的答案，将数据映射到 $\phi(x)$ 看看是否可以线性可分。

```
import numpy as np
import matplotlib.pyplot as plt
#创建数据
X_xor = np.random.randn(40,2)
y_xor = np.logical_xor(X_xor[:,0]>0, X_xor[:,1]>0)
y_xor = np.where(y_xor, 1, -1)
#绘制散点图
plt.scatter(x=X_xor[y_xor==1,0]),  # 横坐标
```

```
        y=X_xor[y_xor==1,1]),  # 纵坐标
        color='g', marker='x', label='1')
plt.scatter(x=X_xor[y_xor==-1,0]),
        y=X_xor[y_xor==-1,1]),
        color='b', marker='o', label='-1')
plt.legend() #显示图例
plt.show()
```

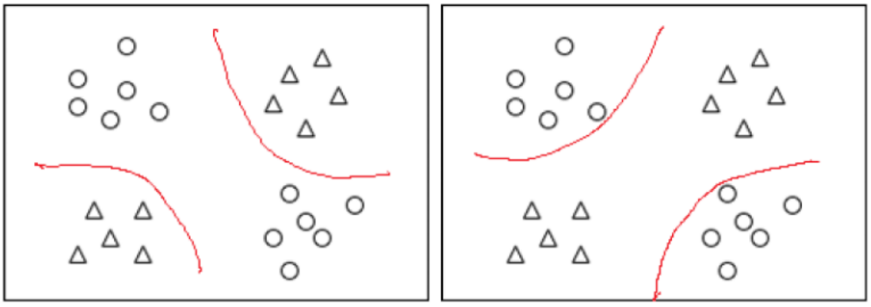
通过上述代码，我们生成了 XOR 数据并将其映射到三维空间。通过可视化可以看到，映射后的数据在三维空间中是**线性可分的**，从而验证了多项式核的有效性。

2(c) 是否可以通过硬边距得到合理的模型

答案是：**可以**。

根据上面的推导和可视化结果，在特征映射之后，数据在新的高维空间中是线性可分的。因此，我们可以在这个高维空间中使用硬边距支持向量机（即没有松弛变量的支持向量机）来训练一个模型，并得到一个合理的分类器。此分类器在原始的二维特征空间中的决策边界将会是非线性的。

如下图所示，在原始特征空间中，硬边距支持向量机得到的决策边界是非线性的：



第二题

Kernel PCA 的推导过程

Kernel PCA（核主成分分析）是一种将传统主成分分析（PCA）扩展到非线性特征空间的方法。其基本思想是通过核技巧将输入数据映射到高维特征空间，然后在该高维空间中进行PCA。

1. 数据预处理

首先，考虑输入数据集 $X = \{x_1, x_2, ..., x_n\}$，其中每个 $x_i \in \mathbb{R}^d$。我们希望将这些数据点映射到一个高维特征空间 $\mathcal{F}$，然后在该空间中进行PCA。

2. 核函数和特征映射

选择一个核函数 $k(x, y)$，它隐式地定义了一个从输入空间到特征空间的映射 $\phi : \mathbb{R}^d \to \mathcal{F}$。常见的核函数包括：

- 多项式核：$k(x, y) = (x \cdot y + c)^d$
- 高斯核（RBF核）：$k(x, y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$

核函数的选择将决定特征空间的性质。

3. 核矩阵的构造

构造核矩阵 $K \in \mathbb{R}^{n \times n}$，其中 $K_{ij} = k(x_i, x_j)$。核矩阵是对所有数据点对之间的核函数值的计算结果。

4. 中心化核矩阵

为了在特征空间中进行PCA，我们需要将核矩阵中心化。中心化后的核矩阵 K' 的计算方法如下：

$$K' = K - \mathbf{1}_n K - K\mathbf{1}_n + \mathbf{1}_n K \mathbf{1}_n$$

其中，$\mathbf{1}_n$ 是 $n \times n$ 的全1矩阵，且 $\mathbf{1}_n = \frac{1}{n}\mathbf{1}\mathbf{1}^T$。

5. 特征分解

对中心化后的核矩阵 K' 进行特征分解：

$$K' = V \Lambda V^T$$

其中，V 是特征向量组成的矩阵，Λ 是特征值组成的对角矩阵。

6. 投影到主成分

选择前 m 个最大特征值对应的特征向量，将原始数据点 x_i 投影到特征空间中的主成分上。投影后的坐标可以通过以下方式获得：

$$\tilde{x}_i = (\lambda_1^{-\frac{1}{2}} v_{i1}, \lambda_2^{-\frac{1}{2}} v_{i2}, ..., \lambda_m^{-\frac{1}{2}} v_{im})$$

其中，λ_j 是第 j 个特征值，v_{ij} 是第 i 个数据点在第 j 个特征向量上的分量。

7. 总结

Kernel PCA 的核心思想是通过核函数将数据映射到高维特征空间，并在该空间中进行主成分分析。其步骤包括：

1. 选择合适的核函数。
2. 计算核矩阵。
3. 中心化核矩阵。
4. 对中心化后的核矩阵进行特征分解。
5. 选择前 m 个主成分，并将数据投影到这些主成分上。

通过这些步骤，我们可以在高维特征空间中发现数据的主要特征，并实现对非线性数据结构的分析和降维。

第三题

(1) 对于 $x, y \in \mathbb{R}^N$，考虑函数 $\kappa(x, y) = \tanh(ax^T y + b)$，其中 a, b 是任意实数。试说明 $a \geq 0, b \geq 0$ 是 κ 为核函数的必要条件。

解答：

要证明 $a \geq 0, b \geq 0$ 是 $\kappa(x, y) = \tanh(ax^T y + b)$ 为核函数的必要条件，我们需要证明，对于任何输入数据集，该核函数生成的核矩阵是半正定的。

我们可以通过构造反例来证明，如果 $a < 0$ 或 $b < 0$ 时，核矩阵可能不是半正定的。

1. **考虑 $m = 1$ 的情况，令 $x = (x, 0, \cdots, 0)$，则核矩阵为 $K = [\tanh(ax^2 + b)]$。**

由于 $\tanh(y) \geq 0$ 且仅当 $y \geq 0$ 时，我们需要证明 $ax^2 + b \geq 0$ 是必要条件。

- **当 $a < 0$ 且 $b \leq 0$ 时，取 $x = 1$，则有 $ax^2 + b = a + b < 0$。**
  - 这意味着如果 $a < 0$ 且 $b \leq 0$，核矩阵的元素可以是负数，导致矩阵非半正定。
- **当 $a < 0$ 且 $b > 0$ 时，取 $x = \sqrt{-b/a}$，则有 $ax^2 + b = a(\sqrt{-b/a})^2 + b = -b + b = 0$。**

- 在这种情况下，虽然核矩阵的元素可以是零，但 $a < 0$ 会导致其他情况时元素为负，从而非半正定。

- **当 $a \geq 0$ 且 $b < 0$ 时，取 $x = 0$，则有 $ax^2 + b = b < 0$。**
  - 如果 $b < 0$，即使 $a \geq 0$，核矩阵也会有负元素，从而非半正定。

因此，我们得出结论，$a \geq 0$ 和 $b \geq 0$ 是 κ 为核函数的必要条件。

(2) 考虑 $\mathbb{R}^N$ 上的函数 $\kappa(x, y) = (x^T y + c)^d$，其中 c 是任意实数，d, N 是任意正整数。试分析函数 κ 何时是核函数，何时不是核函数，并说明理由。

解答：

要分析 $\kappa(x, y) = (x^T y + c)^d$ 是否为核函数，我们需要使用Mercer定理。该定理指出，核矩阵必须是半正定的。

1. **当 $c \geq 0$ 时，κ 是核函数。**
   - κ 是 d 个 $(x^T y + c)$ 的乘积。根据课本定理6.26，如果我们能证明 $x^T y + c$ 是核函数，即可证明 κ 是核函数。
   - 由于 $x^T y$ 是标准的线性核函数，显然是核函数。
   - 根据课本定理6.25，$x^T y + c$ 也是核函数，因为常数 c 不会改变核矩阵的半正定性。
2. **当 $c < 0$ 时，κ 不一定是核函数。**
   - 根据 Mercer 定理，一个函数 $K(x, z)$ 是一个合法的核函数，当且仅当对于任意的输入样本 $x_1, x_2, ..., x_n$，以及对应的 Gram 矩阵 K，矩阵 K 是半正定的。
   - 考虑 Gram 矩阵 K 的元素 K_{ij}，其中 $K_{ij} = K(x_i, x_j) = (x_i \cdot x_j + c)^d$。
   - 当 $c < 0$ 时，我们可以将常数项 c 表示为 $c = -|c|$，其中 $|c|$ 是 c 的绝对值。

   考虑一个简单的示例，假设存在两个样本 x_1 和 x_2，我们可以计算 K_{11} 和 K_{22}：

   $$K_{11} = (x_1 \cdot x_1 + c)^d = (x_1 \cdot x_1 - |c|)^d$$

   $$K_{22} = (x_2 \cdot x_2 + c)^d = (x_2 \cdot x_2 - |c|)^d$$

   - 由于 $x_1 \cdot x_1$ 和 $x_2 \cdot x_2$ 都是非负的，且 $c = -|c| < 0$，所以 K_{11} 和 K_{22} 会包含一个负数项 $(-|c|)^d$。
   - 当 d 是偶数时，负数的偶次幂仍然是正数，所以 K_{11} 和 K_{22} 仍然是非负的。
   - 然而，当 d 是奇数时，负数的奇次幂会变为负数。因此，在这种情况下，K_{11} 和 K_{22} 将包含负数项，导致 Gram 矩阵 K 不再是半正定的。

   根据 Mercer 定理的要求，核函数对应的 Gram 矩阵必须是半正定的，因此，当 $c < 0$ 且 d 是奇数时，多项式核函数 $K(x, z) = (x \cdot z + c)^d$ 不满足 Mercer 定理的条件，因此不是一个有效的核函数。

(3) 当上一个小问中的函数是核函数时，考虑 $d = 2$ 的情况，此时 κ 将 N 维数据映射到了什么空间中？具体的映射函数是什么？更一般的，对 d 不加限制时，κ 将 N 维数据映射到了什么空间中？

1. **考虑 $d = 2$ 的情况：**

   当 $d = 2$ 时，$\kappa(x, y) = (x^T y + c)^2$，对应的映射函数可以通过多项式展开来确定。

   令 $x = (x_1, x_2, \cdots, x_N)$，则：

   $$(x^T y + c)^2 = (x_1 y_1 + x_2 y_2 + \cdots + x_N y_N + c)^2$$

   展开后可以得到：

   $$(x_1 y_1 + x_2 y_2 + \cdots + x_N y_N + c)^2 = \sum_{i=1}^{N} x_i^2 y_i^2 + 2 \sum_{1 \leq i < j \leq N} x_i x_j y_i y_j + 2c \sum_{i=1}^{N} x_i y_i + c^2$$

这意味着 N 维数据 x 被映射到一个维度为 $\left(\frac{N^2+3N+2}{2}\right)$ 的空间中。具体的映射函数 $\phi(x)$ 可以表示为：

$$\phi(x) = \left(x_1^2, x_2^2, \cdots, x_N^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \cdots, \sqrt{2}x_{N-1}x_N, \sqrt{2}cx_1, \sqrt{2}cx_2, \cdots, \sqrt{2}cx_N, c\right)$$

其中，每一项代表的是 x 的不同次幂或交叉项。

2. **更一般的，当 d 为任意正整数时**：

   对于任意正整数 d，多项式核函数 $\kappa(x,y) = (x^Ty + c)^d$ 将 N 维数据 x 映射到维度为 $\left(\frac{N+d-1}{d-1}\right)$ 的空间中。

具体的映射函数 $\phi(x)$ 将包含所有形式为 $x_{i_1}x_{i_2}\cdots x_{i_d}$ 的项，其中 $1 \le i_1 \le i_2 \le \cdots \le i_d \le N$，再加上所有形式为 $cx_{i_1}x_{i_2}\cdots x_{i_{d-1}}$ 的项，依此类推，直至常数项 c^d。

具体来说：

$$\phi(x) = \left(x_1^d, x_2^d, \cdots, x_N^d, \sqrt{\frac{d!}{k_1!k_2!\cdots k_N!}}x_1^{k_1}x_2^{k_2}\cdots x_N^{k_N}, \cdots, \sqrt{\frac{d!}{k_1!k_2!\cdots k_N!}}c^{d-\sum k_i}x_1^{k_1}x_2^{k_2}\cdots x_N^{k_N}, \cdots, c^d\right)$$

其中，$k_1 + k_2 + \cdots + k_N \le d$，每一项系数是组合数 $\sqrt{\frac{d!}{k_1!k_2!\cdots k_N!}}$。

总结

1. 当 $d = 2$ 时，κ 将 N 维数据映射到维度为 $\left(\frac{N^2+3N+2}{2}\right)$ 的空间中。
2. 更一般的，当 d 为任意正整数时，κ 将 N 维数据映射到维度为 $\left(\frac{N+d-1}{d-1}\right)$ 的空间中。

第四题

Kernel LDA 的推导过程

线性判别分析（LDA）是一种经典的降维和分类方法，通过最大化类间散布矩阵与类内散布矩阵的比值来找到最优的投影方向。Kernel LDA 是将 LDA 扩展到非线性空间的版本。其基本思想是通过核技巧将输入数据映射到高维特征空间，然后在该高维空间中进行 LDA。

1. 数据预处理

考虑输入数据集 $X = \{x_1, x_2, ..., x_n\}$，其中每个 $x_i \in \mathbb{R}^d$，并且有 k 个类别。定义每个类别的样本集合为 X_i，其中 $i = 1, 2, ..., k$。

2. 核函数和特征映射

选择一个核函数 $k(x, y)$，它隐式地定义了一个从输入空间到特征空间的映射 $\phi : \mathbb{R}^d \to \mathcal{F}$。常见的核函数包括：

- 多项式核：$k(x, y) = (x \cdot y + c)^d$
- 高斯核（RBF 核）：$k(x, y) = \exp(-\frac{\|x-y\|^2}{2\sigma^2})$

核函数的选择将决定特征空间的性质。

3. 类内散布矩阵和类间散布矩阵

在特征空间中定义类内散布矩阵 S_W 和类间散布矩阵 S_B：

- 类内散布矩阵：

$$S_W = \sum_{i=1}^{k} \sum_{x \in X_i} (\phi(x) - m_i)(\phi(x) - m_i)^T$$

其中，m_i 是第 i 类的均值向量：

$$m_i = \frac{1}{|X_i|} \sum_{x \in X_i} \phi(x)$$

- 类间散布矩阵：

$$S_B = \sum_{i=1}^{k} |X_i| (m_i - m)(m_i - m)^T$$

其中，m 是所有样本的均值向量：

$$m = \frac{1}{n} \sum_{i=1}^{n} \phi(x_i)$$

4. 在特征空间中求解 LDA

为了在特征空间中进行 LDA，我们需要最大化类间散布矩阵 S_B 和类内散布矩阵 S_W 的比值，即：

$$\arg\max_w \frac{w^T S_B w}{w^T S_W w}$$

在特征空间中，我们需要找到一个投影方向 w，使得上述比值最大。由于 w 是特征空间中的向量，难以直接求解，因此我们通过核函数将问题转化。

5. 使用核函数的 LDA

我们引入核矩阵 K，其中 $K_{ij} = k(x_i, x_j)$。利用核矩阵，我们可以将特征空间中的操作转化为输入空间中的操作。

定义中心化核矩阵 K'：

$$K' = K - \mathbf{1}_n K - K \mathbf{1}_n + \mathbf{1}_n K \mathbf{1}_n$$

其中，$\mathbf{1}_n$ 是 $n \times n$ 的全1矩阵。

使用中心化核矩阵，我们可以计算类内散布矩阵和类间散布矩阵在特征空间中的对应形式。

- 类内散布矩阵：

$$S_W = \sum_{i=1}^{k} (K'_i - M_i)(K'_i - M_i)^T$$

其中，K'_i 是第 i 类样本的中心化核矩阵，M_i 是第 i 类样本的中心化均值向量。

- 类间散布矩阵：

$$S_B = \sum_{i=1}^{k} |X_i| (M_i - M)(M_i - M)^T$$

其中，M 是所有样本的中心化均值向量。

6. 求解广义特征值问题

在特征空间中进行 LDA 归结为求解广义特征值问题：

$$S_B w = \lambda S_W w$$

由于我们通过核函数进行了映射，我们在输入空间中实际需要求解的是：

$$\bar{} \qquad \bar{}$$

$$(K'\alpha)^T S_B (K'\alpha) = \lambda (K'\alpha)^T S_W (K'\alpha)$$

通过求解上述广义特征值问题，我们可以得到特征空间中的投影向量 α，并利用核函数将其应用于原始数据进行分类。

Kernel LDA 的核心思想是通过核函数将数据映射到高维特征空间，并在该空间中进行线性判别分析。其步骤包括：

1. 选择合适的核函数。
2. 计算核矩阵并进行中心化处理。
3. 在特征空间中计算类内散布矩阵和类间散布矩阵。
4. 求解广义特征值问题，找到最佳投影方向。

通过这些步骤，我们可以在高维特征空间中实现非线性分类和降维。

Naive Bayes

第一题

(1) 计算每个变量在每个标签中出现的条件概率

我们使用给定的数据集来计算条件概率。使用拉普拉斯平滑来避免零概率。

给定的数据集：

A	B	C	D	y
2	4	10	3	$+1$
3	1	4	2	$+1$
0	2	0	5	-1
2	0	4	0	$+1$
1	6	6	0	-1
0	2	1	7	-1
3	0	0	8	$+1$
6	1	2	7	-1

拉普拉斯平滑后的条件概率

- 计算每个变量在每个标签中出现的次数之和。
- 使用拉普拉斯平滑计算条件概率：

$$p_{A,+1} = \frac{10+1}{46+4} = \frac{11}{50}$$

$$p_{B,+1} = \frac{5+1}{46+4} = \frac{6}{50}$$

$$p_{C,+1} = \frac{18+1}{46+4} = \frac{19}{50}$$

$$p_{D,+1} = \frac{13+1}{46+4} = \frac{14}{50}$$

$$p_{A,-1} = \frac{7+1}{46+4} = \frac{8}{50}$$

$$p_{B,-1} = \frac{11+1}{46+4} = \frac{12}{50}$$

$$p_{C,-1} = \frac{9+1}{46+4} = \frac{10}{50}$$

$$p_{D,-1} = \frac{19+1}{46+4} = \frac{20}{50}$$

(2) 给定新的样本 $(A=3, B=2, C=1, D=2)$，预测其标签。

使用计算出的条件概率来预测新的样本的标签。

我们需要计算：

$P(y=+1 \mid A=3, B=2, C=1, D=2) \propto P(A=3 \mid y=+1)P(B=2 \mid y=+1)P(C=1 \mid y=+1)P(D=2 \mid y=$

$P(y=-1 \mid A=3, B=2, C=1, D=2) \propto P(A=3 \mid y=-1)P(B=2 \mid y=-1)P(C=1 \mid y=-1)P(D=2 \mid y=$

使用拉普拉斯平滑计算每个条件概率：

$$P(A=3 \mid y=+1) = \left(\frac{11}{50}\right)^3$$

$$P(B=2 \mid y=+1) = \left(\frac{6}{50}\right)^2$$

$$P(C=1 \mid y=+1) = \left(\frac{19}{50}\right)$$

$$P(D=2 \mid y=+1) = \left(\frac{14}{50}\right)^2$$

$$P(y=+1) = \frac{4}{8} = \frac{1}{2}$$

所以：

$$P(y=+1 \mid A=3, B=2, C=1, D=2) \propto \left(\frac{11}{50}\right)^3 \left(\frac{6}{50}\right)^2 \left(\frac{19}{50}\right) \left(\frac{14}{50}\right)^2 \times \frac{1}{2}$$

计算：

$$P(A=3 \mid y=-1) = \left(\frac{8}{50}\right)^3$$

$$P(B=2 \mid y=-1) = \left(\frac{12}{50}\right)^2$$

$$P(C=1 \mid y=-1) = \left(\frac{10}{50}\right)$$

$$P(D=2 \mid y=-1) = \left(\frac{20}{50}\right)^2$$

$$P(y=-1) = \frac{4}{8} = \frac{1}{2}$$

所以：

$$P(y=-1 \mid A=3, B=2, C=1, D=2) \propto \left(\frac{8}{50}\right)^3 \left(\frac{12}{50}\right)^2 \left(\frac{10}{50}\right) \left(\frac{20}{50}\right)^2 \times \frac{1}{2}$$

计算：

$$\frac{8^3 \times 12^2 \times 10 \times 20^2 \times 1}{50^8}$$

比较两个概率的大小：

$$P(y=-1 \mid A=3, B=2, C=1, D=2) > P(y=+1 \mid A=3, B=2, C=1, D=2)$$

因此，该样本的标签应该是 -1。

第二题

(1) 最小化分类错误率的贝叶斯最优分类器

贝叶斯最优分类器可以通过最大化后验概率来实现：

$$h^*(x) = \arg\max_y \Pr(y \mid x)$$

(2) 给定样本 x 服从类内相同协方差矩阵的正态分布

假设第 k 类中的样本是从正态分布 $\mathcal{N}(\mu_k, \Sigma)$ 中独立同分布抽取的，其中 $k = 1, 2, \ldots, K$，所有类别共享相同的协方差矩阵。令 m_k 表示第 k 类中的样本数量，先验概率 $P(y=k) = \pi_k$。

对于 $x \in \mathbb{R}^d$，若 $x \sim \mathcal{N}(\mu, \Sigma)$，则其概率密度函数为：

$$p(x) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)$$

根据贝叶斯定理，我们有：

$$\Pr(y=k \mid x) \propto \Pr(y=k) \Pr(x \mid y=k)$$

对于正态分布 $\mathcal{N}(\mu_k, \Sigma)$，我们有：

$$\Pr(x \mid y=k) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^\top \Sigma^{-1}(x-\mu_k)\right)$$

因此，贝叶斯最优分类器可以表示为：

$$h^*(x) = \arg\max_k \left[\ln \pi_k - \frac{1}{2}(x-\mu_k)^\top \Sigma^{-1}(x-\mu_k)\right]$$

(3) 二分类问题

在二分类问题中，若每个类中的样本是从共享相同协方差矩阵的正态分布中独立同分布抽取的，且两个类具有相等的先验概率 $\pi_0 = \pi_1$，则LDA给出贝叶斯最优分类器。

提示：LDA 的最优解为：

$$w = S_w^{-1}(\mu_0 - \mu_1)$$

其中，S_w 是类内散布矩阵，$S_w = \Sigma_0 + \Sigma_1$（$\Sigma_i$ 是第 i 类的协方差矩阵）。

贝叶斯最优决策边界是：

$$g(x) = x^\top \Sigma^{-1}(\mu_0 - \mu_1) - \frac{1}{2}(\mu_0^\top \Sigma^{-1}\mu_0 - \mu_1^\top \Sigma^{-1}\mu_1) + \ln\left(\frac{\pi_0}{\pi_1}\right)$$

因为 $\pi_0 = \pi_1$，所以决策边界可以简化为：

$$g(x) = x^\top \Sigma^{-1}(\mu_0 - \mu_1) - \frac{1}{2}(\mu_0^\top \Sigma^{-1}\mu_0 - \mu_1^\top \Sigma^{-1}\mu_1)$$

当两个类别共享相同的协方差矩阵时，LDA 的最优解是：

$$w = S_w^{-1}(\mu_0 - \mu_1) = (2\Sigma)^{-1}(\mu_0 - \mu_1) = \frac{1}{2}\Sigma^{-1}(\mu_0 - \mu_1)$$

投影类中心的中点是：

$$c = \frac{1}{2}(\mu_0 + \mu_1)^\top w = \frac{1}{4}(\mu_0 + \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1)$$

决策边界是：

$$f(x) = x^\top w - c = \frac{1}{2}x^\top \Sigma^{-1}(\mu_0 - \mu_1) - \frac{1}{4}(\mu_0 + \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1)$$

这等价于：

$$f(x) = x^\top \Sigma^{-1}(\mu_0 - \mu_1) - \frac{1}{2}(\mu_0 + \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1)$$

由于 $\pi_0 = \pi_1$，所以 $f(x)$ 等价于 $g(x)$，因此 LDA 给出了贝叶斯最优分类器。

第三题

(1) 应用最大似然估计 (MLE) 推导优化问题

假设样本点 $X_i \sim D$。标签 y_i 是确定性函数 $f(X_i)$ 加上随机噪声的和：$y_i = f(X_i) + \varepsilon_i$，其中 $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$。我们假设每个样本点的噪声方差 σ_i^2 已知。

我们需要应用最大似然估计 (MLE) 来推导分布参数 f 的最大似然估计。

似然函数为：

$$L = \prod_i \frac{1}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{1}{2\sigma_i^2}(y_i - f(X_i))^2\right)$$

对数似然函数为：

$$\begin{aligned}\log L &= \sum_i \log\left(\frac{1}{\sqrt{2\pi\sigma_i}} \exp\left(-\frac{1}{2\sigma_i^2}(y_i - f(X_i))^2\right)\right) \\ &= \sum_i \left(-\frac{1}{2}\log(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2}(y_i - f(X_i))^2\right)\end{aligned}$$

我们要最大化对数似然函数，因此目标是最小化负对数似然函数：

$$\mathcal{L} = -\log L \propto \sum_i \frac{1}{2\sigma_i^2}(y_i - f(X_i))^2$$

因此，损失函数为：

$$\mathcal{L} = \frac{1}{2} \sum_i \frac{1}{\sigma_i^2} (y_i - f(X_i))^2$$

(2) 线性回归中的等效优化问题

我们决定进行线性回归，因此我们将 $f(X_i)$ 参数化为 $f(X_i) = w \cdot X_i$，其中 w 是 p 维权重向量。

将损失函数用矩阵形式表示，我们定义设计矩阵 X，其中每一行是一个样本点 X_i，标签向量 y，以及对角矩阵 $\Omega = \text{diag}(\sigma_1^{-2}, \sigma_2^{-2}, \ldots, \sigma_n^{-2})$。

损失函数变为：

$$\mathcal{L} = \frac{1}{2} \sum_i \frac{1}{\sigma_i^2} (y_i - w \cdot X_i)^2$$

用矩阵形式表示：

$$\mathcal{L} = \frac{1}{2} (y - Xw)^\top \Omega (y - Xw)$$

(3) 优化问题的解

我们通过求解线性方程组来最小化损失函数 $\mathcal{L}$：

$$\frac{\partial \mathcal{L}}{\partial w} = 0$$

$$\frac{\partial}{\partial w} \left(\frac{1}{2} (y - Xw)^\top \Omega (y - Xw) \right) = 0$$

$$\Rightarrow X^\top \Omega (y - Xw) = 0$$

$$\Rightarrow X^\top \Omega y = X^\top \Omega X w$$

解得：

$$w = (X^\top \Omega X)^{-1} X^\top \Omega y$$

1. 应用 MLE 推导优化问题，得到损失函数 $\mathcal{L} = \frac{1}{2} \sum_i \frac{1}{\sigma_i^2} (y_i - f(X_i))^2$。
2. 对于线性回归，损失函数可以用矩阵形式表示为 $\mathcal{L} = \frac{1}{2} (y - Xw)^\top \Omega (y - Xw)$。
3. 最小化损失函数的解为 $w = (X^\top \Omega X)^{-1} X^\top \Omega y$。