姓名：杜兴豪
学号：201300096

一. (20 points) **没有免费的午餐定理**

1. 根据教材 1.4 节“没有免费的午餐”定理, 所有学习算法的期望性能都和随机胡猜一样, 是否还有必要继续进行研究机器学习算法?
2. 教材 1.4 节在论述“没有免费的午餐”定理时, 默认使用了“分类错误率”作为性能度量来对分类器进行评估. 若换用其他性能度量 ℓ, 则教材中式（1.1）将改为

$$E_{ote}(\mathfrak{L}_a|X,f)=\sum_h\cdot\sum_{x\in\mathcal{X}-X}P(\boldsymbol{x})\ell(h(\boldsymbol{x}),f(\boldsymbol{x}))P(h|\mathcal{X},\mathfrak{L}_a) \quad (1)$$

试证明“没有免费的午餐定理”仍成立.

解:

1. 有必要。

算法对所有问题的期望性能不等价于算法的实际作用。算法是被设计出来解决一些特定问题的。

虽然普适的算法并不存在，但是我们仍然可以设计出针对特定问题有高效解决办法的算法，这对生产生活会产生很好的实际意义。

2. 证明:

对题中所给 $\ell(h(\boldsymbol{x}),f(\boldsymbol{x}))$, 当 f 均匀分布在所有可能的情况中时, 其值应只和 f 的取值空间和“随机胡猜”获得良好性能的概率有关, 因此记

$$\sum_f\ell(h(\boldsymbol{x}),f(\boldsymbol{x}))=L(\mathcal{X})$$

则对所有的 f 按照均匀分布求和:

$$\begin{aligned}\sum_f E_{ote}(\mathfrak{L}_a|X,f)&=\sum_f\sum_h\sum_{x\in\mathcal{X}-X}P(\boldsymbol{x})\ell(h(\boldsymbol{x}),f(\boldsymbol{x}))P(h|\mathcal{X},\mathfrak{L}_a)\\&=\sum_{x\in\mathcal{X}-X}P(\boldsymbol{x})\sum_h P(h|\mathcal{X},\mathfrak{L}_a)\sum_f\ell(h(\boldsymbol{x}),f(\boldsymbol{x}))\\&=\sum_{x\in\mathcal{X}-X}P(\boldsymbol{x})\sum_h P(h|\mathcal{X},\mathfrak{L}_a)L(\mathcal{X})\end{aligned}$$

$$\begin{aligned}\sum_f E_{ote}(\mathfrak{L}_a|X,f) &= L(\mathcal{X})\sum_{x\in\mathcal{X}-X} P(\boldsymbol{x})\sum_h P(h|\mathcal{X},\mathfrak{L}_a)\\ &= L(\mathcal{X})\sum_{x\in\mathcal{X}-X} P(\boldsymbol{x})\cdot 1\end{aligned}$$

可以看出总和与选取的学习算法没有关系。

二. (15 points) **线性回归**

给定包含 m 个样例的数据集 $\boldsymbol{D}=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_m,y_m)\}$, 其中 $\boldsymbol{x}_i=(x_{i1};x_{i2};\cdots;x_{id})\in\mathbb{R}^d$, $y_i\in\mathbb{R}$ 为 $\boldsymbol{x}_i$ 的实数标记. 针对数据集 $\boldsymbol{D}$ 中的 m 个示例, 教材 3.2 节所介绍的 "线性回归" 模型要求该线性模型的预测结果和其对应的标记之间的误差之和最小:

$$\begin{aligned}(\boldsymbol{w}^*,b^*) &= \frac{1}{2}\arg\min_{(\boldsymbol{w},b)}\sum_{i=1}^{m}\left(f\left(\boldsymbol{x}_i\right)-y_i\right)^2\\ &= \frac{1}{2}\arg\min_{(\boldsymbol{w},b)}\sum_{i=1}^{m}\left(y_i-\left(\boldsymbol{w}^\top\boldsymbol{x}_i+b\right)\right)^2 .\end{aligned}\tag{2}$$

即寻找一组权重 $(\boldsymbol{w},b)$, 使其对 $\boldsymbol{D}$ 中示例预测的整体误差最小.[1] 定义 $y=[y_1;\ldots,y_m]\in\mathbb{R}^m$, 且 $\boldsymbol{X}=\left[\boldsymbol{x}_1^\top;\boldsymbol{x}_2^\top;\cdots;\boldsymbol{x}_m^\top\right]\in\mathbb{R}^{m\times d}$, 请将线性回归的优化过程使用矩阵进行表示.

解:

将 w,b 吸收进 $\hat{w}$, 有

$$\hat{\boldsymbol{w}}=(\boldsymbol{w};b)\in\mathbb{R}^{d+1}\text{ and }\hat{\boldsymbol{X}}=(\boldsymbol{X};1)\in\mathbb{R}^{m\times(d+1)}$$

原优化目标可改写为:

$$\frac{1}{2}E(\hat{\boldsymbol{w}})=(\boldsymbol{w}^*,b^*)=\frac{1}{2}\arg\min(y-\hat{\boldsymbol{X}}\hat{\boldsymbol{w}})^\top(y-\hat{\boldsymbol{X}}\hat{\boldsymbol{w}})$$

对 $\hat{w}$ 求偏导有:

$$\frac{\partial E(\hat{\boldsymbol{w}})}{\partial\hat{\boldsymbol{w}}}=2\hat{\boldsymbol{X}}^\top(\hat{\boldsymbol{X}}\hat{\boldsymbol{w}}-y)=0\quad\Rightarrow\hat{\boldsymbol{X}}^\top\hat{\boldsymbol{X}}\hat{\boldsymbol{w}}^*=\hat{\boldsymbol{X}}^\top y$$

[1]公式 2中系数 $\frac{1}{2}$ 是为了化简后续推导. 有时也会乘上 $\frac{1}{m}$ 以计算均方误差 (Mean Square Error), 由于平均误差和误差和在优化过程中只相差一个常数, 不影响优化结果, 因此在后续讨论中省略这一系数.

当 $\hat{\boldsymbol{X}}^\top X$ 为满秩矩阵或正定矩阵时，可以进行进一步化简得到 $\hat{w}$ 的闭式解：

$$\hat{\boldsymbol{w}}^* = (\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}})^{-1} \hat{\boldsymbol{X}}^\top y$$

得到原来问题的解为：

$$f(\hat{x}_i) = \hat{x}_i^\top (\hat{\boldsymbol{X}}^\top \hat{\boldsymbol{X}})^{-1} \hat{\boldsymbol{X}}^\top y$$

三. (25 points) **正则化**

在实际问题中, 我们常常会遇到示例相对较少, 而特征很多的场景. 在这类情况中如果直接求解线性回归模型, 较少的示例无法获得唯一的模型参数, 会具有多个模型能够" 完美" 拟合训练集中的所有样例, 实现插值 (interpolation). 此外, 模型很容易过拟合. 为缓解这些问题, 常在线性回归的闭式解中引入正则化项 $\Omega(\boldsymbol{w})$, 通常形式如下:

$$\boldsymbol{w}^*_{\textbf{Ridge}}, b^*_{\textbf{Ridge}} = \arg\min_{\boldsymbol{w},b} \frac{1}{2} \|\boldsymbol{X}\boldsymbol{w} + \mathbf{1}b - \mathbf{y}\|_2^2 + \lambda\Omega(\boldsymbol{w}) \ . \tag{3}$$

其中, $\lambda > 0$ 为正则化参数. 正则化表示了对模型的一种偏好, 例如 $\Omega(\boldsymbol{w})$ 一般对模型的复杂度进行约束, 因此相当于从多个在训练集上表现同等预测结果的模型中选出模型复杂度最低的一个.

考虑岭回归 (ridge regression) 问题, 即设置公式(3)中正则项 $\Omega(\boldsymbol{w}) = \|\boldsymbol{w}\|_2^2$. 本题中将对岭回归的闭式解以及正则化的影响进行探讨.

1. 请给出岭回归的最优解 $\boldsymbol{w}^*_{\textbf{Ridge}}$ 和 $b^*_{\textbf{Ridge}}$ 的闭式解表达式, 并使用矩阵形式表示, 分析其最优解和原始线性回归最优解 $\boldsymbol{w}^*_{\textbf{LS}}$ 和 $b^*_{\textbf{LS}}$ 的区别;
2. 请证明对于任何矩阵 $\boldsymbol{X}$, 下式均成立

   $$\left(\boldsymbol{X}\boldsymbol{X}^\top + \lambda\boldsymbol{I}_m\right)^{-1}\boldsymbol{X} = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}_d\right)^{-1} \ . \tag{4}$$

   请思考, 上述的结论是否能够帮助岭回归的计算, 在何种情况下能够带来帮助?
3. 针对波士顿房价预测数据 (`boston`), 编程实现原始线性回归模型和岭回归模型, 基于闭式解在训练集上构建模型, 计算测试集上的均方误差 (Mean Square Error, MSE). 请参考`LinearRegression.py`进行模型构造.

```python
from sklearn.datasets import load_boston
from sklearn.model_selection import train_test_split
import numpy as np

X, y = load_boston(return_X_y = True)
trainx, testx, trainy, testy = train_test_split(X, y, test_size = 0.33, random_state = 42)

# linear regression
def linReg(X_train:np.ndarray, y_train:np.ndarray) -> np.ndarray:
    pass

def linRegMSE(X_train:np.ndarray, y_train:np.ndarray, X_test:np.ndarray, y_test:np.ndarray) -> float:
    pass
reportLinRegMSE= lambda : linRegMSE(trainx,trainy,testx,testy)

# ridge regression
def ridgeReg(X_train:np.ndarray, y_train:np.ndarray, lmbd:float) -> np.ndarray:
    pass

def ridgeRegMSE(X_train:np.ndarray, y_train:np.ndarray, X_test:np.ndarray, y_test:np.ndarray, lmbd:float) -> float:
    pass
reportRidgeRegMSE= lambda lmbd : ridgeRegMSE(trainx,trainy,testx,testy,lmbd)
```

(a) 对于线性回归模型, 请直接计算测试集上的 MSE;

(b) 对于岭回归问题, 请考察不同正则项权重 λ 的取值范围, 并观察训练集 MSE、测试集 MSE 和 λ 的取值的关系, 总结变化的规律;

除示例代码中使用到的 sklearn 库函数外, 不能使用其他的 sklearn 函数, 需要基于 numpy 实现线性回归模型和 MSE 的计算.

解:

1. 令:

$$f(\boldsymbol{w}, b) = \boldsymbol{w}^*_{\textbf{Ridge}}, b^*_{\textbf{Ridge}} = \arg\min_{w,b} \frac{1}{2}\|\boldsymbol{X}\boldsymbol{w} + \mathbf{1}b - \mathbf{y}\|_2^2 + \lambda\Omega(\boldsymbol{w})$$

$$= \arg\min_{w,b} \frac{1}{2}(\boldsymbol{X}\boldsymbol{w} + \mathbf{1}b - \mathbf{y})^\top(\boldsymbol{X}\boldsymbol{w} + \mathbf{1}b - \mathbf{y}) + \lambda\boldsymbol{w}^\top\boldsymbol{w}$$

由 $b \in \mathbb{R}^m$, 有 $\mathbf{1} \in \mathbb{R}^{m\times m}, \mathbf{1}^\top\mathbf{1} = m$, 则对两参数分别求偏导有:

$$\frac{\partial f(\boldsymbol{w}, b)}{\partial \boldsymbol{w}} = (\boldsymbol{X}^\top\boldsymbol{X} + 2\lambda)\boldsymbol{w} + \boldsymbol{X}^\top(\mathbf{1}b - \mathbf{y})$$

$$\frac{\partial f(\boldsymbol{w}, b)}{\partial b} = \mathbf{1}^\top\boldsymbol{X}\boldsymbol{w} + \mathbf{1}^\top\mathbf{1}b - \mathbf{1}^\top\mathbf{y} = \mathbf{1}^\top\boldsymbol{X}\boldsymbol{w} - \mathbf{1}^\top\mathbf{y} + mb$$

令上两式为 0，联立可得 w,b 的闭式解：

$$\begin{cases} \boldsymbol{w}^*_{\textbf{Ridge}} = (\boldsymbol{X}^\top \boldsymbol{X} + 2\lambda \boldsymbol{I}_d - \dfrac{1}{m}\boldsymbol{X}^\top \mathbf{1}\mathbf{1}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top (\boldsymbol{I}_m - \dfrac{1}{m}\mathbf{1}\mathbf{1}^\top)\mathbf{y} \\ b^*_{\textbf{Ridge}} = \dfrac{1}{m}(1^\top \mathbf{y} - 1^\top \boldsymbol{X}\boldsymbol{w}^*_{\textbf{Ridge}}) \end{cases}$$

和原始线性回归解 $\hat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \mathbf{y}$ 相比，这样得到的解中含有一个 $\frac{2m}{m-1}\lambda \boldsymbol{I}_d$ 正则化项，反映了当前分类器的归纳偏好，通过调节正则项权重 λ，容易产生较为符合预期的模型。

2.

$$\begin{aligned} &(\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_m)^{-1}\boldsymbol{X} = \boldsymbol{X}(X^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d)^{-1} \\ \iff\ &\boldsymbol{X} = (\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_m)\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d)^{-1} \\ \iff\ &\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_d) = (\boldsymbol{X}\boldsymbol{X}^\top + \lambda \boldsymbol{I}_m)\boldsymbol{X} \\ \iff\ &\boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{X}\boldsymbol{I}_d = \boldsymbol{X}\boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}_m\boldsymbol{X} \\ \iff\ &\boldsymbol{X}\boldsymbol{I}_d = \boldsymbol{I}_m\boldsymbol{X} \\ \iff\ &\boldsymbol{X} = \boldsymbol{X} \end{aligned}$$

上式显然正确

当计算岭回归的 $\boldsymbol{w}^*_{\textbf{Ridge}}$ 时，为方便描述，令

$$\boldsymbol{M} = (\boldsymbol{X}^\top \boldsymbol{X} + \frac{2m}{m-1}\lambda)^{-1} \in \mathbb{R}^{d\times d}$$

则用 $\boldsymbol{M}\boldsymbol{X}^\top \mathbf{y}(\mathbb{R}^{d\times d}\times\mathbb{R}^{d\times m}\times\mathbb{R}^{m\times 1})$ 计算时，计算复杂度至少为 $dm^2 + d^3$（先算 $\boldsymbol{X}^\top \mathbf{y} = \boldsymbol{K}$，再算 $\boldsymbol{M}\boldsymbol{K}$）

而由本题式子化简后，令化简后：

$$\boldsymbol{M} = (\boldsymbol{X}\boldsymbol{X}^\top + \frac{2m}{m-1}\lambda)^{-1} \in \mathbb{R}^{m\times m}$$

计算 $\boldsymbol{X}^\top \boldsymbol{M}\mathbf{y}(\mathbb{R}^{d\times m}\times\mathbb{R}^{m\times m}\times\mathbb{R}^{m\times 1})$ 时，最小计算复杂度为 $m^3 + dm^2$

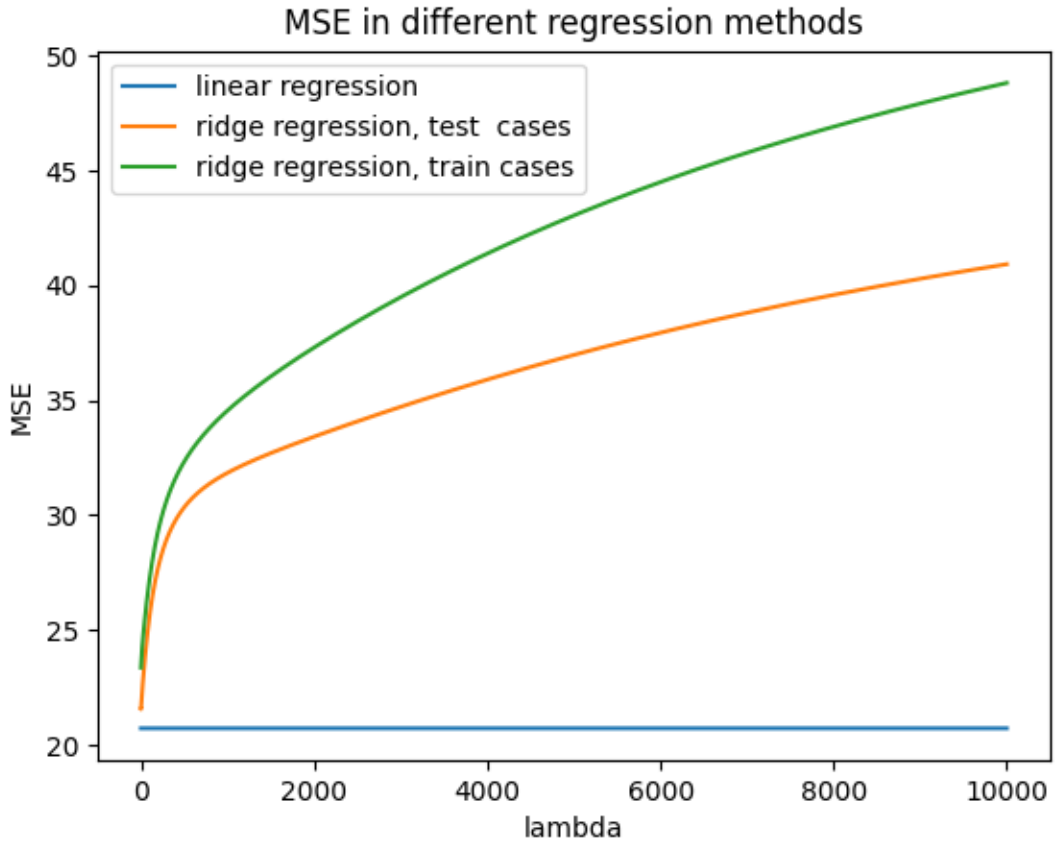
当样例很少，而特征很多时 (即 $m < d$)，采用这个结论可以帮助岭回归减少计算量。

3.

a. 测试集上的 MSE 为：

$$\textbf{MSE} = 20.72402343734257$$

b. 观察到当权重 $\lambda < 0$ 时，测试集上的均方误差很大，远大于图上显示的部分，因此绘制图表时仅考虑 $\lambda > 0$，如图：



随着权重 λ 不断增大，测试集 MSE 和训练集 MSE 都在逐渐变大，但增速逐渐放缓。考虑应该是增加的正则化项影响到了模型对真实情况的模拟，正则化项占比越大，对真实情况的拟合程度越差，反映出人为的偏好。测试代码见附件。

四. (20 points) **线性判别分析**

教材 3.4 节介绍了“线性判别分析”模型 LDA (Linear Discriminative Analysis), 本题首先针对 LDA 从分布假设的角度进行推导和分析. 考虑 N 分类问题, 训练集 $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$, 其中, 第 n 类样例从高斯分布 $\mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ 中独立同分布采样得到 (其中, $n = 1, 2, \cdots, N$). 记该类样例数量为 m_n. 类别先验为 $p(y = n) = \pi_n$, 反映了各类别出现

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

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{5}$$

假设不同类别的条件概率为高斯分布, 当不同类别的协方差矩阵 $\boldsymbol{\Sigma}_n$ 相同时, 对于类别的预测转化为类别中心之间的线性问题, 下面对这一模型进行进一步分析. 假设 $\boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}$, 分析 LDA 的分类方式以及参数估计步骤.

1. 样例 $\boldsymbol{x}$ 的后验概率 $p(y = n \mid \boldsymbol{x})$ 表示了样例属于第 n 类的可能性, 当计算样例针对 N 个类别的后验概率后, 找出后验概率最大的类别对样例的标记进行预测, 即 $\arg\max_n p(y = n \mid \boldsymbol{x})$. 等价于考察 $\ln p(y = n \mid \boldsymbol{x})$ 的大小, 请证明在此假设下,

$$\arg\max_y p(y \mid \boldsymbol{x}) = \arg\max_n \underbrace{\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_n - \frac{1}{2}\boldsymbol{\mu}_n^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_n + \ln \pi_n}_{\delta_n(\boldsymbol{x})}. \tag{6}$$

   其中 $\delta_n(\boldsymbol{x})$ 为 LDA 在分类时的判别函数.

2. 在 LDA 模型中, 需要估计各类别的先验概率, 以及条件概率中高斯分布的参数. 针对二分类问题 ($N = 2$), 使用如下方式估计类别先验、均值与协方差矩阵:

$$\hat{\pi}_n = \frac{m_n}{m}; \quad \hat{\boldsymbol{\mu}}_n = \frac{1}{m_n}\sum_{y_i = n} \boldsymbol{x}_i, \tag{7}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{m-N}\sum_{n=1}^{N}\sum_{y_i=n} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_n)^\top. \tag{8}$$

   LDA 使用这些经验量替代真实参数, 计算判别式 $\delta_n(\boldsymbol{x})$ 并按照第1问中的准则做出预测. 请证明:

$$\boldsymbol{x}^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) > \frac{1}{2}(\hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1)^\top \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1) - \ln(m_2/m_1) \tag{9}$$

   时 LDA 将样例预测为第 2 类. 请分析这一判别方式的几何意义.

3. 在 LDA 中, 对样例 $\boldsymbol{x}$ 的判别可视为在投影的空间中和某个阈值进行比较. 上述推导通过最大后验概率的方法得到对投影后样例分布的需求, 而 Fisher 判别分析 (Fisher Discriminant Analysis, FDA) 也是

一种常见的线性判别分析方法, 直接对样例投影后数据的分布情况进行约束. FDA 一般通过广义瑞利商进行求解, 请基于教材 3.4 节对“线性判别分析” 的介绍，对广义瑞利商的性质进行分析, 探讨 FDA 多分类推广的性质. 下面请说明对于 N 类分类问题, FDA 投影的维度最多为 $N-1$, 即投影矩阵 $\boldsymbol{W} \in \mathbb{R}^{d\times(N-1)}$.
提示：矩阵的秩具有如下性质：对于矩阵 $\mathbf{A} \in \mathbb{R}^{m\times n}$, 矩阵 $\mathbf{B} \in \mathbb{R}^{n\times r}$, 则

$$\operatorname{rank}(\mathbf{A})+\operatorname{rank}(\mathbf{B})-n \leq \operatorname{rank}(\mathbf{AB}) \leq \min\{\operatorname{rank}(\mathbf{A}), \operatorname{rank}(\mathbf{B})\} . \quad (10)$$

对于任意矩阵 $\boldsymbol{A}$, 以下公式成立

$$\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}\left(\mathbf{A}^{\top}\right) = \operatorname{rank}\left(\mathbf{A}\mathbf{A}^{\top}\right) = \operatorname{rank}\left(\mathbf{A}^{\top}\mathbf{A}\right) . \quad (11)$$

解：

1.
当各个类别的协方差矩阵相同时，若 $x \in \mathbb{R}^d \sim \mathcal{N}(\mu_i, \Sigma)$，则 $p(x)$ 可以转化为更简单的形式：

$$p(x) = k \exp\left(-\frac{1}{2}(x-\mu_i)^{\top}\Sigma^{-1}(x-\mu_i)\right)$$

其中 k 为和类别无关的常数，设置为 1 不影响结论，由贝叶斯公式：

$$p(y \mid \boldsymbol{x}) = \frac{p(y)p(\boldsymbol{x} \mid y)}{p(\boldsymbol{x})}$$

可以得到：

$$\begin{aligned}
\arg\max_y p(y \mid \boldsymbol{x}) &= \arg\max_n \frac{p(y=n)p(\boldsymbol{x} \mid y=n)}{p(\boldsymbol{x})} \\
&= \arg\max_n \frac{p(y=n)p(\boldsymbol{x}_n)}{p(\boldsymbol{x})} \\
&= \arg\max_n \ln \frac{p(y=n)p(\boldsymbol{x}_n)}{p(\boldsymbol{x})} \\
&= \arg\max_n \ln p(y=n) + \ln p(\boldsymbol{x}_n) - \ln p(\boldsymbol{x})
\end{aligned}$$

因为 $p(\boldsymbol{x})$ 不随 n 的改变而变化，由此原式转化为：

$$\begin{aligned}\arg\max_y p(y \mid \boldsymbol{x}) &= \arg\max_n \ \ln p(y=n) + \ln p(\boldsymbol{x}_n) \\ &= \arg\max_n \ \ln \pi_n - \frac{1}{2}(x-\mu_n)^\top \Sigma^{-1}(x-\mu_n) \\ &= \arg\max_n \ \ln \pi_n + \boldsymbol{x}^\top \Sigma^{-1}\mu_n - \frac{1}{2}\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{x} - \frac{1}{2}\mu_n^\top\Sigma^{-1}\mu_n\end{aligned}$$

加入不随 n 变化的项 $\frac{1}{2}\boldsymbol{x}^\top\Sigma^{-1}\boldsymbol{x}$ 不改变原优化问题，因此有：

$$\arg\max_y p(y \mid \boldsymbol{x}) = \arg\max_n \ \ln \pi_n + \boldsymbol{x}^\top \Sigma^{-1}\mu_n - \frac{1}{2}\mu_n^\top\Sigma^{-1}\mu_n$$

2. 由判别函数

$$\delta_n(x) = x^\top \hat{\Sigma}^{-1}\hat{\mu}_n - \frac{1}{2}\hat{\mu}_n^\top\hat{\Sigma}^{-1}\hat{\mu}_n + \ln\hat{\pi}_n$$

可以写出预测为 1 类和 2 类的表达式：

$$\delta_1(x) = x^\top \hat{\Sigma}^{-1}\hat{\mu}_1 - \frac{1}{2}\hat{\mu}_1^\top\hat{\Sigma}^{-1}\hat{\mu}_1 + \ln(m_1/m)$$

$$\delta_2(x) = x^\top \hat{\Sigma}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_2^\top\hat{\Sigma}^{-1}\hat{\mu}_2 + \ln(m_2/m)$$

在预测中，预测的 $\delta_n(x)$ 值越大，最终预测为这个分类的概率就越大，因此我们有：

$$\begin{aligned}&\delta_2(x) > \delta_1(x) \\ \iff\ &\delta_2(x) - \delta(_1x) > 0 \\ \iff\ &x^\top\hat{\Sigma}^{-1}(\hat{\mu}_2-\hat{\mu}_1) - (\frac{1}{2}\hat{\mu}_2^\top\hat{\Sigma}^{-1}\hat{\mu}_2 - \frac{1}{2}\hat{\mu}_1^\top\hat{\Sigma}^{-1}\hat{\mu}_1) + \ln(m_2/m_1) > 0 \\ \iff\ &x^\top\hat{\Sigma}^{-1}(\hat{\mu}_2-\hat{\mu}_1) > \frac{1}{2}(\hat{\mu}_2^\top\hat{\Sigma}^{-1}\hat{\mu}_2 - \hat{\mu}_1^\top\hat{\Sigma}^{-1}\hat{\mu}_1) - \ln(m_2/m_1)\end{aligned}$$

$\therefore$ 为证明原来预测正确，只需要证明：

$$\frac{1}{2}(\hat{\mu}_2^\top\hat{\Sigma}^{-1}\hat{\mu}_2 - \hat{\mu}_1^\top\hat{\Sigma}^{-1}\hat{\mu}_1) \le \frac{1}{2}(\hat{\mu}_2+\hat{\mu}_1)^\top\hat{\Sigma}^{-1}(\hat{\mu}_2-\hat{\mu}_1)$$

即证明

$$\hat{\mu}_1^\top\Sigma^{-1}\hat{\mu}_2 \ge \hat{\mu}_2^\top\Sigma^{-1}\hat{\mu}_1 \qquad (a)$$

因为左表达式为一个数，$\hat{\mu}_1^\top \Sigma^{-1} \hat{\mu}_2 \in \mathbb{R}$

∴ 我们有：$\hat{\mu}_1^\top \Sigma^{-1} \hat{\mu}_2 = (\hat{\mu}_1^\top \Sigma^{-1} \hat{\mu}_2)^\top = \hat{\mu}_2^\top \Sigma^{-1} \hat{\mu}_1$

∴(a) 式成立，因此原预测值成立
预测方法 $\arg\max_n \delta_n(x)$ 的几何意义为在向量空间中，使其在 n 号分布上的出现频率最大时的取值 n，也就是反映其分布在此处的疏密程度，将样例判别为在此分布最密集的。
3.
. 原问题的求解等价于求解广义特征值问题：

$$S_b W = \lambda S_w W$$

其闭式解为 $S_w^{-1} S_b$ 的 d' 个最大非零广义特征值所对应的特征向量组成的矩阵, 记这个矩阵为 $\mathfrak{W}$

$$\because S_b = \sum_{i=1}^{N} m_i (\mu_i - \mu)(\mu_i - \mu)^\top$$

且 $rank((\mu_i - \mu)(\mu_i - \mu)^\top) = rank((\mu_i - \mu)) = 1$
∴ S_b 为 N 个秩为 1 的矩阵之和，即

$$rank(S_b) \leq N$$

又 $\mu_i - \mu$ 并非线性无关，因此其和之秩小于 N，有

$$rank(S_b) < N \iff rank(S_b) \leq N - 1$$

又

$$rank(S_w^{-1} S_b) \leq \min(rank(S_w^{-1}), rank(S_b)) \leq rank(S_b) \leq N - 1$$

∴ $S_w^{-1} S_b$ 一共有不超过 N-1 个特征值, 因此有

$$rank(\mathfrak{W}) \leq N - 1$$

五. (20 points) **多分类学习**

教材 3.5 节介绍了“多分类学习”的多种方式, 本题针对 OvO 和 OvR

两种多分类学习方法进行分析:

1. 分析两种多分类方法的优劣. 思考这两种多分类推广方式是否存在难以处理的情况?
2. 在 OvR 的每一个二分类子任务中, 目标类别作为正类, 而其余所有类别作为负类. 此时, 是否需要显式考虑正负类别的不平衡带来的影响?

解:

1.
OvR 只需要训练 N 个分类器，而 OvO 需要 $N(N-1)/2$ 个分类器，因此 OvO 一般比 OVR 需要更大的存储开销和测试时间开销。然而在训练分类器时，OvR 的所有分类器都需要使用所有的训练用例，而 OvO 的每个分类器只需要用到涉及到的两个用例的训练样本，因此在类别很多时，训练时的时间开销 OvO 要小于 OvR
2.
不需要显式考虑。我们对每个类都进行了相同的处理，其产生的类别不平衡效果会相互抵消，通常不会对分类器的性能造成影响。