机器学习（西瓜书）

注　解

（第 3 章　线性模型）

https://blog.csdn.net/jbb0523

前言

经常听人说南大周老师所著的《机器学习》（以下统称为西瓜书）是一本入门教材，是一本科普性质的教科书。在该书第十次印刷之际，周老师在“如何使用本书”中也提到“这是一本入门级教科书”。然而，本人读起来却感觉该书远不止“科普”“入门”那么简单，书中的很多公式需要思考良久方能推导，很多概念需要反复咀嚼才能消化。边读边想着是不是应该将自己学习时遇到的一些知识难点的解析分享出来，以帮助更多的人入门。自己的确也随手做过一些笔记，但由于怀疑这仅是自己的个别现象，毕竟读书期间，思考更多的是如何使用单片机、DSP、ARM、FGPA 等，而这些基本是不需要推导任何公式的，因此作罢。偶然间在周老师的新浪微博看到如下对话：

此时方知，可能“读不懂”并不是个别现象。因此决定写一本“西瓜书注解”或者称为“西瓜书读书笔记”，对自己研读西瓜书时遇到的“台阶”进行解释和推导，以帮助更多的人能够更快地进入到这个领域。另外，近期越来越强地意识到，扎扎实实地推导一些基础算法的公式，无论是对于理解算法本身机理还是进行学术研究，都是非常有必要的。

自己会根据个人学习进度和研究需要按章发布，不知道能不能坚持写完，加油！

毕竟自己也是一名初学者，所以可能一些概念解释并不完整、一些公式推导并不优美，甚至会存在错误，这是不可避免的，不接受漫骂，但欢迎将问题反馈给我，共同学习进步！

（网盘链接：https://pan.baidu.com/s/1QtEiNnk8jMzmbs0KPBN-_w）

第 3 章目录

第 3 章 线性模型.....1
3.1 基本形式.....1
3.2 线性回归.....1
1、离散属性连续化问题.....1
2、式(3.4)的解释.....1
3、式(3.7)的推导.....2
4、式(3.15)的解释.....3
3.3 对数几率回归.....3
1、式(3.19)的推导.....3
2、式(3.20)和式(3.21)的解释.....3
3、式(3.23)和式(3.24)的推导.....4
4、式(3.25)的解释.....4
5、式(3.26)的推导.....4
6、式(3.27)的推导.....4
7、式(3.29)的解释.....6
8、式(3.30)的推导.....6
9、式(3.31)的推导.....7
3.4 线性判别分析.....8
1、式(3.36)的推导.....8
2、式(3.37)的推导.....8
3、式(3.38)的推导.....8
4、式(3.39)的推导.....9
5、式(3.40)的解释.....9
6、式(3.43)的推导.....9
7、式(3.44)的解释.....10
8、式(3.45)的推导[?].....10
3.5 多分类学习.....11
1、图 3.4 的解释.....11
2、图 3.5 的解释.....11
3、ECOC 编码长度的解释.....11
3.6 类别不平衡问题.....12
1、式(3.48)的推导.....12
3.7 本章小节.....12

第 3 章 线性模型

如西瓜书前言所述，本章仍属于第 1 部分机器学习基础知识；作为西瓜书介绍机器学习模型的开篇，线性模型也是机器学习中最为基础的模型，很多复杂模型均可认为由线性模型衍生而得。

3.1 基本形式

式(3.2)中$\boldsymbol{w}=(w_1;w_2;\ldots;w_d)$，这里向量元素之间分号“;”表示列元素分隔符，即$\boldsymbol{w}$表示$d\times 1$的列向量；相应的，逗号“,”表示行元素分隔符，如$\boldsymbol{\alpha}=(a_1,a_2,\ldots,a_m)$表示$1\times m$的行向量；该规则同样适用于矩阵，这与 Matlab 软件中的语法习惯一致。

3.2 线性回归

本节符号“$\simeq$”是“约等于”。

1、离散属性连续化问题

本节第 2 段提到离散属性连续化问题，这里补充一下。

对于存在“序”关系的属性值，很好理解，额外要注意的仅是转化时要结合属性规范化（参见 P48 习题 2.8）确定上下限；对于不存在“序”关系的属性值，则将该维属性展为多个属性，例如书中提到的“瓜类”属性，在原本仅为一维，而展开后变为三维，每一维代表原属性的一个离散值。再举一个例子，对于 P76 表 4.1 的西瓜数据集 2.0 来说，属性包括色泽、根蒂、敲声、纹理、脐部、触感共六维特征属性，各维特征取值均为无序离散值（此处举例就这么认为而已，例如你也可以认为色泽的浅白、青绿、乌黑属于有序属性），因此离散属性连续化后，色泽展为三维（浅白、青绿、乌黑）、根蒂展为三维（蜷缩、稍蜷、硬挺）、敲声展为三维（沉闷、浊响、清脆）、纹理展为三维（模糊、稍糊、清晰）、脐部展为三维（凹陷、稍凹、平坦）、触感展为两维（硬滑、软粘），共计 17 维 0/1 二值属性；注意，数据集中 “好瓜”列属于类别标记，编号什么都不算~

2、式(3.4)的解释

解释一下符号“arg min”，其中“arg”是 “argument”的前三个字母，“min” 是“minimum”的前三个字母。维基百科中有“arg max”的解释：https://en.wikipedia.org/wiki/Arg_max，是其反义符号。概括起来，式(3.4)表示求出使目标函数$\sum_{i=1}^{m}(y_i-wx_i-b)^2$最小的参数$(w,b)$，注意目标函数中$x_i,y_i$是常量，即训练集样本。

类似的符号还有“min”，例如将式(3.4)改为

$$\min_{(w,b)}\sum_{i=1}^{m}(y_i-wx_i-b)^2$$

则表示求目标函数的最小值。对比知道，“min”和“arg min”的区别在于，前者输出目标函数的最小值，而后者输出取得目标函数最小值时的参数(w,b)，即“argument”。

若进一步修改式(3.4)为（此处修改并不针线性回归，仅说明符号含义）：

$$\min_{(w,b)} \sum_{i=1}^{m}(y_i - wx_i - b)^2, \text{ s.t. } w > 0, b < 0$$

则表示在$w > 0, b < 0$范围内寻找目标函数$\sum_{i=1}^{m}(y_i - wx_i - b)^2$的最小值并输出，“s.t.”是“subject to”的简写，意思是“受约束于”，$w > 0, b < 0$即为约束条件。

以上介绍的符号都是**最优化方法**中的内容，若想进一步了解可找一本最优化方法的教材进行系统性地学习。另外，实际使用时“min”和“arg min”的区别并不十分明显，因为二者的目标都是解得参数(w, b)，所以有时不必较真。

3、式(3.7)的推导

推导之前先重点强调一下“闭式解(closed-form)”。所谓闭式解，即可以通过某个具体的表达式解出待解参数，如可根据式(3.7)直接解得w。机器学习算法很少有闭式解，线性回归是一个特例。接下来推导式(3.7)：

令式(3.5)为零，得

$$w\sum_{i=1}^{m}x_i^2 = \sum_{i=1}^{m}(y_i - b)x_i$$

将b的表达式(3.8)代入等号右侧（令式(3.6)等于 0，移项，两边同除以m即得式(3.8)），得

$$\begin{aligned}
w\sum_{i=1}^{m}x_i^2 &\overset{①}{=} \sum_{i=1}^{m}\left(y_i - \frac{1}{m}\sum_{j=1}^{m}(y_j - wx_j)\right)x_i \overset{②}{=} \sum_{i=1}^{m}y_ix_i - \sum_{i=1}^{m}\frac{1}{m}\sum_{j=1}^{m}(y_j - wx_j)x_i \\
&\overset{③}{=} \sum_{i=1}^{m}y_ix_i - \frac{1}{m}\sum_{i=1}^{m}x_i\sum_{j=1}^{m}(y_j - wx_j) \overset{④}{=} \sum_{i=1}^{m}y_ix_i - \bar{x}\sum_{j=1}^{m}(y_j - wx_j) \\
&\overset{⑤}{=} \sum_{i=1}^{m}y_ix_i - \bar{x}\sum_{i=1}^{m}(y_i - wx_i) \overset{⑥}{=} \sum_{i=1}^{m}y_ix_i - \bar{x}\sum_{i=1}^{m}y_i + \bar{x}\sum_{i=1}^{m}wx_i \\
&\overset{⑦}{=} \sum_{i=1}^{m}y_ix_i - \sum_{i=1}^{m}y_i\bar{x} + w\bar{x}\sum_{i=1}^{m}x_i \\
&\overset{⑧}{=} \sum_{i=1}^{m}y_i(x_i - \bar{x}) + w\cdot(\frac{1}{m}\sum_{i=1}^{m}x_i)\cdot(\sum_{i=1}^{m}x_i) \\
&\overset{⑨}{=} \sum_{i=1}^{m}y_i(x_i - \bar{x}) + w\frac{1}{m}\left(\sum_{i=1}^{m}x_i\right)^2
\end{aligned}$$

上式中，第 1 个等号中代入式(3.8)时之所以换用求和变量j是因为外层还有求和；第 2 个等号即将求和项分别求和；第 3 个等号是因为x_i与求和变量j无关，可以作为常量提到求和号外面；第 4 个等号用$\bar{x}$代替了平均值公式；第 5 个等号将求和变量j换为变量i，求和变量用谁来表示并不影响，实在不明白把求和号拆开就是了；第 6 个等号类似于第 2 个等号，将求和项分别求和；第 7 个等号由于$\bar{x}$和w是常量，所以放在求和号内部（中间项）和外部（第三项）；第 8 个等号首先是合并了前两项，后面又把$\bar{x}$换回了平均值公式，其实第二项的最后求和部分再除以m就是$\bar{x}$了，因此第二项实际是$wm\bar{x}^2$，在第 9 个等号只需把$\bar{x}$换回平均值公式再消去一个m就是了。此时即有

$$w\sum_{i=1}^{m}x_i^2 = \sum_{i=1}^{m}y_i(x_i - \bar{x}) + w\frac{1}{m}\left(\sum_{i=1}^{m}x_i\right)^2$$

移项，得

$$w\left(\sum_{i=1}^{m}x_i^2-\frac{1}{m}\left(\sum_{i=1}^{m}x_i\right)^2\right)=\sum_{i=1}^{m}y_i(x_i-\bar{x})$$

解得即式(3.7)

$$w=\frac{\sum_{i=1}^{m}y_i(x_i-\bar{x})}{\sum_{i=1}^{m}x_i^2-\frac{1}{m}\left(\sum_{i=1}^{m}x_i\right)^2}$$

4、式(3.15)的解释

将该式变形为$g(y)=\boldsymbol{w}^\top\boldsymbol{x}+b$，当$g(y)=\ln y$时即为式(3.14)。也就是说，线性回归虽然是线性模型，但可以求解$(\boldsymbol{w},b)$使$\boldsymbol{w}^\top\boldsymbol{x}+b$逼近$g(y)$，然后可得$y=g^{-1}(y)$，相当于使用线性回归实现了一个非线性模型。

具体来说，对于数据集$D=\{\boldsymbol{x}_i,y_i\}_{i=1}^m$，先将其转化为$\widetilde{D}=\{\boldsymbol{x}_i,g(y_i)\}_{i=1}^m$，在此函数$g(\cdot)$形式已知；然后根据$\widetilde{D}$训练一个线性回归模型$(\boldsymbol{w},b)$；对于未见示例$\boldsymbol{x}$，线性回归模型输出的预测值为$g(\hat{y})=\boldsymbol{w}^\top\boldsymbol{x}+b$，即$\hat{y}=g^{-1}(\boldsymbol{w}^\top\boldsymbol{x}+b)$。

3.3 对数几率回归

对数几率回归，简称对率回归，即 logistic regression，国内大多译为逻辑回归，在《统计学习方法》(李航 著)一书中则直接音译为“逻辑斯谛回归”。

1、式(3.19)的推导

将式(3.18)代入式(3.19)的等号左侧，得

$$\ln\frac{y}{1-y}=\ln\frac{\frac{1}{1+e^{-(\boldsymbol{w}^\top\boldsymbol{x}+b)}}}{1-\frac{1}{1+e^{-(\boldsymbol{w}^\top\boldsymbol{x}+b)}}}$$

对$\ln(\cdot)$内部进行整理，分子分母同乘以$1+e^{-(\boldsymbol{w}^\top\boldsymbol{x}+b)}$，得

$$\ln\frac{y}{1-y}=\ln\frac{1}{1+e^{-(\boldsymbol{w}^\top\boldsymbol{x}+b)}-1}=\ln\frac{1}{e^{-(\boldsymbol{w}^\top\boldsymbol{x}+b)}}=\ln e^{\boldsymbol{w}^\top\boldsymbol{x}+b}=\boldsymbol{w}^\top\boldsymbol{x}+b$$

由上式可知，对应到式(3.15)，此处$g(y)$为

$$g(y)=\ln\frac{y}{1-y}$$

但此处与图 3.1 中的例子有很大不同。对率回归用来处理二分类任务，$D=\{\boldsymbol{x}_i,y_i\}_{i=1}^m$中的$y_i\in\{0,1\}$，若使用$g(y)$对数据集$D$进行转化，则$g(y)$要么为正无穷$(y_i=1)$，要么为负无穷$(y_i=0)$，因此不能简单类似于式(3.13)~式(3.15)的方法求解模型$(\boldsymbol{w},b)$。实际上$g(y)$中的y是样本为正例(1)的可能性，而非二值化的 0/1。

2、式(3.20)和式(3.21)的解释

书中，将正例的可能性与反例的可能性的比例称为“几率”，即式(3.20)。有关“几率”的概念，在网上搜了好久，很多搜索结果将“几率”和“概率”等价，也有部分搜索结果按式(3.20)解释，但这些结果成文年份大多晚于西瓜书出版时间(2016 年 1 月)，因此很可能是

看完西瓜书后所写。式(3.20)是“几率”，式(3.21)则是对数几率，而本节是用$\boldsymbol{w}^\top \boldsymbol{x}+b$逼近式(3.21)的对数几率，因此将 logistic regression 译为“对率回归”实为信达雅之译~

3、式(3.23)和式(3.24)的推导

注意本页(Page59)第一段 “若将式(3.18)中的y视为类后验概率估计$p(y=1 \mid \boldsymbol{x})$”，即$p(y=1 \mid \boldsymbol{x})=y$，而$p(y=0 \mid \boldsymbol{x})=1-y$。式(3.23)就是式(3.18)的变形，式(3.24)就是用 1 减去式(3.23)的结果。

4、式(3.25)的解释

由式(3.23)和式(3.24)可知$p(y=1 \mid \boldsymbol{x})$和$p(y=0 \mid \boldsymbol{x})$与$(\boldsymbol{w}, b)$有关。式(3.25)只是将式(3.23)和式(3.24)换了一种写法，将$(\boldsymbol{w}, b)$放入了$p(y=1 \mid \boldsymbol{x})$和$p(y=0 \mid \boldsymbol{x})$之中而已。

5、式(3.26)的推导

注意此处y_i要么等于 0，要么等于 1，分别将$y_i=0$和$y_i=1$代入即可：

$$p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b)=\begin{cases} p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) & \text{if } y_i=1 \\ p_0(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) & \text{if } y_i=0 \end{cases}$$

6、式(3.27)的推导

对于式(3.27)，将式(3.23)和式(3.24)代入式(3.26)，然后将式(3.26)代入式(3.25)，得

$$\begin{aligned}
\ell(\boldsymbol{w}, b) &\overset{①}{=} \sum_{i=1}^{m} \ln p(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b) \\
&\overset{②}{=} \sum_{i=1}^{m} \ln \left(y_i p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})+(1-y_i) p_0(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})\right) \\
&\overset{③}{=} \sum_{i=1}^{m} \ln \left(y_i \frac{e^{\boldsymbol{w}^\top \boldsymbol{x}_i+b}}{1+e^{\boldsymbol{w}^\top \boldsymbol{x}_i+b}}+(1-y_i) \frac{1}{1+e^{\boldsymbol{w}^\top \boldsymbol{x}_i+b}}\right) \\
&\overset{④}{=} \sum_{i=1}^{m} \ln \left(y_i \frac{e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i}}+(1-y_i) \frac{1}{1+e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i}}\right) \\
&\overset{⑤}{=} \sum_{i=1}^{m} \ln \left(\frac{y_i e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i}+(1-y_i)}{1+e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i}}\right) \\
&\overset{⑥}{=} \sum_{i=1}^{m} \left(\ln (y_i e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i}+(1-y_i))-\ln (1+e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i})\right)
\end{aligned}$$

其中，第 1 个等号即式(3.25)；第 2 个等号即式(3.26)代入式(3.25)中；第 3 个等号即将式(3.23)和式(3.24)代入；第 4 个等号即将$\boldsymbol{w}^\top \boldsymbol{x}_i+b$用$\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i$代替；第 5 个等号即通分；第 6 个等号利用对数的性质$\ln(a/b)=\ln a-\ln b$。

如书中式(3.27)左侧所注，考虑$y_i \in \{0,1\}$，当$y_i=0$时：

$$\begin{aligned}\ell(\boldsymbol{w},b)&=\sum_{i=1}^{m}\Big(\ln(y_i e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i}+(1-y_i))-\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})\Big)\\&=\sum_{i=1}^{m}\Big(\ln(0\cdot e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i}+(1-0))-\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})\Big)\\&=\sum_{i=1}^{m}\Big(0-\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})\Big)\\&=\sum_{i=1}^{m}\Big(0\cdot\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i-\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})\Big)\\&=\sum_{i=1}^{m}\Big(y_i\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i-\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})\Big)\end{aligned}$$

当$y_i=1$时：

$$\begin{aligned}\ell(\boldsymbol{w},b)&=\sum_{i=1}^{m}\Big(\ln(y_i e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i}+(1-y_i))-\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})\Big)\\&=\sum_{i=1}^{m}\Big(\ln(1\cdot e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i}+(1-1))-\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})\Big)\\&=\sum_{i=1}^{m}\Big(\ln(e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})-\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})\Big)\\&=\sum_{i=1}^{m}\Big(1\cdot\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i-\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})\Big)\\&=\sum_{i=1}^{m}\Big(y_i\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i-\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})\Big)\end{aligned}$$

即无论$y_i=0$还是$y_i=1$，均有$\ell(\boldsymbol{w},b)=\sum_{i=1}^{m}\Big(y_i\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i-\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})\Big)$。

最大化式(3.25)的$\ell(\boldsymbol{w},b)$等价于最小化$-\ell(\boldsymbol{w},b)$（原话参见式(3.27)上面一句话），令$\ell(\boldsymbol{\beta})=-\ell(\boldsymbol{w},b)$，将刚才$y_i=0$和$y_i=1$的化简结果代入，得式(3.27)：

$$\begin{aligned}\ell(\boldsymbol{\beta})&=-\ell(\boldsymbol{w},b)\\&=-\sum_{i=1}^{m}\Big(y_i\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i-\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})\Big)\\&=\sum_{i=1}^{m}\Big(-y_i\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i+\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i})\Big)\end{aligned}$$

其实式(3.27)的化简结果形式比较别扭，尤其是到了第六章与支持向量机作对比的时候就更显得捉襟见肘了，因此将其进一步整理为

$$\ell(\boldsymbol{\beta})=\begin{cases}\sum_{i=1}^{m}\ln(1+e^{-\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i}) & \text{if } y_i=1\\ \sum_{i=1}^{m}\ln(1+e^{\boldsymbol{\beta}^\top\hat{\boldsymbol{x}}_i}) & \text{if } y_i=0\end{cases}$$

整理过程中，$y_i=0$时直接代入即得；对于$y_i=1$，将$y_i=1$代入：

$$\begin{aligned}\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{m}\left(-\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i + \ln(1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i})\right)\\ &= \sum_{i=1}^{m}\left(\ln(e^{-\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}) + \ln(1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i})\right)\\ &= \sum_{i=1}^{m}\ln\left(e^{-\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}(1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i})\right)\\ &= \sum_{i=1}^{m}\ln\left(1+e^{-\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}\right)\end{aligned}$$

整理过程使用了对数的性质$\ln(a\cdot b)=\ln a+\ln b$。这样变形后的好处在于，若正例标记仍用$y_i=1$表示，而反例标记换用$y_i=-1$表示，此时可将这种形式统一写为：

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m}\ln(1+e^{-y_i\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i})$$

而这几乎就是对率损失函数（P130 式(6.33)），详见第 6 章。

7、式(3.29)的解释

其实，这就是牛顿法的迭代公式；注意这里的“牛顿法”是无约束最优化方法中的一种，而不是求解方程$f(x)=0$的根的牛顿迭代。不同于梯度下降法只需目标函数的一阶导数信息，牛顿法需要目标函数的二阶导数信息。以下依据《最优化方法(第二版)》（孙文瑜、徐成贤、朱德通 编著，高等教育出版社 2010 年出版），简单说明牛顿法原理。

设$f(\boldsymbol{x})$二次连续可微，$\boldsymbol{x}_k\in\mathbb{R}^n$，Hesse 矩阵$\nabla^2 f(\boldsymbol{x}_k)$正定。我们在$\boldsymbol{x}_k$附近用二次泰勒展开近似$f(\boldsymbol{x})$，

$$f(\boldsymbol{x})\approx f(\boldsymbol{x}_k)+\nabla f(\boldsymbol{x}_k)^{\top}(\boldsymbol{x}-\boldsymbol{x}_k)+\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_k)^{\top}\nabla^2 f(\boldsymbol{x}_k)(\boldsymbol{x}-\boldsymbol{x}_k)$$

若令$\boldsymbol{s}=\boldsymbol{x}-\boldsymbol{x}_k$（自变量迭代增量，即$\boldsymbol{x}=\boldsymbol{x}_k+\boldsymbol{s}$），则上式变为

$$f(\boldsymbol{x}_k+\boldsymbol{s})\approx f(\boldsymbol{x}_k)+\nabla f(\boldsymbol{x}_k)^{\top}\boldsymbol{s}+\tfrac{1}{2}\boldsymbol{s}^{\top}\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{s}$$

令$q^k(\boldsymbol{s})=f(\boldsymbol{x}_k)+\nabla f(\boldsymbol{x}_k)^{\top}\boldsymbol{s}+\frac{1}{2}\boldsymbol{s}^{\top}\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{s}$为$f(\boldsymbol{x})$的二次近似，求解自变量迭代增量$\boldsymbol{s}$使$q^k(\boldsymbol{s})$极小化，即对$q^k(\boldsymbol{s})$求导并令导数等于零：

$$\nabla q^k(\boldsymbol{s})=\nabla f(\boldsymbol{x}_k)+\nabla^2 f(\boldsymbol{x}_k)\boldsymbol{s}=0$$

解得

$$\boldsymbol{s}=-[\nabla^2 f(\boldsymbol{x}_k)]^{-1}\nabla f(\boldsymbol{x}_k)$$

代入$\boldsymbol{x}=\boldsymbol{x}_k+\boldsymbol{s}$可得迭代公式

$$\boldsymbol{x}_{k+1}=\boldsymbol{x}_k-[\nabla^2 f(\boldsymbol{x}_k)]^{-1}\nabla f(\boldsymbol{x}_k)$$

对应到式(3.29)中

$$\boldsymbol{\beta}^t=\boldsymbol{x}_k,\ [\nabla^2 f(\boldsymbol{x}_k)]^{-1}=\left(\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\top}}\right)^{-1},\ \nabla f(\boldsymbol{x}_k)=\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}$$

8、式(3.30)的推导

先来练习一个高等数学中的复合函数求导习题：

求$f(x)=\ln(1+e^{ax})$的导数。

【分析】这是一个复合函数求导问题，可分解为$f(\cdot)=\ln u(\cdot)$和$u(x)=1+e^{ax}$

【解】$f'(x)=\frac{1}{1+e^{ax}}\cdot ae^{ax}$

有了以上基础，现在可以来推导式(3.30)了，其实就是一个函数求导：

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} =& \frac{\partial \sum_{i=1}^{m}\left(-y_i\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i+\ln(1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i})\right)}{\partial \boldsymbol{\beta}}\\ =&\sum_{i=1}^{m}\left(\frac{\partial(-y_i\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i)}{\partial\boldsymbol{\beta}}+\frac{\partial\ln(1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i})}{\partial\boldsymbol{\beta}}\right)\\ =&\sum_{i=1}^{m}\left(-y_i\hat{\boldsymbol{x}}_i+\frac{1}{1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}\cdot\hat{\boldsymbol{x}}_i e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}\right)\\ =&-\sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\left(y_i-\frac{e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}\right)\\ =&-\sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\left(y_i-p_1(\hat{\boldsymbol{x}}_i;\boldsymbol{\beta})\right)\end{aligned}$$

最后一步即将式(3.23)中$\boldsymbol{w}^{\top}\boldsymbol{x}_i+b$用$\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i$代替，代入即可。

9、式(3.31)的推导

复习一个求导公式：$\left(\frac{u}{v}\right)'=\frac{u'v-uv'}{v^2}$

继续对式(3.30)倒数第二个等号的结果求导：

$$\begin{aligned}\frac{\partial^2\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^{\top}}=&-\frac{\partial\sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\left(y_i-\frac{e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}\right)}{\partial\boldsymbol{\beta}^{\top}}=-\sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\frac{\partial\left(y_i-\frac{e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}\right)}{\partial\boldsymbol{\beta}^{\top}}\\ =&-\sum_{i=1}^{m}\hat{\boldsymbol{x}}_i\left(\frac{\partial y_i}{\partial\boldsymbol{\beta}^{\top}}-\frac{\partial\left(\frac{e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}\right)}{\partial\boldsymbol{\beta}^{\top}}\right)\end{aligned}$$

其中第一项偏导$\frac{\partial y_i}{\partial\boldsymbol{\beta}^{\top}}=0$，第二项偏导

$$\begin{aligned}\frac{\partial\left(\frac{e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}\right)}{\partial\boldsymbol{\beta}^{\top}}=&\frac{\frac{\partial e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}{\partial\boldsymbol{\beta}^{\top}}\cdot\left(1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}\right)-e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}\cdot\frac{\partial\left(1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}\right)}{\partial\boldsymbol{\beta}^{\top}}}{\left(1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}\right)^2}\\ =&\frac{\hat{\boldsymbol{x}}_i^{\top}e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}\cdot\left(1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}\right)-e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}\cdot\hat{\boldsymbol{x}}_i^{\top}e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}{\left(1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}\right)^2}\\ =&\hat{\boldsymbol{x}}_i^{\top}e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}\frac{\left(1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}\right)-e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}{\left(1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}\right)^2}\\ =&\hat{\boldsymbol{x}}_i^{\top}\cdot\frac{e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}\cdot\frac{1}{1+e^{\boldsymbol{\beta}^{\top}\hat{\boldsymbol{x}}_i}}\end{aligned}$$

将两项偏导结果代入，

$$\begin{aligned}\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\top}} &= -\sum_{i=1}^{m} \hat{\boldsymbol{x}}_i \left(0 - \hat{\boldsymbol{x}}_i^{\top} \cdot \frac{e^{\boldsymbol{\beta}^{\top} \hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\top} \hat{\boldsymbol{x}}_i}} \cdot \frac{1}{1+e^{\boldsymbol{\beta}^{\top} \hat{\boldsymbol{x}}_i}}\right) \\ &= \sum_{i=1}^{m} \hat{\boldsymbol{x}}_i \cdot \hat{\boldsymbol{x}}_i^{\top} \cdot \frac{e^{\boldsymbol{\beta}^{\top} \hat{\boldsymbol{x}}_i}}{1+e^{\boldsymbol{\beta}^{\top} \hat{\boldsymbol{x}}_i}} \cdot \frac{1}{1+e^{\boldsymbol{\beta}^{\top} \hat{\boldsymbol{x}}_i}} \\ &= \sum_{i=1}^{m} \hat{\boldsymbol{x}}_i \hat{\boldsymbol{x}}_i^{\top} p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta}) \left(1 - p_1(\hat{\boldsymbol{x}}_i; \boldsymbol{\beta})\right)\end{aligned}$$

3.4 线性判别分析

首先，应该弄清楚线性判别分析的功能：可以认为线性判别分析同上一节的对数几率回归(logistic regression)一样，也是一种二分类算法。对数几率回归在训练集上学得预测模型

$$y = \frac{1}{1+e^{-(\boldsymbol{w}^{\top}\boldsymbol{x}+b)}}$$

对于新样本$\boldsymbol{x}$，将$\boldsymbol{x}$代入以上预测模型得到一个y值，若$y > 0.5$判为正例，反之判为反例；

线性判别分析首先在训练集上学得预测模型$\boldsymbol{w}$

$$y = \boldsymbol{w}^{\top}\boldsymbol{x}$$

训练集中的每一个$\boldsymbol{x}$均由此模型均对应一个y（其实是$\boldsymbol{x}$在直线$\boldsymbol{w}$上的投影），线性判别分析能够保证训练集的同类样本在$\boldsymbol{w}$上的投影y很近，而异类样本在$\boldsymbol{w}$上的投影y很远。对于新样本$\boldsymbol{x}$，首先得到它在直接$\boldsymbol{w}$上的投影，然后判别这个投影与哪一类投影更近，则得到$\boldsymbol{x}$的预测类别。

最后，线性判别分析也是一种降维方法，但不同于第 10 章介绍的无监督降维方法，线性判别分析是一种监督降维方法，即降维过程需要样本类别信息。

1、式(3.36)的推导

式(3.36)是将式(3.35)的分母等于 1 作为了约束条件，然后添加了一个负号，将最大化式(3.35)变成了式(3.36)中最小化目标函数。

2、式(3.37)的推导

由拉格朗日乘子法，式(3.36)的拉格朗日函数为

$$L(\boldsymbol{w}, \lambda) = -\boldsymbol{w}^{\top}\mathbf{S}_b\boldsymbol{w} + \lambda(\boldsymbol{w}^{\top}\mathbf{S}_w\boldsymbol{w} - 1)$$

对变量$\boldsymbol{w}$求导，得

$$\frac{\partial L(\boldsymbol{w}, \lambda)}{\partial \boldsymbol{w}} = -\mathbf{S}_b\boldsymbol{w} + \lambda\mathbf{S}_w\boldsymbol{w}$$

令导数等于零，移项即得式(3.37)。

3、式(3.38)的推导

对于式(3.38)，注意到式(3.34)

$$\mathbf{S}_b = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\top}$$

其中$\boldsymbol{\mu}_0$和$\boldsymbol{\mu}_1$均与$\boldsymbol{x}$维度相同，均为d维列向量，因此$\mathbf{S}_b$是一个$d \times d$维的矩阵；而$\boldsymbol{w}$也是d维列向量，因此

$$\mathbf{S}_b\boldsymbol{w} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\top}\boldsymbol{w} = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)\left((\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^{\top}\boldsymbol{w}\right)$$

注意到后面用括号括在一起的实际是一个标量（就是一个数而已），因此得式(3.38)

$$\mathbf{S}_b\boldsymbol{w}=\zeta(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)$$

这里将式(3.38)中的符号λ换为了ζ（在下式的推导中再解释原因）。

4、式(3.39)的推导

将刚刚得到的式(3.38)表达式$\mathbf{S}_b\boldsymbol{w}=\zeta(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)$替换式(3.37)等号左侧内容，得

$$\zeta(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)=\lambda\mathbf{S}_w\boldsymbol{w}$$

根据线性代数的知识，很容易得到

$$\boldsymbol{w}=\frac{\zeta}{\lambda}\mathbf{S}_w^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)$$

注意，前面曾提到式(3.35)的解与$\boldsymbol{w}$的长度无关，因此令$\frac{\zeta}{\lambda}=1$则得式(3.39)。

个人认为这样解释更通俗易懂一些，因为式(3.37)与式(3.38)中的λ并不是同一个数，如果直接将书中的式(3.38)代入式(3.37)则会让人产生错觉，以为是两边都有λ进而消掉了；所以本人在推导式(3.38)时将符号λ换为了ζ，说明与式(3.37)中的拉格拉日常数不同。

5、式(3.40)的解释

对于式(3.40)，最大的混淆在于别跟前面的式(3.33)和式(3.34)扯上关系就好了，这就是一个定义，注意此式前面有句话是“我们先定义‘全局散度矩阵’”，即它并不是由谁推导而来的，而是在多分类场景下定义的“全局散度矩阵”，仅此而已。

另外，多分类任务是下一节的内容，而作者在此却先讲多分类 LDA，个人感觉把 LDA 放在本章最后一节可能更合适。

6、式(3.43)的推导

首先将式(3.40)进行变形

$$\mathbf{S}_t=\sum_{i=1}^{m}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\top=\sum_{i=1}^{N}\sum_{\boldsymbol{x}\in X_i}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top$$

其中X_i表示第$i\in\{1,\dots,N\}$类示例的集合。这个变形道理很简单，就是将原来依次遍历每个示例求和的循环，分为了N个按类别求和的小循环（求和与求和次序无关）。

将式(3.42)代入式(3.41)，将式(3.40)的变形和式(3.41)代入式(3.43)

$$\begin{aligned}
\mathbf{S}_b&=\mathbf{S}_t-\mathbf{S}_w\\
&=\sum_{i=1}^{N}\sum_{\boldsymbol{x}\in X_i}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top-\sum_{i=1}^{N}\sum_{\boldsymbol{x}\in X_i}(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top\\
&=\sum_{i=1}^{N}\sum_{\boldsymbol{x}\in X_i}\left((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^\top-(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top\right)\\
&=\sum_{i=1}^{N}\sum_{\boldsymbol{x}\in X_i}\left((\boldsymbol{x}\boldsymbol{x}^\top-\boldsymbol{x}\boldsymbol{\mu}^\top-\boldsymbol{\mu}\boldsymbol{x}^\top+\boldsymbol{\mu}\boldsymbol{\mu}^\top)-(\boldsymbol{x}\boldsymbol{x}^\top-\boldsymbol{x}\boldsymbol{\mu}_i^\top-\boldsymbol{\mu}_i\boldsymbol{x}^\top+\boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top)\right)\\
&=\sum_{i=1}^{N}\sum_{\boldsymbol{x}\in X_i}\left(-\boldsymbol{x}(\boldsymbol{\mu}^\top-\boldsymbol{\mu}_i^\top)-(\boldsymbol{\mu}-\boldsymbol{\mu}_i)\boldsymbol{x}^\top+(\boldsymbol{\mu}\boldsymbol{\mu}^\top-\boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top)\right)
\end{aligned}$$

设第i类示例数为m_i，将上式第二个求和号由$\sum_{x\in X_i}$变为$\sum_{j=1}^{m_i}$，则求和项中的$\boldsymbol{x}$变为$\boldsymbol{x}_j$，

而且注意到所有示例的均值向量$\boldsymbol{\mu}$为常量，第i类示例的均值向量$\boldsymbol{\mu}_i$相对于$\sum_{j=1}^{m_i}$来说也是常量，因此上式可继续变形整理为

$$
\begin{aligned}
\mathbf{S}_b &= \sum_{i=1}^{N} \sum_{\boldsymbol{x} \in X_i} \left(-\boldsymbol{x}(\boldsymbol{\mu}^\top - \boldsymbol{\mu}_i^\top) - (\boldsymbol{\mu} - \boldsymbol{\mu}_i)\boldsymbol{x}^\top + (\boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top) \right) \\
&= \sum_{i=1}^{N} \sum_{j=1}^{m_i} \left(-\boldsymbol{x}_j(\boldsymbol{\mu}^\top - \boldsymbol{\mu}_i^\top) - (\boldsymbol{\mu} - \boldsymbol{\mu}_i)\boldsymbol{x}_j^\top + (\boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top) \right) \\
&= \sum_{i=1}^{N} \left(-(\sum_{j=1}^{m_i} \boldsymbol{x}_j)(\boldsymbol{\mu}^\top - \boldsymbol{\mu}_i^\top) - (\boldsymbol{\mu} - \boldsymbol{\mu}_i)(\sum_{j=1}^{m_i} \boldsymbol{x}_j^\top) + (\boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top)(\sum_{j=1}^{m_i} 1) \right) \\
&= \sum_{i=1}^{N} \left(-m_i\boldsymbol{\mu}_i(\boldsymbol{\mu}^\top - \boldsymbol{\mu}_i^\top) - (\boldsymbol{\mu} - \boldsymbol{\mu}_i)(m_i\boldsymbol{\mu}_i^\top) + m_i(\boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top) \right) \\
&= \sum_{i=1}^{N} m_i \left(-\boldsymbol{\mu}_i\boldsymbol{\mu}^\top + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top - \boldsymbol{\mu}\boldsymbol{\mu}_i^\top + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top - \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top \right) \\
&= \sum_{i=1}^{N} m_i \left(-\boldsymbol{\mu}_i\boldsymbol{\mu}^\top + \boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top - \boldsymbol{\mu}\boldsymbol{\mu}_i^\top + \boldsymbol{\mu}\boldsymbol{\mu}^\top \right) \\
&= \sum_{i=1}^{N} m_i \left(-\boldsymbol{\mu}_i(\boldsymbol{\mu}^\top - \boldsymbol{\mu}_i^\top) - \boldsymbol{\mu}(\boldsymbol{\mu}_i^\top - \boldsymbol{\mu}^\top) \right) \\
&= \sum_{i=1}^{N} m_i \left(\boldsymbol{\mu}_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top - \boldsymbol{\mu}(\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top \right) \\
&= \sum_{i=1}^{N} m_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^\top
\end{aligned}
$$

7、式(3.44)的解释

由式(3.41)和式(3.43)可知$\mathbf{S}_b$和$\mathbf{S}_w$都是$d \times d$的矩阵，而$\mathbf{W} \in \mathbb{R}^{d \times (N-1)}$，设$\mathbf{W}$的第$i$列为$\boldsymbol{w}_i \in \mathbb{R}^{d \times 1}$，即$\mathbf{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{N-1})$，因此

$$
\operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_b \mathbf{W}\right) = \sum_{i=1}^{N-1} \boldsymbol{w}_i^\top \mathbf{S}_b \boldsymbol{w}_i, \ \operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_w \mathbf{W}\right) = \sum_{i=1}^{N-1} \boldsymbol{w}_i^\top \mathbf{S}_w \boldsymbol{w}_i
$$

这里其实就是式(3.35)的推广。对比发现，式(3.35)中只有 1 个$\boldsymbol{w}$，因为它是一个二分类问题；而式(3.44)有$N-1$个$\boldsymbol{w}_i, i = 1, 2, \ldots, N-1$，因为它是一个$N$分类问题。式(3.35)分子只有 1 个$\boldsymbol{w}^\top \mathbf{S}_b \boldsymbol{w}$，而式(3.44)有$N-1$个$\boldsymbol{w}_i^\top \mathbf{S}_b \boldsymbol{w}_i, i = 1, 2, \ldots, N-1$。

8、式(3.45)的推导[?]

类似于式(3.36)，令$\operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_w \mathbf{W}\right) = 1$，将式(3.44)变形为

$$
\begin{aligned}
&\min_{\mathbf{W}} \ -\operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_b \mathbf{W}\right) \\
&\text{s.t.} \ \operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_w \mathbf{W}\right) = 1
\end{aligned}
$$

应用拉格朗日乘子法，拉格朗日函数为

$$L(\mathbf{W},\lambda)=-\mathrm{tr}\left(\mathbf{W}^{\top}\mathbf{S}_b\mathbf{W}\right)+\lambda\left(\mathrm{tr}\left(\mathbf{W}^{\top}\mathbf{S}_w\mathbf{W}\right)-1\right)$$

对$\mathbf{W}$求导，得

$$\begin{aligned}\frac{\partial L(\mathbf{W},\lambda)}{\partial\mathbf{W}}&=-\frac{\partial\mathrm{tr}\left(\mathbf{W}^{\top}\mathbf{S}_b\mathbf{W}\right)}{\partial\mathbf{W}}+\lambda\frac{\partial\mathrm{tr}\left(\mathbf{W}^{\top}\mathbf{S}_w\mathbf{W}\right)}{\partial\mathbf{W}}\\&=-\left(\mathbf{S}_b\mathbf{W}+\mathbf{S}_b^{\top}\mathbf{W}\right)+\lambda\left(\mathbf{S}_w\mathbf{W}+\mathbf{S}_w^{\top}\mathbf{W}\right)\\&=-2\mathbf{S}_b\mathbf{W}+2\lambda\mathbf{S}_w\mathbf{W}\end{aligned}$$

有关矩阵的迹求导可以搜索《The Matrix Cookbook(Version: November 15, 2012)》（式 108）：
http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf
令导数等于零，即得式(3.45)

$$-2\mathbf{S}_b\mathbf{W}+2\lambda\mathbf{S}_w\mathbf{W}=0\Rightarrow\mathbf{S}_b\mathbf{W}=\lambda\mathbf{S}_w\mathbf{W}$$

两边同时左乘$\mathbf{S}_w^{-1}$，得

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

这就是一个广义特征值求解问题【？】。

3.5 多分类学习

1、图 3.4 的解释

左图为 OvO，共产生$\frac{4\times(4-1)}{2}=6$个二分类任务，其中 C1 得 2 票，C2 得 1 票，C3 得 3 票，C4 得 0 票，因此预测为 C3（把预测得最多的类别作为最终分类结果）。

右图为 OvR，共产生 4 个二分类任务，其中 C3 预测为正类，因此预测为 C3。

2、图 3.5 的解释

先解释图(a)，在五个二分类器中，C1 类样本分别被指定为反类、正类、反类、正类、正类，C2 类样本分别被指定为正类、反类、反类、正类、反类，C3 类样本分别被指定为反类、正类、正类、反类、正类，C4 类样本分别被指定为反类、反类、正类、正类、反类，而测试示例被五个二分类器分别预测为反类、反类、正类、反类、正类，与 C1~C4 类样本在五个二分类器中的角色谁最像呢？直觉上也应该是 C3 类，因为只有f_2对测试示例的预测输出（反类）与 C3 类样本在f_2训练时充当的角色（正类）不同，其余四个分类器对测试示例的输出均与 C3 类样本在相应分类器训练时充当的角色相同。

图(b)与图(a)的唯一不同是有停用类，即三元码。

上面解释图(a)中测试样本的类别时判断为C3类，实际用的是海明距离。所谓海明距离，即两个码对应位置不相同的个数，如图(a)中第 1 行为{-1, +1, -1, +1, +1}，测试示例一行为{-1, -1, +1, -1, +1}，其中第 2 个、第 3 个、第 4 个元素不相同，所以它们的海明距离为 3；同理第 2 行与测试示例一行除第 2 个元素相同外其余均不同，所以它们的海明距离为 4；第 3 行前面已经说过只有第 2 个元素不同，因此海明距离为 1；第 4 行与测试示例一行第 4 个、第 5 个元素不相同，所以它们的海明距离是 2；即第 3 行与测试示例一行的海明距离最小，因此预测结果将是 C3 类；在图(b)中新增了停用类，此时海明距离对应加 0.5，其余与图(a)相同。欧氏距离即对应位置相减再平方，求和后再开方就是了，即差向量的 2 范数。

3、ECOC 编码长度的解释

西瓜书第 65 页倒数第 6 行中提到：“对有限的类别数，可能的组合数是有限的，码长超

过一定范围后就失去了意义”，比如对于图 3.5(a)，最大的组合数为$2^4-2=14$种，其中2^4指每种类均有正类和反类两种可能，减 2 指 C1~C4 不能全是正类或反类。

3.6 类别不平衡问题

对于类别平衡问题，2.3.1 节的性能度量“精度”并不能满足该特殊任务的需求，例如本节第一段的举例：998 个反例和 2 个正例，若返回一个永远将新样本预测为反例的学习器则能达到 99.8%的精度；此时 2.3.2 节的查准率、查全率，尤其是 F1 将闪亮登场。

本节第二段提到了 OvR、OvM 策略产生的二分类任务可能出现类别不平衡现象，而左侧的边注中特别提到“对 OvR、OvM 来说，由于对每个类进行了相同的处理，其拆解出的二分类任务中类别不平衡的影响会相互抵消，因此通常不需专门处理”，这也是习题 3.9 的问题，谈一点个人理解：比如对于 0~9 数字的 10 分类问题，各类样本占总样本数的十分之一，针对每类样本实施 OvR 时，它们面临的都是 1:9 的不平衡分类，轮流一圈后大家遭受的不平衡待遇实际是一样的。

1、式(3.48)的推导

式(3.48)实际上是想将式(3.47)变为式(3.46)的样子，即大于号的右侧变为 1，因此实际很简单，只需对式(3.47)两边同乘以m^-/m^+即可，此即式(3.48)。

3.7 本章小节

本章是全书第 1 部分（介绍机器学习基础知识）的最后一章，章节标题为“线性模型”；

抛开 3.1 节不谈，3.2 节、3.3 节、3.4 节分别介绍了一个线性模型，分别是线性回归、对率回归、线性判别分析，其中 3.2 节是回归任务，3.3 节和 3.4 节是分类任务，分类任务是机器学中更常见的任务；但 3.3 节、3.4 节的对率回归和线性判别分析是针对二分类任务的，3.5 节则介绍了如何利用二分类学习器解决多分类任务；

前面也曾提到 3.4 节后半部分介绍了多分类线性判别分析与 3.5 节多分类学习顺序有些颠倒，但也实属无奈之举，因为 3.5 节介绍的多分类学习是**问题转换型解决方法**（即将多分类任务转换为二分类任务，如 OvR、OvO、MvM、ECOC），而 3.4 节后半部分介绍的多分类线性判别分析是**算法适应型解决方法**（即对现有某特定二分类算法进行改进以适应多分类任务，如多分类线性判别分析以及西瓜书 P68 第二段最后提到的多类支持向量机[Crammer and Singer, 2001; Lee et al., 2004]等；另外，第 4 章介绍的决策树和第 7 章介绍的朴素贝叶斯分类器本身可以很好地解决多分类问题）；也就是说从 3.4 节后半部分就开始介绍多分类任务了，因此实际可将 3.5 节开头多分类任务的开篇介绍提前一些~

多分类任务一般可以通过 3.5 节的转换方法解决，一般来说算法效率和精度都还很好，例如著明的 LIBSVM 中实现多分类时并没有直接实现多类支持向量机算法，而是使用 OvO 将多分类转换为多个二分类后再解决的。[Hsu, Chih-Wei, and Chih-Jen Lin. "A comparison of methods for multiclass support vector machines." IEEE transactions on Neural Networks 13.2 (2002): 415-425.]则给出实验结论，“OvO 更适合实际中使用”。总之，直接解决多分类任务还是一个相对比较困难的问题，但很多场景并不需要直接解决，只要解决就可以了~

3.6 节的介绍的类别不平衡问题确实现实生活中很常见的实际问题，而且西瓜书作者在这方面也有一些该领域代表性的研究成果；注意该节提到的 SMOTE 算法，比较常用~

梳理一下本章结构：3.1 节引言，3.2 节、3.3 节、3.4 节三种线性模型，3.5 节、3.6 节两类特别的分类问题。