ML-02

201300086

一、NFL

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?

有必要:

- NFL定理有一个重要前提：所有"问题"出现的机会相同、或所有问题同等重要。但实际情形并不是这样。很多时候，我们只关注自己正在试图解决的问题（例如某个具体应用任务），希望为它找到一个解决方案，至于这个解决方案在别的问题、甚至在相似的问题上是否为好方案，我们并不关心。
- 脱离具体问题，空泛地谈论"什么学习算法更好"毫无意义，因为若考虑所有潜在的问题，则所有学习算法都一样好。要谈论算法的相对优劣，必须要针对具体的学习问题。

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.

对于任意的k分类问题和任意性能度量，假设f均匀分布，满足（注意到与学习算法无关）：

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

c：预测结果的期望

$$\text{特别地，当}k=2\text{时，}\sum_f\ell(h(\boldsymbol{x}),f(\boldsymbol{x}))=\frac{1}{2}2^{|\mathcal{X}|}$$

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

我们发现总误差与学习算法无关，NFL仍成立

二、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.

[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.

令$E_{\boldsymbol{w}}=(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}-\mathbf{1}b)^{\top}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}-\mathbf{1}b)$

$$\frac{\partial E_{\boldsymbol{w}}}{\partial\boldsymbol{w}}=2\mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{w}-\boldsymbol{y}+\mathbf{1}b)$$

$$\frac{\partial E_{\boldsymbol{w}}}{\partial b}=2\mathbf{1}^{\mathrm{T}}(\mathbf{X}\boldsymbol{w}-\boldsymbol{y}+\mathbf{1}b)$$

当$\mathbf{X}^{\mathrm{T}}\mathbf{X}$满秩时，令$\boldsymbol{T}=(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$，偏导为0：

$$\begin{cases}\mathbf{1}^{\top}\mathbf{1}b^*=b^*=\mathbf{1}^{\top}\boldsymbol{y}-\mathbf{1}^{\top}\boldsymbol{X}\boldsymbol{w}^*\\\boldsymbol{w}^*=\boldsymbol{T}\boldsymbol{X}^{\top}(\boldsymbol{y}-\mathbf{1}b^*)\end{cases}$$

最后解出$\boldsymbol{w}^*=\boldsymbol{T}\boldsymbol{X}^{\top}\left(\boldsymbol{y}-\mathbf{1}\dfrac{\mathbf{1}^{\top}\boldsymbol{X}\boldsymbol{T}\boldsymbol{X}^{\top}\boldsymbol{y}-\mathbf{1}^{\top}\boldsymbol{y}}{\mathbf{1}^{\top}\boldsymbol{X}\boldsymbol{T}\boldsymbol{X}^{\top}\mathbf{1}-1}\right)$

三、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:

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

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 $\boldsymbol{w}^*_{\mathbf{Ridge}}$ 和 $b^*_{\mathbf{Ridge}}$ 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 $\boldsymbol{w}^*_{\mathbf{LS}}$ 和 $b^*_{\mathbf{LS}}$ 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?

(1)

$$令E_{\boldsymbol{w}} = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - \mathbf{1}b)^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} - \mathbf{1}b) + \lambda||\boldsymbol{w}||_2^2$$

$$\frac{\partial E_{\boldsymbol{w}}}{\partial \boldsymbol{w}} = \mathbf{X}^{\mathrm{T}}(\mathbf{X}\boldsymbol{w} - \boldsymbol{y} + \mathbf{1}b) + 2\lambda\boldsymbol{w}$$

$$\frac{\partial E_{\boldsymbol{w}}}{\partial b} = \mathbf{1}^{\mathrm{T}}(\mathbf{X}\boldsymbol{w} - \boldsymbol{y} + \mathbf{1}b)$$

$$当\mathbf{X}^{\mathrm{T}}\mathbf{X} + 2\lambda\boldsymbol{I}_d满秩时，令\boldsymbol{T} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + 2\lambda\boldsymbol{I}_d)^{-1}，令偏导为0:$$

$$\begin{cases} b^*_{\mathbf{Ridge}} = \mathbf{1}^\top\boldsymbol{y} - \mathbf{1}^\top\boldsymbol{X}\boldsymbol{w}^*_{\mathbf{Ridge}} \\ \boldsymbol{w}^*_{\mathbf{Ridge}} = \boldsymbol{T}\boldsymbol{X}^\top(\boldsymbol{y} - \mathbf{1}b^*_{\mathbf{Ridge}}) \end{cases}$$

$$最后解出\boldsymbol{w}^*_{\mathbf{Ridge}} = \boldsymbol{T}\boldsymbol{X}^\top\left(\boldsymbol{y} - \mathbf{1}\frac{\mathbf{1}^\top\boldsymbol{X}\boldsymbol{T}\boldsymbol{X}^\top\boldsymbol{y} - \mathbf{1}^\top\boldsymbol{y}}{\mathbf{1}^\top\boldsymbol{X}\boldsymbol{T}\boldsymbol{X}^\top\mathbf{1} - 1}\right)$$

分析:

- 我们发现最优解和原始最优解形式相同，只有矩阵$\boldsymbol{T}$不同
- 特别的，当λ=0时，我们可以认为二者完全相同

(2)

$$
\begin{aligned}
&(\boldsymbol{X}\boldsymbol{X}^\top + \lambda I_m)^{-1}\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda I_d)^{-1} \\
\Leftrightarrow\ &\boldsymbol{X} = (\boldsymbol{X}\boldsymbol{X}^\top + \lambda I_m)\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda I_d)^{-1} \\
\Leftrightarrow\ &\boldsymbol{X} = (\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{X})(\boldsymbol{X}^\top\boldsymbol{X} + \lambda I_d)^{-1} \\
\Leftrightarrow\ &\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + \lambda I_d)(\boldsymbol{X}^\top\boldsymbol{X} + \lambda I_d)^{-1} \\
\Leftrightarrow\ &\boldsymbol{X} = \boldsymbol{X}
\end{aligned}
$$

等价条件显然，证毕

可能有帮助：

两式等价，但矩阵求逆计算开销不同。

当m和d大小有差别时，实际应用中可以选取m和d中较小的对应的求逆进行实际计算，加快计算速度

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.

(3)

（附代码LinearRegression.py）

(i)

线性回归模型在测试集上的MSE：20.724023437340996

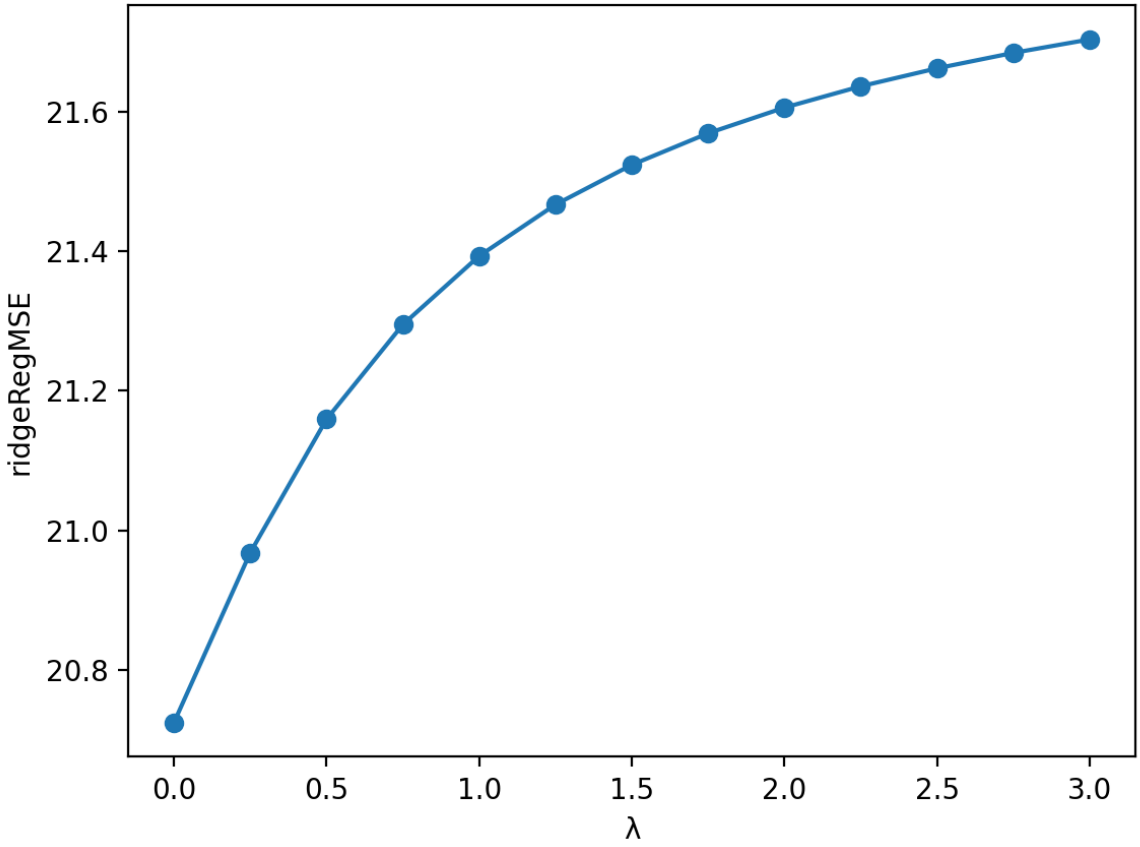
(ii)

λ取值：`lmbds = np.arange(0, 3.1, 0.25)`

得到可视化MSE与λ关系图如下

分析：

- MSE随λ增大而增大，增速逐渐放缓
- 较小的λ对MSE的影响更小，λ大时放大了偏好对MSE的影响



四、LDA

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)

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,

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

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

由贝叶斯公式，并除去与n无关项

$$\begin{aligned}
\arg\max_y p(y \mid \boldsymbol{x}) &= \arg\max_n \frac{p(y=n)p(\boldsymbol{x} \mid y=n)}{p(\boldsymbol{x})} \\
&= \arg\max_n \frac{\pi_n \left((2\pi)^{\frac{d}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}} \right)^{-1} \exp\left(\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_n - \frac{1}{2} \boldsymbol{\mu}_n^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_n - \frac{1}{2} \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \right)}{p(\boldsymbol{x})} \\
&= \arg\max_n \frac{\pi_n \exp\left(\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_n - \frac{1}{2} \boldsymbol{\mu}_n^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_n \right)}{p(\boldsymbol{x})} \\
&= \arg\max_n \pi_n \exp\left(\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_n - \frac{1}{2} \boldsymbol{\mu}_n^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_n \right) \\
&= \arg\max_n \boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_n - \frac{1}{2} \boldsymbol{\mu}_n^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_n + \ln \pi_n
\end{aligned}$$

(2)

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:

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

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

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:

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

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.

证明:

预测为第二类说明：$p(y=2 \mid \boldsymbol{x}) > p(y=1 \mid \boldsymbol{x})$

带入第一问的结论：

$$\Leftrightarrow \boldsymbol{x}^{\top} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_2 - \frac{1}{2} \hat{\boldsymbol{\mu}}_2^{\top} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_2 + \ln \pi_2 > \boldsymbol{x}^{\top} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_1 - \frac{1}{2} \hat{\boldsymbol{\mu}}_1^{\top} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_1 + \ln \pi_1$$

$$\Leftrightarrow \boldsymbol{x}^{\top} \hat{\boldsymbol{\Sigma}}^{-1} \left(\hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1\right) > \frac{1}{2} (\hat{\boldsymbol{\mu}}_2^{\top} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_2 - \hat{\boldsymbol{\mu}}_1^{\top} \boldsymbol{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_1) - \ln\left(\pi_2/\pi_1\right)$$

$\left(\text{由于} \hat{\boldsymbol{\Sigma}}^{-1} \text{是对称矩阵，所以} \hat{\boldsymbol{\mu}}_1^{\top} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_2 = \hat{\boldsymbol{\mu}}_2^{\top} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}_1\right)$

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

即等价为了原不等式

几何意义:

- $\frac{1}{2}\left(\hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1\right)$ 是 $\hat{\boldsymbol{\mu}}_1$ 和 $\hat{\boldsymbol{\mu}}_2$ 这两个类别中心的中点
- 上述不等式成立说明了在引入对数几率$ln(m_2/m_1)$时，$\boldsymbol{x}$和中点$\frac{1}{2}(\hat{\boldsymbol{\mu}}_2 + \hat{\boldsymbol{\mu}}_1)$相比，$\boldsymbol{x}$离$\hat{\boldsymbol{\mu}}_2$更近

(3)

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

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

原优化目标：$\max_{\mathbf{W}} \frac{\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{S}_b\mathbf{W}\right)}{\operatorname{tr}\left(\mathbf{W}^{\mathrm{T}}\mathbf{S}_w\mathbf{W}\right)}$

转化为广义特征值问题：$\mathbf{S}_b\mathbf{W} = \lambda\mathbf{S}_w\mathbf{W}$

闭式解$\mathbf{W} = \mathbf{S}_w^{-1}\mathbf{S}_b$

由于 $\sum_{i=1}^{N} m_i\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}\right) = \sum_{i=1}^{N} m_i\boldsymbol{\mu}_i - \sum_{i=1}^{N} m_i\boldsymbol{\mu} = 0$

说明$\boldsymbol{\mu}_1 - \boldsymbol{\mu}$，$\boldsymbol{\mu}_2 - \boldsymbol{\mu}$，……，$\boldsymbol{\mu}_N - \boldsymbol{\mu}$线性相关，任一项可由其他线性表示

$$\begin{aligned}
\text{即} rank(\boldsymbol{S}_b) &= rank(\sum_{i=1}^{N} m_i\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}\right)(\boldsymbol{\mu}_i - \boldsymbol{\mu})^{\top}) \\
&= rank(\sum_{i=1}^{N-1} m_i\left(\boldsymbol{\mu}_i - \boldsymbol{\mu}\right)(\boldsymbol{\mu}_i - \boldsymbol{\mu})^{\top}) \\
&\leq N-1 \\
rank(\mathbf{S}_w^{-1}\mathbf{S}_b) &\leq \min\left\{\operatorname{rank}\left(\boldsymbol{S}_w^{-1}\right), \operatorname{rank}\left(\boldsymbol{S}_b\right)\right\} \\
&\leq rank(\boldsymbol{S}_b) \\
&\leq N-1
\end{aligned}$$

因此 $\boldsymbol{S}_w^{-1}\boldsymbol{S}_b$ 最多只能对应 N-1 个特征向量，而 $\boldsymbol{W}$的每个列向量都线性无关，所以 FDA 投影的维度最多为 N-1

五、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?

(1)

1、存储开销和测试时间开销：

OvR只需训练N个分类器，而 OvO需训练N（N-1）/2个分类器，因此OvO通常比 OvR 更大

2、训练时间开销：

训练时OvR每个分类器均使用全部训练样例，而 OvO 每个分类器仅用到两个类的样例，因此类别很多时，OvO通常比OvR 更小

3、预测性能：

取决于具体的数据分布，多数情形下两者差不多

4、难以处理的情况：

OvR： 如果类别特别多，则每次训练都要使用全部数据，时间开销$O(n^2)$，难以处理

OvO： 如果样例非常不平均，如第一个类别占了一半的样例，而 OvO需训练N（N-1）/2个分类器，存储和测试时间开销难以处理

(2)

不需要，因为OvR遍历所有类进行相似处理，其过程的对称性可以消除正负类别不平衡所带来的影响