机器学习（西瓜书）

注　解

（第 13 章　半监督学习）

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 13 章目录

第 13 章 半监督学习

13.1 未标记样本

本节内容简单易懂，几乎不需要什么注解。两张插图可谓本节亮点：图 13.1 直观地说明了使用未标记样本后带来的好处；图 13.2 对比了主动学习、(纯)半监督学习和直推学习，尤其是巧妙地将主动学习的概念融入进来。

直推学习是综合运用手头上已有的少量有标记样本和大量未标记样本，对**这些大量未标记样本**预测其标记；而(纯)半监督学习是综合运用手头上已有的少量有标记样本和大量未标记样本，对**新的未标记样本**预测其标记。

对于直推学习，当然可以仅利用有标记样本训练一个学习器，再对未标记样本进行预测，此即传统的监督学习；对于(纯)半监督学习，当然也可以舍弃大量未标记样本，仅利用有标记样本训练一个学习器，再对新的未标记样本进行预测。但图 13.1 直观地说明了使用未标记样本后带来的好处，然而利用了未标记样本后是否真的会如图 13.1 所示带来预期的好处呢？此即 13.7 节阅读材料中提到的安全半监督学习。

接下来在 13.2 节、13.3 节、13.4 节、13.5 节介绍的四种半监督学习方法，都可以应用于直推学习，但若要应用于(纯)半监督学习，则要有额外的考虑，尤其是 13.4 节介绍的图半监督学习，因为该节最后一段也明确提到“构图过程仅能考虑训练样本集，难以判知新样本在图中的位置，因此，在接收到新样本时，或是将其加入原数据集对图进行重构并重新进行标记传播，或是需引入额外的预测机制”。

13.2 生成式方法

本节与 9.4.3 节的高斯混合聚类密切相关，有关 9.4.3 节的公式推导参见附录，建议将高斯混合聚类的内容理解之后再学习本节算法。

1、式(13.1)的解释

该式即为 9.4.3 节的式(9.29)，式(9.29)中的k个混合成分对应于此处的N个可能的类别。

2、式(13.2)的推导

首先，该式的变量$\Theta \in \{1, 2, \ldots, N\}$即为 9.4.3 节的式(9.30)中的$z_j \in \{1, 2, \ldots, k\}$。

从公式第 1 行到第 2 行是做了边际化（marginalization）；具体来说第 2 行比第 1 行多了变量Θ，为了消掉Θ对其进行求和（若是连续变量则为积分）$\sum_{i=1}^{N}$；

从公式第 2 行到第 3 行推导如下：

$$
\begin{aligned}
p(y=j, \Theta=i \mid \boldsymbol{x}) &= \frac{p(y=j, \Theta=i, \boldsymbol{x})}{p(\boldsymbol{x})} \\
&= \frac{p(y=j, \Theta=i, \boldsymbol{x})}{p(\Theta=i, \boldsymbol{x})} \cdot \frac{p(\Theta=i, \boldsymbol{x})}{p(\boldsymbol{x})} \\
&= p(y=j \mid \Theta=i, \boldsymbol{x}) \cdot p(\Theta=i \mid \boldsymbol{x})
\end{aligned}
$$

$p(y=j \mid \boldsymbol{x})$表示$\boldsymbol{x}$的类别$y$为第$j$个类别标记的后验概率（注意条件是已知$\boldsymbol{x}$）；

$p(y=j, \Theta=i \mid \boldsymbol{x})$表示$\boldsymbol{x}$的类别$y$为第$j$个类别标记且由第$i$个高斯混合成分生成的后验概率（注意条件是已知$\boldsymbol{x}$）；

$p(y=j \mid \Theta=i, \boldsymbol{x})$表示第$i$个高斯混合成分生成的$\boldsymbol{x}$其类别$y$为第$j$个类别标记的概率（注意条件是已知$\Theta$和$\boldsymbol{x}$，这里修改了西瓜书式(13.3)下方对$p(y=j \mid \Theta=i, \boldsymbol{x})$的表述）；

$p(\Theta=i \mid \boldsymbol{x})$表示$\boldsymbol{x}$由第$i$个高斯混合成分生成的后验概率（注意条件是已知$\boldsymbol{x}$）。

西瓜书第 296 页第 2 行提到“假设样本由高斯混合模型生成，且每个类别对应一个高斯混合成分”，也就是说，如果已知$\boldsymbol{x}$是由哪个高斯混合成分生成的，也就知道了其类别。而$p(y=j \mid \Theta=i, \boldsymbol{x})$表示已知$\Theta$和$\boldsymbol{x}$的条件概率（已知$\Theta$就足够，不需$\boldsymbol{x}$的信息），因此

$$p(y=j \mid \Theta=i, \boldsymbol{x})=\begin{cases}1, & i=j \\ 0, & i \neq j\end{cases}$$

3、式(13.3)的解释

该式即为 9.4.3 节的式(9.30)，具体推导参见附录有关式(9.30)的解释。

4、式(13.4)的推导

由式(13.2)对概率$p(y=j \mid \Theta=i, \boldsymbol{x})$的分析，式中第 1 项中的$p(y_j \mid \Theta=i, \boldsymbol{x}_j)$为

$$p(y_j \mid \Theta=i, \boldsymbol{x}_j)=\begin{cases}1, & y_j=i \\ 0, & y_j \neq i\end{cases}$$

该式第 1 项针对有标记样本$(\boldsymbol{x}_j, y_j) \in D_l$来说的，因为有标记样本的类别是确定的，因此在计算它的对数似然时，它只可能来自N个高斯混合成分中的一个（西瓜书第 296 页第 2 行提到“假设样本由高斯混合模型生成，且每个类别对应一个高斯混合成分”），所以计算第 1 项计算有标记样本似然时乘以了$p(y_j \mid \Theta=i, \boldsymbol{x}_j)$；

该式第 2 项针对未标记样本$\boldsymbol{x}_j \in D_u$来说的，因为未标记样本的类别不确定，即它可能来自N个高斯混合成分中的任何一个，所以第 1 项使用了式(13.1)。

5、式(13.5)的解释

该式与式(13.3)相同，即后验概率。

可通过有标记数据对模型参数$(\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$进行初始化，具体来说：

$$\alpha_i=\frac{l_i}{|D_l|}, \text{where } |D_l|=\sum_{i=1}^{N} l_i$$

$$\boldsymbol{\mu}_i=\frac{1}{l_i} \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j=i} \boldsymbol{x}_j$$

$$\boldsymbol{\Sigma}_i=\frac{1}{l_i} \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j=i} (\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^{\top}$$

其中l_i表示第i类样本的有标记样本数目，$|D_l|$为有标记样本集样本总数，$\wedge$为“逻辑与”。

6、式(13.6)的解释

类似于式(9.34)，该式由$\frac{\partial LL(D_l \cup D_u)}{\partial \boldsymbol{\mu}_i}=0$而得，将式(13.4)的两项分别记为：

$$LL(D_l) = \sum_{(\boldsymbol{x}_j, y_j) \in D_l} \ln \left(\sum_{s=1}^{N} \alpha_s \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) \cdot p(y_j \mid \Theta = s, \boldsymbol{x}_j) \right)$$

$$LL(D_u) = \sum_{\boldsymbol{x}_j \in D_u} \ln \left(\sum_{s=1}^{N} \alpha_s \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) \right)$$

对于式(13.4)中的第 2 项$LL(D_u)$，求导结果与式(9.33)的推导过程一样：

$$\begin{aligned} \frac{\partial LL(D_u)}{\partial \boldsymbol{\mu}_i} &= \sum_{\boldsymbol{x}_j \in D_u} \frac{\alpha_i}{\sum_{s=1}^{N} \alpha_s \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)} \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \cdot \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) \\ &= \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) \end{aligned}$$

对于式(13.4)中的第 1 项$LL(D_l)$，由于$p(y_j \mid \Theta = s, \boldsymbol{x}_j)$的取值非 1 即 0（详见式(13.2)和式(13.4)的分析），因此

$$LL(D_l) = \sum_{(\boldsymbol{x}_j, y_j) \in D_l} \ln \left(\alpha_{y_j} \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_{y_j}, \boldsymbol{\Sigma}_{y_j}) \right)$$

若求$LL(D_l)$对$\boldsymbol{\mu}_i$的偏导，则$LL(D_l)$求和号中只有$y_j = i$的项能留下来，即

$$\begin{aligned} \frac{\partial LL(D_l)}{\partial \boldsymbol{\mu}_i} &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{\partial \ln \left(\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right)}{\partial \boldsymbol{\mu}_i} \\ &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{1}{p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \cdot \frac{\partial p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\partial \boldsymbol{\mu}_i} \\ &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{1}{p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \cdot \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) \\ &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) \end{aligned}$$

综合两项结果，则$\frac{\partial LL(D_l \cup D_u)}{\partial \boldsymbol{\mu}_i}$为

$$\begin{aligned} \frac{\partial LL(D_l \cup D_u)}{\partial \boldsymbol{\mu}_i} &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) \\ &= \boldsymbol{\Sigma}_i^{-1} \left(\sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot (\boldsymbol{x}_j - \boldsymbol{\mu}_i) \right) \\ &= \boldsymbol{\Sigma}_i^{-1} \left(\sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{x}_j + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{x}_j - \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{\mu}_i - \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{\mu}_i \right) \end{aligned}$$

令$\frac{\partial LL(D_l \cup D_u)}{\partial \boldsymbol{\mu}_i} = 0$，两边同时左乘$\boldsymbol{\Sigma}_i$可将$\boldsymbol{\Sigma}_i^{-1}$消掉，移项即得

$$\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{\mu}_i + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{\mu}_i = \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot \boldsymbol{x}_j + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \boldsymbol{x}_j$$

上式中，$\boldsymbol{\mu}_i$可以作为常量提到求和号外面，而$\sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} 1 = l_i$，即第$i$类样本的有标记样本数目，因此

$$\left(\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}1\right)\boldsymbol{\mu}_i=\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\boldsymbol{x}_j+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\boldsymbol{x}_j$$

即得式(13.6)：

$$\boldsymbol{\mu}_i=\frac{1}{\sum\limits_{\boldsymbol{x}_j\in D_u}\gamma_{ji}+l_i}\left(\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\boldsymbol{x}_j+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\boldsymbol{x}_j\right)$$

7、式(13.7)的解释

类似于式(9.35)，该式由$\frac{\partial LL(D_l\cup D_u)}{\partial\boldsymbol{\Sigma}_i}=0$而得。对于式(13.4)中的第 2 项$LL(D_u)$，求导结果与式(9.35)的推导过程一样：

$$\frac{\partial LL(D_u)}{\partial\boldsymbol{\Sigma}_i}=\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top-\boldsymbol{I}\right)\cdot\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}$$

对于式(13.4)中的第 1 项$LL(D_l)$，类似于刚才式(13.6)的推导过程：

$$\begin{aligned}\frac{\partial LL(D_l)}{\partial\boldsymbol{\Sigma}_i}&=\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\frac{\partial\ln\left(\alpha_i\cdot p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\right)}{\partial\boldsymbol{\Sigma}_i}\\&=\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\frac{1}{p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}\cdot\frac{\partial p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{\partial\boldsymbol{\Sigma}_i}\\&=\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\frac{1}{p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}\cdot p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\cdot\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top-\boldsymbol{I}\right)\cdot\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\\&=\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top-\boldsymbol{I}\right)\cdot\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\end{aligned}$$

综合两项结果，则$\frac{\partial LL(D_l\cup D_u)}{\partial\boldsymbol{\Sigma}_i}$为

$$\begin{aligned}\frac{\partial LL(D_l\cup D_u)}{\partial\boldsymbol{\mu}_i}&=\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top-\boldsymbol{I}\right)\cdot\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\\&\quad+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top-\boldsymbol{I}\right)\cdot\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\\&=\left(\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top-\boldsymbol{I}\right)\right.\\&\quad\left.+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top-\boldsymbol{I}\right)\right)\cdot\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\end{aligned}$$

令$\frac{\partial LL(D_l\cup D_u)}{\partial\boldsymbol{\Sigma}_i}=0$，两边同时右乘以$2\boldsymbol{\Sigma}_i$可将$\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}$消掉，移项即得

$$\begin{aligned}&\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top\\&=\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}\cdot\boldsymbol{I}+\sum_{(\boldsymbol{x}_j,y_j)\in D_l\wedge y_j=i}\boldsymbol{I}\\&=\left(\sum_{\boldsymbol{x}_j\in D_u}\gamma_{ji}+l_i\right)\boldsymbol{I}\end{aligned}$$

两边同时左乘以$\mathbf{\Sigma}_i$，上式变为

$$\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top = \left(\sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i \right) \mathbf{\Sigma}_i$$

即得式(13.7)：

$$\begin{aligned} \mathbf{\Sigma}_i = \frac{1}{\sum\limits_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + l_i} \Bigg(& \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} \cdot (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \\ & + \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \Bigg) \end{aligned}$$

8、式(13.8)的解释

类似于式(9.36)，写出$LL(D_l \cup D_u)$的拉格朗日形式

$$\begin{aligned} \mathcal{L}(D_l \cup D_u, \lambda) &= LL(D_l \cup D_u) + \lambda \left(\sum_{s=1}^{N} \alpha_s - 1 \right) \\ &= LL(D_l) + LL(D_u) + \lambda \left(\sum_{s=1}^{N} \alpha_s - 1 \right) \end{aligned}$$

类似于式(9.37)，对α_i求偏导。对于$LL(D_u)$，求导结果与式(9.37)的推导过程一样：

$$\frac{\partial LL(D_u)}{\partial \alpha_i} = \sum_{\boldsymbol{x}_j \in D_u} \frac{1}{\sum_{s=1}^{N} \alpha_s \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_s, \mathbf{\Sigma}_s)} \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$$

对于$LL(D_u)$，类似于刚才式(13.6)和式(13.7)的推导过程：

$$\begin{aligned} \frac{\partial LL(D_l)}{\partial \alpha_i} &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{\partial \ln \left(\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \right)}{\partial \alpha_i} \\ &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{1}{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \mathbf{\Sigma}_i)} \cdot \frac{\partial \left(\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \right)}{\partial \alpha_i} \\ &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{1}{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \mathbf{\Sigma}_i)} \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \\ &= \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} \frac{1}{\alpha_i} = \frac{1}{\alpha_i} \cdot \sum_{(\boldsymbol{x}_j, y_j) \in D_l \wedge y_j = i} 1 = \frac{l_i}{\alpha_i} \end{aligned}$$

上式推导过程中，重点注意变量是α_i，$p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \mathbf{\Sigma}_i)$是常量；最后一行$\alpha_i$相对于求和变量为常量，因此作为公因子提到求和号外面；l_i为第i类样本的有标记样本数目。

综合两项结果，则$\frac{\partial \mathcal{L}(D_l \cup D_u, \lambda)}{\partial \alpha_i}$为

$$\frac{\partial \mathcal{L}(D_l \cup D_u, \lambda)}{\partial \alpha_i} = \frac{l_i}{\alpha_i} + \sum_{\boldsymbol{x}_j \in D_u} \frac{p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \mathbf{\Sigma}_i)}{\sum_{s=1}^{N} \alpha_s \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_s, \mathbf{\Sigma}_s)} + \lambda$$

令$\frac{\partial \mathcal{L}(D_l \cup D_u, \lambda)}{\partial \alpha_i} = 0$，并且两这边同乘以$\alpha_i$，得

$$\alpha_i \cdot \frac{l_i}{\alpha_i} + \sum_{\boldsymbol{x}_j \in D_u} \frac{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{s=1}^{N} \alpha_s \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)} + \lambda \cdot \alpha_i = 0$$

结合式(9.30)发现，求和号内即为后验概率γ_{ji}，即

$$l_i + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + \lambda \alpha_i = 0$$

对所有**混合成分**求和，得

$$\sum_{i=1}^{N} l_i + \sum_{i=1}^{N} \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + \sum_{i=1}^{N} \lambda \alpha_i = 0$$

注意$\sum_{i=1}^{N} \alpha_i = 1$，因此$\sum_{i=1}^{N} \lambda \alpha_i = \lambda \sum_{i=1}^{N} \alpha_i = \lambda$，根据式(9.30)中$\gamma_{ji}$表达式可知

$$\begin{aligned}
\sum_{i=1}^{N} \gamma_{ji} &= \sum_{i=1}^{N} \frac{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{s=1}^{N} \alpha_s \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)} \\
&= \frac{\sum_{i=1}^{N} \alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{s=1}^{N} \alpha_s \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)} \\
&= 1
\end{aligned}$$

再结合加法满足交换律，所以

$$\sum_{i=1}^{N} \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} = \sum_{\boldsymbol{x}_j \in D_u} \sum_{i=1}^{N} \gamma_{ji} = \sum_{\boldsymbol{x}_j \in D_u} 1 = u$$

以上分析过程中，$\sum_{\boldsymbol{x}_j \in D_u}$形式与$\sum_{j=1}^{u}$等价，其中$u$为未标记样本集的样本个数；$\sum_{i=1}^{N} l_i = l$，其中$l$为有标记样本集的样本个数；将这些结果代入

$$\sum_{i=1}^{N} l_i + \sum_{i=1}^{N} \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + \sum_{i=1}^{N} \lambda \alpha_i = 0$$

即$l + u + \lambda = 0$；又$l + u = m$，其中m样本总个数，移项即得$\lambda = -m$。

将以上分析结果代入$\sum_{i=1}^{k} \sum_{j=1}^{m} \gamma_{ji} + \sum_{i=1}^{k} \lambda \alpha_i = 0$，移项即得$\lambda = -m$。将此结果代入

$$l_i + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} + \lambda \alpha_i = 0$$

即$l_i + \sum_{\boldsymbol{x}_j \in D_u} \gamma_{ji} - m\alpha_i = 0$，整理即得式(13.8)。

13.3 半监督 SVM

从本节名称“半监督 SVM”即可知道与第 6 章的 SVM 内容联系紧密。建议理解了 SVM 之后再学习本节算法，会发现实际很简单；否则会感觉无从下手，难以理解。

由本节开篇的两段介绍可知，S3VM 是 SVM 在半监督学习上的推广，是此类算法的总称而非某个具体的算法，其最著名的代表是 TSVM。

1、图 13.3 的解释

注意对比 S3VM 划分超平面穿过的区域与 SVM 划分超平面穿过的区域的差别，明显 S3VM 划分超平面周围样本较少，也就是“数据低密度区域”，即“低密度分隔”。

2、式(13.9)的解释

与第 6 章式(6.35)对比可以发现，二者几乎一模一样，区别在于此处利用了未标记样本，约束条件中未标记样本的标记使用的预测标记$\hat{y}_i$；当然目标函数将有标记样本和未标记样本的松弛变量的权重系数不同，分别用C_l和C_u表示。

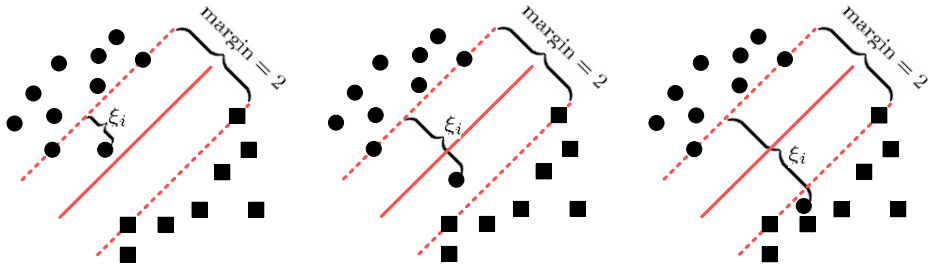
因此，欲理解本节内容应该先理解 SVM，否则会感觉无从下手，难以理解。

3、图 13.4 的解释

解释一下第 6 行：

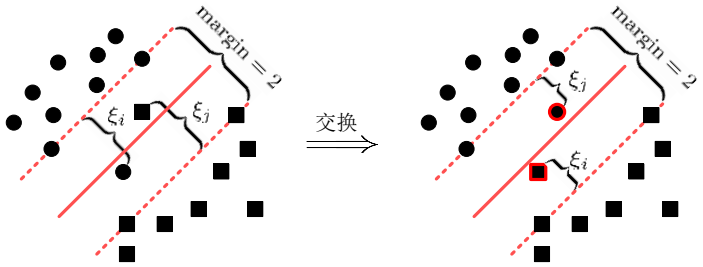
(1) $\hat{y}_i\hat{y}_j < 0$意味着未标记样本$\boldsymbol{x}_i, \boldsymbol{x}_j$在此次迭代中被指派的标记$\hat{y}_i, \hat{y}_j$相反（正例+1和反例−1各 1 个）；

(2) $\xi_i > 0$意味着未标记样本$\boldsymbol{x}_i$在此次迭代中为支持向量：(a)在间隔带内但仍与自己标记同侧$(0 < \xi_i < 1)$，(b)在间隔带内但与自己标记异侧$(1 < \xi_i < 2)$，(c)不在间隔带且与自己标记异侧$(\xi_i > 2)$；三种情况分别如下图(a)(b)(c)所示：



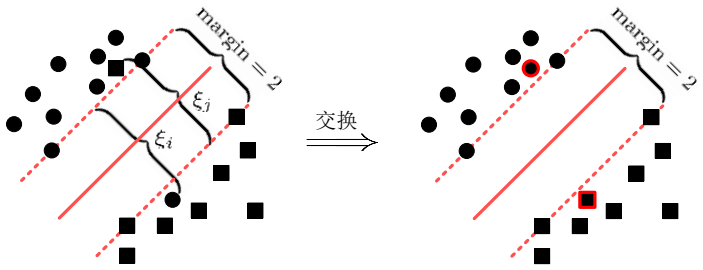
(a)$(0 < \xi_i < 1)$ (b)$(1 < \xi_i < 2)$ (c)$(\xi_i > 2)$

(3) $\xi_i + \xi_j > 2$分两种情况：(I)$(\xi_i > 1) \wedge (\xi_j > 1)$，表示都位于自己指派标记异侧，交换它们的标记后，二者就都位于自己新指派标记同侧了，如下图所示$(1 < \xi_i, \xi_j < 2)$：



(I.1)$(1 < \xi_i, \xi_j < 2)$

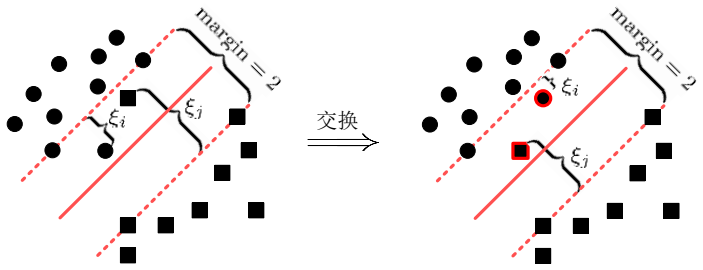
可以发现，当$1 < \xi_i, \xi_j < 2$时，交换之后虽然松弛变量仍然大于 0，但至少$\xi_i + \xi_j$比交换之前变小了；若进一步的，当$\xi_i, \xi_j > 2$时，则交换之后$\xi_i + \xi_j$将变为 0，如下图所示：



(I.2)$(\xi_i > 2) \wedge (\xi_j > 2)$

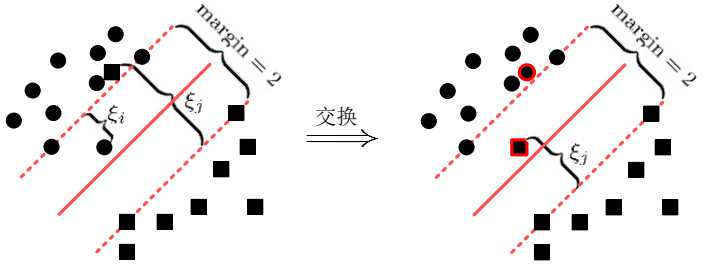
可以发现，交换之后两个样本均被分类正确，因此松弛变量均等于 0；至于ξ_i, ξ_j其中之一位于 1~2 之间，另一个大于 2，情况类似，不单列出分析。

(II)$(0 < \xi_i < 1) \wedge (\xi_j > 2 - \xi_i)$，表示有一个与自己标记同侧，有一个与自己标记异侧，此时可分两种情况：(II.1)$1 < \xi_j < 2$，表示样本与自己标记异侧，但仍在间隔带内：



(II.1)$(\xi_i + \xi_j > 2) \wedge (0 < \xi_i < 1) \wedge (1 < \xi_j < 2)$

可以发现，此时两个样本位置超平面同一侧，交换标记之后似乎没发生什么变化，但是仔细观察会发现交换之后$\xi_i + \xi_j$ 比交换之前变小了；(II.2)$\xi_j > 2$，表示样本在间隔带外：



(II.2)$(\xi_i + \xi_j > 2) \wedge (0 < \xi_i < 1) \wedge (\xi_j > 2)$

可以发现，交换之后其中之一被正确分类，$\xi_i + \xi_j$ 比交换之前也变小了。

综上所述，当$\xi_i + \xi_j > 2$时，交换指派标记$\hat{y}_i, \hat{y}_j$可以使$\xi_i + \xi_j$下降，也就是说分类结果会得到改善。

再解释一下第 11 行：逐步增长C_u，但不超过C_l，未标记样本的权重小于有标记样本。

4、式(13.10)的解释

将该式变形为$\frac{C_u^+}{C_u^-} = \frac{u_-}{u_+}$，即样本个数多的权重小，样本个数少的权重大，总体上保持二者的作用相同。

13.4 图半监督学习

本节共讲了两种方法，其中式(13.11)~式(13.17)讲述了一个针对二分类问题的标记传播方法，式(13.18)~式(13.21)讲述了一个针对多分类问题的标记传播方法，两种方法的原理均

为“相似的样本应具有相似的标记”，只是面向的问题不同，而且具体实现的方法也不同。

1、式(13.12)的推导

注意，该方法针对二分类问题的标记传播方法。我们希望能量函数$E(f)$越小越好，注意到式(13.11)的$0<(\mathbf{W})_{ij}\leqslant 1$，且样本$\boldsymbol{x}_i$和样本$\boldsymbol{x}_j$越相似(即$\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2$越小)则$(\mathbf{W})_{ij}$越大，因此要求式(13.12)中的$(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j))^2$相应地越小越好（即“相似的样本应具有相似的标记”），如此才能达到能量函数$E(f)$越小的目的。

首先对式(13.12)的第 1 行式子进行展开整理：

$$\begin{aligned}E(f)&=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}\left(f(\boldsymbol{x}_i)-f(\boldsymbol{x}_j)\right)^2\\&=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}\left(f^2(\boldsymbol{x}_i)-2f(\boldsymbol{x}_i)f(\boldsymbol{x}_j)+f^2(\boldsymbol{x}_j)\right)\\&=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f^2(\boldsymbol{x}_i)+\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f^2(\boldsymbol{x}_j)-\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f(\boldsymbol{x}_i)f(\boldsymbol{x}_j)\end{aligned}$$

然后证明$\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f^2(\boldsymbol{x}_i)=\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f^2(\boldsymbol{x}_j)$，并变形：

$$\begin{aligned}\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f^2(\boldsymbol{x}_j)&=\sum_{j=1}^{m}\sum_{i=1}^{m}(\mathbf{W})_{ji}f^2(\boldsymbol{x}_i)=\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f^2(\boldsymbol{x}_i)\\&=\sum_{i=1}^{m}f^2(\boldsymbol{x}_i)\sum_{j=1}^{m}(\mathbf{W})_{ij}\end{aligned}$$

其中，第 1 个等号是把变量i,j分别用j,i替代（统一替换公式中的符号并不影响公式本身）；第 2 个等号是由于$\mathbf{W}$是对称矩阵（即$(\mathbf{W})_{ij}=\mathbf{W})_{ji}$），并交换了求和号次序（类似于多重积分中交换积分号次序），到此完成了该步骤的证明；第 3 个等号是由于$f^2(\boldsymbol{x}_i)$与求和变量j无关，因此拿到了该求和号外面（与求和变量无关的项相对于该求和变量相当于常数），该步骤的变形主要是为了得到d_i。

令$d_i=\sum_{j=1}^{m}(\mathbf{W})_{ij}$（既是$\mathbf{W}$第$i$行元素之和，实际亦是第$i$列元素之和，因为由于$\mathbf{W}$是对称矩阵，即$(\mathbf{W})_{ij}=\mathbf{W})_{ji}$，因此$d_i=\sum_{j=1}^{m}(\mathbf{W})_{ji}$，即第$i$列元素之和），则

$$E(f)=\sum_{i=1}^{m}d_if^2(\boldsymbol{x}_i)-\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f(\boldsymbol{x}_i)f(\boldsymbol{x}_j)$$

即式(13.12)的第 3 行，其中第一项$\sum_{i=1}^{m}d_if^2(\boldsymbol{x}_i)$可以写为如下矩阵形式：

$$\begin{aligned}\sum_{i=1}^{m}d_if^2(\boldsymbol{x}_i)&=\begin{bmatrix}f(\boldsymbol{x}_1) & f(\boldsymbol{x}_2) & \cdots & f(\boldsymbol{x}_m)\end{bmatrix}\begin{bmatrix}d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_m\end{bmatrix}\begin{bmatrix}f(\boldsymbol{x}_1)\\ f(\boldsymbol{x}_2)\\ \vdots\\ f(\boldsymbol{x}_m)\end{bmatrix}\\&=\boldsymbol{f}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{f}\end{aligned}$$

第二项$\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f(\boldsymbol{x}_i)f(\boldsymbol{x}_j)$也可以写为如下矩阵形式：

$$\sum_{i=1}^{m}\sum_{j=1}^{m}(\mathbf{W})_{ij}f(\boldsymbol{x}_i)f(\boldsymbol{x}_j)$$

$$=\begin{bmatrix}f(\boldsymbol{x}_1) & f(\boldsymbol{x}_2) & \cdots & f(\boldsymbol{x}_m)\end{bmatrix}\begin{bmatrix}(\mathbf{W})_{11} & (\mathbf{W})_{12} & \cdots & (\mathbf{W})_{1m}\\(\mathbf{W})_{21} & (\mathbf{W})_{22} & \cdots & (\mathbf{W})_{2m}\\\vdots & \vdots & \ddots & \vdots\\(\mathbf{W})_{m1} & (\mathbf{W})_{m2} & \cdots & (\mathbf{W})_{mm}\end{bmatrix}\begin{bmatrix}f(\boldsymbol{x}_1)\\f(\boldsymbol{x}_2)\\\vdots\\f(\boldsymbol{x}_m)\end{bmatrix}$$

$$=\boldsymbol{f}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{f}$$

所以$E(f)=\boldsymbol{f}^{\mathrm{T}}\boldsymbol{D}-\boldsymbol{f}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{f}=\boldsymbol{f}^{\mathrm{T}}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{f}$，即式(13.12)。

2、式(13.13)的推导

本式就是将式(13.12)用分块矩阵形式表达而已，拆分为标记样本和未标记样本两部分。

另外解释一下该式之前一段话中第一句的含义：“具有最小能量的函数f在有标记样本上满足$f(\boldsymbol{x}_i)=y_i\ (i=1,2,\ldots,l)$，在未标记样本上满足$\boldsymbol{\Delta f}=\mathbf{0}$”，前半句是很容易理解的，有标记样本上满足$f(\boldsymbol{x}_i)=y_i\ (i=1,2,\ldots,l)$，这时未标记样本的$f(\boldsymbol{x}_i)$是待求变量且应该使$E(f)$最小，因此应将式(13.12)对未标记样本的$f(\boldsymbol{x}_i)$求导并令导数等于 0 即可，此即表达式$\boldsymbol{\Delta f}=\mathbf{0}$，此处可以查看该算法的原始文献。

3、式(13.14)的推导

将式(13.13)根据矩阵运算规则进行变形：

$$\begin{aligned}E(f)&=\begin{bmatrix}\boldsymbol{f}_l^{\mathrm{T}} & \boldsymbol{f}_u^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\boldsymbol{D}_{ll}-\boldsymbol{W}_{ll} & -\boldsymbol{W}_{lu}\\-\boldsymbol{W}_{ul} & \boldsymbol{D}_{uu}-\boldsymbol{W}_{uu}\end{bmatrix}\begin{bmatrix}\boldsymbol{f}_l\\\boldsymbol{f}_u\end{bmatrix}\\&=\begin{bmatrix}\boldsymbol{f}_l^{\mathrm{T}}(\boldsymbol{D}_{ll}-\boldsymbol{W}_{ll})-\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul} & -\boldsymbol{f}_l^{\mathrm{T}}\boldsymbol{W}_{lu}+\boldsymbol{f}_u^{\mathrm{T}}(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu})\end{bmatrix}\begin{bmatrix}\boldsymbol{f}_l\\\boldsymbol{f}_u\end{bmatrix}\\&=(\boldsymbol{f}_l^{\mathrm{T}}(\boldsymbol{D}_{ll}-\boldsymbol{W}_{ll})-\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul})\boldsymbol{f}_l+(-\boldsymbol{f}_l^{\mathrm{T}}\boldsymbol{W}_{lu}+\boldsymbol{f}_u^{\mathrm{T}}(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu}))\boldsymbol{f}_u\\&=\boldsymbol{f}_l^{\mathrm{T}}(\boldsymbol{D}_{ll}-\boldsymbol{W}_{ll})\boldsymbol{f}_l-\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul}\boldsymbol{f}_l-\boldsymbol{f}_l^{\mathrm{T}}\boldsymbol{W}_{lu}\boldsymbol{f}_u+\boldsymbol{f}_u^{\mathrm{T}}(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu})\boldsymbol{f}_u\\&=\boldsymbol{f}_l^{\mathrm{T}}(\boldsymbol{D}_{ll}-\boldsymbol{W}_{ll})\boldsymbol{f}_l-2\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul}\boldsymbol{f}_l+\boldsymbol{f}_u^{\mathrm{T}}(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu})\boldsymbol{f}_u\end{aligned}$$

其中最后一步应用了$\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul}\boldsymbol{f}_l=(\boldsymbol{f}_l^{\mathrm{T}}\boldsymbol{W}_{lu}\boldsymbol{f}_u)^{\mathrm{T}}=\boldsymbol{f}_l^{\mathrm{T}}\boldsymbol{W}_{lu}\boldsymbol{f}_u$，这是因为对于一个数（即大小为$1\times 1$的矩阵）来说，其转置等于本身，此结论经常使用，应该熟悉。

4、式(13.15)的推导

首先，基于式(13.14)对$\boldsymbol{f}_u$求导：

$$\begin{aligned}\frac{\partial E(f)}{\partial \boldsymbol{f}_u}&=\frac{\partial \boldsymbol{f}_l^{\mathrm{T}}(\boldsymbol{D}_{ll}-\boldsymbol{W}_{ll})\boldsymbol{f}_l-2\boldsymbol{f}_u^{\mathrm{T}}\boldsymbol{W}_{ul}\boldsymbol{f}_l+\boldsymbol{f}_u^{\mathrm{T}}(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu})\boldsymbol{f}_u}{\partial \boldsymbol{f}_u}\\&=-2\boldsymbol{W}_{ul}\boldsymbol{f}_l+2(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu})\boldsymbol{f}_u\end{aligned}$$

上式求导过程中，第 1 项与$\boldsymbol{f}_u$无关，因此求导后等于 0。令$\frac{\partial E(f)}{\partial \boldsymbol{f}_u}=0$，即

$$-2\boldsymbol{W}_{ul}\boldsymbol{f}_l+2(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu})\boldsymbol{f}_u=0$$

移项

$$(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu})\boldsymbol{f}_u=\boldsymbol{W}_{ul}\boldsymbol{f}_l$$

两边同时左乘$(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu})$的逆矩阵，得式(13.15)：

$$\boldsymbol{f}_u=(\boldsymbol{D}_{uu}-\boldsymbol{W}_{uu})^{-1}\boldsymbol{W}_{ul}\boldsymbol{f}_l$$

注意式中各项的含义：

$\boldsymbol{f}_u$即函数f在未标记样本上的预测结果；

$\boldsymbol{D}_{uu}, \boldsymbol{W}_{uu}, \boldsymbol{W}_{ul}$均可以由式(13.11)得到；

$\boldsymbol{f}_l$即函数f在有标记样本上的预测结果（即已知标记，详见 P301 倒数第 3 行）；

也就是说可以根据式(13.15)根据D_l上的标记信息（即$\boldsymbol{f}_l$）求得未标记样本的标记（即$\boldsymbol{f}_u$），式(13.17)仅是式(13.15)的进一步变形化简，不再细述。

仔细回顾该方法，实际就是根据“相似的样本应具有相似的标记”的原则，构建了目标函数式(13.12)，求解式(13.12)得到了使用标记样本信息表示的未标记样本的预测标记。

5、式(13.18)的解释

矩阵$\mathbf{F}$是一个$m \times |\mathcal{Y}|$的矩阵，其中$m = l + u$。式(13.18)就是简单地将$\mathbf{F}$初始化为$\mathbf{Y}$，其中$\mathbf{Y}$的第i行表示第i个样本的类别；具体来说，对于前l个有标记样本来说，若第i个样本的类别为$j(1 \leq j \leq |\mathcal{Y}|)$，则$\mathbf{Y}$的第$i$行第$j$列即为 1，第$i$行其余元素为 0；对于后$u$个未标记样本来说，$\mathbf{Y}$统一为零。注意$|\mathcal{Y}|$表示集合$\mathcal{Y}$的势，即包含元素（类别）的个数。

6、式(13.19)的解释

该迭代计算式的目的是经迭代收敛至式(13.20)，而式(13.20)可由式(13.21)得到。

写出式(13.19)迭代公式当t为不同值时的情况：（注意，$\mathbf{F}(0) = \mathbf{Y}$）

$$t = 1: \ \mathbf{F}(1) = \alpha\mathbf{S}\mathbf{F}(0) + (1-\alpha)\mathbf{Y} = \alpha\mathbf{S}\mathbf{Y} + (1-\alpha)\mathbf{Y}$$

$$\begin{aligned} t = 2: \ \mathbf{F}(2) &= \alpha\mathbf{S}\mathbf{F}(1) + (1-\alpha)\mathbf{Y} = \alpha\mathbf{S}(\alpha\mathbf{S}\mathbf{Y} + (1-\alpha)\mathbf{Y}) + (1-\alpha)\mathbf{Y} \\ &= (\alpha\mathbf{S})^2\mathbf{Y} + (1-\alpha)\left(\sum_{i=0}^{1}(\alpha\mathbf{S})^i\right)\mathbf{Y} \end{aligned}$$

$$\begin{aligned} t = 3: \ \mathbf{F}(3) &= \alpha\mathbf{S}\mathbf{F}(2) + (1-\alpha)\mathbf{Y} \\ &= \alpha\mathbf{S}\left((\alpha\mathbf{S})^2\mathbf{Y} + (1-\alpha)\left(\sum_{i=0}^{1}(\alpha\mathbf{S})^i\right)\mathbf{Y}\right) + (1-\alpha)\mathbf{Y} \\ &= (\alpha\mathbf{S})^3\mathbf{Y} + (1-\alpha)\left(\sum_{i=0}^{2}(\alpha\mathbf{S})^i\right)\mathbf{Y} \end{aligned}$$

……

由以上可以发现规律（推导式(13.20)时要用到的）：

$$\mathbf{F}(t) = (\alpha\mathbf{S})^t\mathbf{Y} + (1-\alpha)\left(\sum_{i=0}^{t-1}(\alpha\mathbf{S})^i\right)\mathbf{Y}$$

注意：原文献“Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency[C]//Advances in neural information processing systems. 2004: 321-328.”中的公式(1)应该有笔误：

$$F(t) = (\alpha S)^{t-1}Y + (1-\alpha)\sum_{i=0}^{t-1}(\alpha S)^i Y. \tag{1}$$

7、式(13.20)的推导

根据上面刚刚得到的$\mathbf{F}(t)$表达式来推导式(13.20)的关键是证明$\lim_{t\to\infty}(\alpha\mathbf{S})^t = \mathbf{0}$，而

要证明此极限结果的关键是证明$\mathbf{S}$的特征值在区间$[-1,1]$内，以下开始证明：

由于$\mathbf{S}=\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}=\mathbf{D}^{\frac{1}{2}}\left(\mathbf{D}^{-1}\mathbf{W}\right)\mathbf{D}^{-\frac{1}{2}}$，即$\mathbf{S}$与$\mathbf{D}^{-1}\mathbf{W}$相似（对应到以下定义 7 的表示形式，$\mathbf{P}=\mathbf{D}^{-\frac{1}{2}}$）。

定义 7 设 A, B 都是 n 阶矩阵，若有可逆矩阵 P，使

$$P^{-1}AP=B,$$

则称 B 是 A 的相似矩阵，或说矩阵 A 与 B 相似. 对 A 进行运算 $P^{-1}AP$ 称为对 A 进行相似变换，可逆矩阵 P 称为把 A 变成 B 的相似变换矩阵.

定理 3 若 n 阶矩阵 A 与 B 相似，则 A 与 B 的特征多项式相同，从而 A 与 B 的特征值亦相同.

（摘自同济大学《线性代数（第 5 版）》第 121 页）

由以上定理 3，相似矩阵特征值相同，只需证明$\mathbf{D}^{-1}\mathbf{W}$的特征值在区间$[-1,1]$内即可。

注意$\mathbf{D}^{-1}$是一个对角矩阵，根据$\mathbf{D}$的定义，$\mathbf{D}^{-1}=\mathrm{diag}(d_1^{-1},d_2^{-1},\cdots,d_m^{-1})$，其中$d_i=\sum_{j=1}^{m}(\mathbf{W})_{ij}$，$m=l+u$，因此

$$\mathbf{D}^{-1}\mathbf{W}=\begin{bmatrix} d_1^{-1}\mathbf{W}_{1\cdot} \\ d_2^{-1}\mathbf{W}_{2\cdot} \\ \vdots \\ d_m^{-1}\mathbf{W}_{m\cdot} \end{bmatrix}$$

其中$\mathbf{W}_{i\cdot}$表示$\mathbf{W}$的第i行，且根据式(13.11)中$\mathbf{W}$的定义可知该矩阵是非负矩阵，即所有元素不小于零。注意到d_i为$\mathbf{W}$的第i行元素之和，因此矩阵$\mathbf{D}^{-1}\mathbf{W}$的每行元素之和均等于 1（即接下来要说的矩阵列和范数等于 1）。

定义 5.3.3 设 $A\in C^{n\times n}$，A 的 n 个特征值为 $\lambda_1,\lambda_2,\cdots,\lambda_n$，称 $\rho(A)=\max\{|\lambda_1|,|\lambda_2|,\cdots,|\lambda_n|\}$ 是 A 的谱半径.

定理 5.3.4 $A\in C^{n\times n}$，则

$$\rho(A)\leqslant\|A\|$$

其中 $\|A\|$ 是 A 的任何一种范数.

证明 设 λ 是 A 的任何一个特征值，即

$$Ax=\lambda x \qquad x\neq 0$$

故 $$\|\lambda x\|=|\lambda|\,\|x\|=\|Ax\|\leqslant\|A\|\,\|x\|$$

于是 $$|\lambda|\leqslant\|A\|$$

由于 λ 是 A 的任一个特征值，故

$$\rho(A)\leqslant\|A\|$$

（摘自史荣昌《矩阵分析（第 3 版）》第 186 页）

定理 5.3.2 设 $A=(a_{ij})_{m\times n}$，则

(1) $\|A\|_1=\max\limits_j\left(\sum\limits_{i=1}^{m}|a_{ij}|\right)\quad(j=1,2,\cdots,n)$

称 $\|A\|_1$ 是列和范数.

(2) $\|A\|_2=\max\limits_j(\lambda_j(A^HA))^{\frac{1}{2}}$，$\lambda_j(A^HA)$表示矩阵 A^HA 的第 j 个特征值. 称 $\|A\|_2$ 是谱范数. 即 $\|A\|_2$ 是 A 的最大正奇异值.

(3) $\|A\|_\infty=\max\limits_i\left(\sum\limits_{j=1}^{n}|a_{ij}|\right)\quad(i=1,2,\cdots,m)$

称 $\|A\|_\infty$ 是行和范数.

（摘自史荣昌《矩阵分析（第 3 版）》第 185 页）

当取列和范数时，谱半径$\rho(\mathbf{D}^{-1}\mathbf{W})\leqslant\|\mathbf{D}^{-1}\mathbf{W}\|_1=1$，由谱半径的定义可知$\mathbf{D}^{-1}\mathbf{W}$的特征值在区间$[-1,1]$内。又由于$\mathbf{S}$与$\mathbf{D}^{-1}\mathbf{W}$相似，而相似矩阵的特征值相同，即$\mathbf{S}$的特征值在区间$[-1,1]$内（回忆线性代数知识：如果$\mathbf{A}=\mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$，则$\mathbf{A}^k=\mathbf{P}\mathbf{\Lambda}^k\mathbf{P}^{-1}$，其中$\mathbf{\Lambda}$为对角阵），因此$\lim_{t\to\infty}\mathbf{S}^t$必不会发散至无穷大；又因为参数$\alpha\in(0,1)$，因此$\lim_{t\to\infty}\alpha^t$肯定等于 0；综合两项结果，所以

$$\lim_{t\to\infty}(\alpha\mathbf{S})^t=\mathbf{0}$$

根据等比数列公式

$$\lim_{t\to\infty}\sum_{i=0}^{t-1}(\alpha\mathbf{S})^i=\frac{\mathbf{I}-\lim_{t\to\infty}(\alpha\mathbf{S})^t}{\mathbf{I}-\alpha\mathbf{S}}=\frac{\mathbf{I}}{\mathbf{I}-\alpha\mathbf{S}}=(\mathbf{I}-\alpha\mathbf{S})^{-1}$$

所以

$$\begin{aligned}\mathbf{F}^*&=\lim_{t\to\infty}\mathbf{F}(t)\\&=\lim_{t\to\infty}\left[(\alpha\mathbf{S})^t\mathbf{Y}+(1-\alpha)\left(\sum_{i=0}^{t-1}(\alpha\mathbf{S})^i\right)\mathbf{Y}\right]\\&=(1-\alpha)(\mathbf{I}-\alpha\mathbf{S})^{-1}\mathbf{Y}\end{aligned}$$

即式(13.20) ，证毕！

8、式(13.21)的解释

这里主要是推导式(13.21)的最优解即为式(13.20)。将式(13.21)的目标函数进行变形：

第 1 部分：

先将范数平方拆开为四项：

$$\begin{aligned}\left\|\frac{1}{\sqrt{d_i}}\mathbf{F}_i-\frac{1}{\sqrt{d_j}}\mathbf{F}_j\right\|^2&=\left(\frac{1}{\sqrt{d_i}}\mathbf{F}_i-\frac{1}{\sqrt{d_j}}\mathbf{F}_j\right)\left(\frac{1}{\sqrt{d_i}}\mathbf{F}_i-\frac{1}{\sqrt{d_j}}\mathbf{F}_j\right)^\top\\&=\frac{1}{d_i}\mathbf{F}_i\mathbf{F}_i^\top+\frac{1}{d_j}\mathbf{F}_j\mathbf{F}_j^\top-\frac{1}{\sqrt{d_id_j}}\mathbf{F}_i\mathbf{F}_j^\top-\frac{1}{\sqrt{d_jd_i}}\mathbf{F}_j\mathbf{F}_i^\top\end{aligned}$$

其中$\mathbf{F}_i\in\mathbb{R}^{1\times|\mathcal{Y}|}$表示矩阵$\mathbf{F}$的第$i$行，即第$i$个示例$\boldsymbol{x}_i$的标记向量。将第 1 项中的$\sum_{i,j=1}^m$写为两个和求号$\sum_{i=1}^m\sum_{j=1}^m$的形式，并将上面拆分的四项中的前两项代入，得

$$\sum_{i,j=1}^{m}(\mathbf{W})_{ij}\frac{1}{d_i}\mathbf{F}_i\mathbf{F}_i^\top=\sum_{i=1}^{m}\frac{1}{d_i}\mathbf{F}_i\mathbf{F}_i^\top\sum_{j=1}^{m}(\mathbf{W})_{ij}=\sum_{i=1}^{m}\frac{1}{d_i}\mathbf{F}_i\mathbf{F}_i^\top\cdot d_i=\sum_{i=1}^{m}\mathbf{F}_i\mathbf{F}_i^\top$$

$$\sum_{i,j=1}^{m}(\mathbf{W})_{ij}\frac{1}{d_j}\mathbf{F}_j\mathbf{F}_j^\top=\sum_{j=1}^{m}\frac{1}{d_j}\mathbf{F}_j\mathbf{F}_j^\top\sum_{i=1}^{m}(\mathbf{W})_{ij}=\sum_{j=1}^{m}\frac{1}{d_j}\mathbf{F}_j\mathbf{F}_j^\top\cdot d_j=\sum_{j=1}^{m}\mathbf{F}_j\mathbf{F}_j^\top$$

以上化简过程中，两个求和号可以交换求和次序；又因为$\mathbf{W}$为对称阵，因此对行求和与对列求和效果一样，即$d_i=\sum_{j=1}^{m}(\mathbf{W})_{ij}=\sum_{j=1}^{m}(\mathbf{W})_{ji}$（已在式(13.12)推导时说明）。显然，

$$\sum_{i=1}^{m}\mathbf{F}_i\mathbf{F}_i^\top=\sum_{j=1}^{m}\mathbf{F}_j\mathbf{F}_j^\top=\sum_{i=1}^{m}\|\mathbf{F}_i\|^2=\|\mathbf{F}\|_{\mathrm{F}}^2=\mathrm{tr}\left(\mathbf{F}\mathbf{F}^\top\right)$$

以上推导过程中，第 1 个等号显然成立，因为二者仅是求和变量名称不同；第 2 个等号即将$\mathbf{F}_i\mathbf{F}_i^\top$写为$\|\mathbf{F}_i\|^2$形式；从第 2 个等号的结果可以看出这明显是在求矩阵$\mathbf{F}$各元素平方之和，也就是矩阵$\mathbf{F}$的 Frobenius 范数（简称 F 范数）的平方，即第 3 个等号；根据矩阵 F 范数与矩阵的迹的关系有第 4 个等号（详见第 10 章注解开篇的预备知识：矩阵的 F 范数与迹）。接下来，将上面拆分的四项中的第三项代入，得

$$\sum_{i,j=1}^{m}(\mathbf{W})_{ij}\frac{1}{\sqrt{d_id_j}}\mathbf{F}_i\mathbf{F}_j^\top=\sum_{i,j=1}^{m}(\mathbf{S})_{ij}\mathbf{F}_i\mathbf{F}_j^\top=\mathrm{tr}\left(\mathbf{S}^\top\mathbf{F}\mathbf{F}^\top\right)=\mathrm{tr}\left(\mathbf{S}\mathbf{F}\mathbf{F}^\top\right)$$

具体来说，以上化简过程为：

$$\begin{aligned}\mathbf{S}&=\begin{bmatrix}(\mathbf{S})_{11}&(\mathbf{S})_{12}&\cdots&(\mathbf{S})_{1m}\\(\mathbf{S})_{21}&(\mathbf{S})_{22}&\cdots&(\mathbf{S})_{2m}\\\vdots&\vdots&\ddots&\vdots\\(\mathbf{S})_{m1}&(\mathbf{S})_{m2}&\cdots&(\mathbf{S})_{mm}\end{bmatrix}\\&=\mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}\\&=\begin{bmatrix}\frac{1}{\sqrt{d_1}}&&&\\&\frac{1}{\sqrt{d_2}}&&\\&&\ddots&\\&&&\frac{1}{\sqrt{d_m}}\end{bmatrix}\begin{bmatrix}(\mathbf{W})_{11}&(\mathbf{W})_{12}&\cdots&(\mathbf{W})_{1m}\\(\mathbf{W})_{21}&(\mathbf{W})_{22}&\cdots&(\mathbf{W})_{2m}\\\vdots&\vdots&\ddots&\vdots\\(\mathbf{W})_{m1}&(\mathbf{W})_{m2}&\cdots&(\mathbf{W})_{mm}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{d_1}}&&&\\&\frac{1}{\sqrt{d_2}}&&\\&&\ddots&\\&&&\frac{1}{\sqrt{d_m}}\end{bmatrix}\\&=\begin{bmatrix}\frac{1}{\sqrt{d_1d_1}}(\mathbf{W})_{11}&\frac{1}{\sqrt{d_1d_2}}(\mathbf{W})_{12}&\cdots&\frac{1}{\sqrt{d_1d_m}}(\mathbf{W})_{1m}\\\frac{1}{\sqrt{d_2d_1}}(\mathbf{W})_{21}&\frac{1}{\sqrt{d_2d_2}}(\mathbf{W})_{22}&\cdots&\frac{1}{\sqrt{d_2d_m}}(\mathbf{W})_{2m}\\\vdots&\vdots&\ddots&\vdots\\\frac{1}{\sqrt{d_md_1}}(\mathbf{W})_{m1}&\frac{1}{\sqrt{d_md_2}}(\mathbf{W})_{m2}&\cdots&\frac{1}{\sqrt{d_md_m}}(\mathbf{W})_{mm}\end{bmatrix}\end{aligned}$$

由以上推导可以看出$(\mathbf{S})_{ij}=\frac{1}{\sqrt{d_id_j}}(\mathbf{W})_{ij}$，即第 1 个等号；而

$$\mathbf{F}\mathbf{F}^\top=\begin{bmatrix}\mathbf{F}_1\\\mathbf{F}_2\\\vdots\\\mathbf{F}_m\end{bmatrix}\begin{bmatrix}\mathbf{F}_1^\top&\mathbf{F}_2^\top&\cdots&\mathbf{F}_m^\top\end{bmatrix}=\begin{bmatrix}\mathbf{F}_1\mathbf{F}_1^\top&\mathbf{F}_1\mathbf{F}_2^\top&\cdots&\mathbf{F}_1\mathbf{F}_m^\top\\\mathbf{F}_2\mathbf{F}_1^\top&\mathbf{F}_2\mathbf{F}_2^\top&\cdots&\mathbf{F}_2\mathbf{F}_m^\top\\\vdots&\vdots&\ddots&\vdots\\\mathbf{F}_m\mathbf{F}_1^\top&\mathbf{F}_m\mathbf{F}_2^\top&\cdots&\mathbf{F}_m\mathbf{F}_m^\top\end{bmatrix}$$

若令$\mathbf{A}=\mathbf{S}\circ\mathbf{F}\mathbf{F}^\top$，其中$\circ$表示 Hadmard 积，即矩阵$\mathbf{S}$与矩阵$\mathbf{F}\mathbf{F}^\top$元素对应相乘（参见百度百科哈达玛积），因此

$$\sum_{i,j=1}^{m}(\mathbf{S})_{ij}\mathbf{F}_i\mathbf{F}_j^\top=\sum_{i,j=1}^{m}(\mathbf{A})_{ij}$$

可以验证，上式的矩阵$\mathbf{A}=\mathbf{S}\circ\mathbf{F}\mathbf{F}^\top$元素之和$\sum_{i,j=1}^{m}(\mathbf{A})_{ij}$等于$\mathrm{tr}\left(\mathbf{S}^\top\mathbf{F}\mathbf{F}^\top\right)$，这是因为

$$\begin{aligned}
&\mathrm{tr}\left(\begin{bmatrix}(\mathbf{S})_{11}&(\mathbf{S})_{12}&\cdots&(\mathbf{S})_{1m}\\(\mathbf{S})_{21}&(\mathbf{S})_{22}&\cdots&(\mathbf{S})_{2m}\\\vdots&\vdots&\ddots&\vdots\\(\mathbf{S})_{m1}&(\mathbf{S})_{m2}&\cdots&(\mathbf{S})_{mm}\end{bmatrix}^\top\cdot\begin{bmatrix}\mathbf{F}_1\mathbf{F}_1^\top&\mathbf{F}_1\mathbf{F}_2^\top&\cdots&\mathbf{F}_1\mathbf{F}_m^\top\\\mathbf{F}_2\mathbf{F}_1^\top&\mathbf{F}_2\mathbf{F}_2^\top&\cdots&\mathbf{F}_2\mathbf{F}_m^\top\\\vdots&\vdots&\ddots&\vdots\\\mathbf{F}_m\mathbf{F}_1^\top&\mathbf{F}_m\mathbf{F}_2^\top&\cdots&\mathbf{F}_m\mathbf{F}_m^\top\end{bmatrix}\right)\\
&=\begin{bmatrix}(\mathbf{S})_{11}\\(\mathbf{S})_{21}\\\vdots\\(\mathbf{S})_{m1}\end{bmatrix}^\top\cdot\begin{bmatrix}\mathbf{F}_1\mathbf{F}_1^\top\\\mathbf{F}_2\mathbf{F}_1^\top\\\vdots\\\mathbf{F}_m\mathbf{F}_1^\top\end{bmatrix}+\begin{bmatrix}(\mathbf{S})_{12}\\(\mathbf{S})_{22}\\\vdots\\(\mathbf{S})_{m2}\end{bmatrix}^\top\cdot\begin{bmatrix}\mathbf{F}_1\mathbf{F}_2^\top\\\mathbf{F}_2\mathbf{F}_2^\top\\\vdots\\\mathbf{F}_m\mathbf{F}_2^\top\end{bmatrix}+\ldots+\begin{bmatrix}(\mathbf{S})_{1m}\\(\mathbf{S})_{2m}\\\vdots\\(\mathbf{S})_{mm}\end{bmatrix}^\top\cdot\begin{bmatrix}\mathbf{F}_1\mathbf{F}_m^\top\\\mathbf{F}_2\mathbf{F}_m^\top\\\vdots\\\mathbf{F}_m\mathbf{F}_m^\top\end{bmatrix}\\
&=\sum_{i=1}^{m}(\mathbf{S})_{i1}\mathbf{F}_i\mathbf{F}_1^\top+\sum_{i=1}^{m}(\mathbf{S})_{i2}\mathbf{F}_i\mathbf{F}_2^\top+\ldots+\sum_{i=1}^{m}(\mathbf{S})_{im}\mathbf{F}_i\mathbf{F}_m^\top\\
&=\sum_{i,j=1}^{m}(\mathbf{S})_{ij}\mathbf{F}_i\mathbf{F}_j^\top
\end{aligned}$$

即第 2 个等号；易知矩阵$\mathbf{S}$是对称阵$\left(\mathbf{S}^\top=\mathbf{S}\right)$，即得第 3 个等号。又由于内积$\mathbf{F}_i\mathbf{F}_j^\top$是一个数（即大小为$1\times1$的矩阵），因此其转置等于本身，

$$\mathbf{F}_i\mathbf{F}_j^\top=\left(\mathbf{F}_i\mathbf{F}_j^\top\right)^\top=\left(\mathbf{F}_j^\top\right)^\top\left(\mathbf{F}_i\right)^\top=\mathbf{F}_j\mathbf{F}_i^\top$$

因此

$$\frac{1}{\sqrt{d_id_j}}\mathbf{F}_i\mathbf{F}_j^\top=\frac{1}{\sqrt{d_jd_i}}\mathbf{F}_j\mathbf{F}_i^\top$$

进而上面拆分的四项中的第三项和第四项相等：

$$\sum_{i,j=1}^{m}(\mathbf{W})_{ij}\frac{1}{\sqrt{d_id_j}}\mathbf{F}_i\mathbf{F}_j^\top=\sum_{i,j=1}^{m}(\mathbf{W})_{ij}\frac{1}{\sqrt{d_jd_i}}\mathbf{F}_j\mathbf{F}_i^\top$$

综上所述（以上拆分的四项中前两项相等、后两项相等，正好抵消系数$\frac{1}{2}$）：

$$\frac{1}{2}\left(\sum_{i,j=1}^{m}(\mathbf{W})_{ij}\left\|\frac{1}{\sqrt{d_i}}\mathbf{F}_i-\frac{1}{\sqrt{d_j}}\mathbf{F}_j\right\|^2\right)=\mathrm{tr}\left(\mathbf{F}\mathbf{F}^\top\right)-\mathrm{tr}\left(\mathbf{S}\mathbf{F}\mathbf{F}^\top\right)$$

第 2 部分：

西瓜书中式(13.21)的第 2 部分与原文献“Zhou D, Bousquet O, Lal T N, et al. Learning with local and global consistency[C]//Advances in neural information processing systems. 2004: 321-328.”中式(4)的第 2 部分不同：

$$\mathcal{Q}(F)=\frac{1}{2}\sum_{i,j=1}^{n}W_{ij}\left\|\frac{F_i}{\sqrt{D_{ii}}}-\frac{F_j}{\sqrt{D_{jj}}}\right\|^2+\mu\sum_{i=1}^{n}\left\|F_i-Y_i\right\|^2,\tag{4}$$

原文献中第 2 部分包含了所有样本（求和变量上限为n），而西瓜书只包含有标记样本，并且第 304 页第二段提到“式(13.21)右边第二项是迫使学得结果在有标记样本上的预测与真实

标记尽可能相同”；若按原文献式(4)在第二项中将未标记样本也包含进来，由于对于未标记样本$\mathbf{Y}_i=\mathbf{0}$，因此直观上理解是迫使未标记样本学习结果尽可能接近 0，这显然是不对的；有关这一点作者在第 24 次印刷勘误中进行了补充：“考虑到有标记样本通常很少而未标记样本很多，为缓解过拟合，可在式(13.21)中引入针对未标记样本的L_2范数项$\mu\sum_{i=l+1}^{l+u}\|\mathbf{F}_i\|^2$”，式(13.21) 加上此项之后就与原文献的式(4)完全相同了。将第二项写为 F 范数形式：

$$\sum_{i=1}^{m}\|\mathbf{F}_i-\mathbf{Y}_i\|^2=\|\mathbf{F}-\mathbf{Y}\|_{\mathrm{F}}^2$$

综上，式(13.21)目标函数$\mathcal{Q}(\mathbf{F})=\mathrm{tr}\left(\mathbf{F}\mathbf{F}^\top\right)-\mathrm{tr}\left(\mathbf{S}\mathbf{F}\mathbf{F}^\top\right)+\mu\|\mathbf{F}-\mathbf{Y}\|_{\mathrm{F}}^2$，求导：

$$\begin{aligned}\frac{\partial\mathcal{Q}(\mathbf{F})}{\partial\mathbf{F}}&=\frac{\partial\mathrm{tr}\left(\mathbf{F}\mathbf{F}^\top\right)}{\partial\mathbf{F}}-\frac{\partial\mathrm{tr}\left(\mathbf{S}\mathbf{F}\mathbf{F}^\top\right)}{\partial\mathbf{F}}+\mu\frac{\partial\|\mathbf{F}-\mathbf{Y}\|_{\mathrm{F}}^2}{\partial\mathbf{F}}\\&=2\mathbf{F}-2\mathbf{S}\mathbf{F}+2\mu(\mathbf{F}-\mathbf{Y})\end{aligned}$$

令$\mu=\frac{1-\alpha}{\alpha}$，并令$\frac{\partial\mathcal{Q}(\mathbf{F})}{\partial\mathbf{F}}=2\mathbf{F}-2\mathbf{S}\mathbf{F}+2\frac{1-\alpha}{\alpha}(\mathbf{F}-\mathbf{Y})=0$，移项化简即可得式(13.20)，即式(13.20)是正则化框架式(13.21)的解。

有关矩阵的迹及 F 范数求导可以搜索《The Matrix Cookbook(Version: November 15, 2012)》：
http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf
其中以上第 1 项求导参见式(111)，第 2 项求导参见式(109)，第 3 项根据式(132)再结合复合函数求导规则即可。

13.5 基于分歧的方法

西瓜书的伟大之处在于巧妙地融入了很多机器学习的研究分支，而非仅简单介绍经典的机器学习算法。比如本节处于半监督学习章节范围内，巧妙地将机器学习的研究热点之一多视图学习(multi-view learning) (https://arxiv.org/pdf/1304.5634.pdf)融入进来，类似地还有本章第一节将主动学习融入进来，在第 10 章第一节将 *k* 近邻算法融入进来，在最后一节巧妙地将度量学习(metric learning)融入进来等等。

协同训练是多视图学习代表性算法之一，本章叙述简单易懂。

1、图 13.6 的解释

第 2 行表示从样本集D_u中去除缓冲池样本D_s；

第 4 行，当$j=1$时$\langle\boldsymbol{x}_i^j,\boldsymbol{x}_i^{3-j}\rangle$即为$\langle\boldsymbol{x}_i^1,\boldsymbol{x}_i^2\rangle$，当$j=2$时$\langle\boldsymbol{x}_i^j,\boldsymbol{x}_i^{3-j}\rangle$即为$\langle\boldsymbol{x}_i^2,\boldsymbol{x}_i^1\rangle$，往后的$3-j$与此相同；注意本页左上角的注释：$\langle\boldsymbol{x}_i^1,\boldsymbol{x}_i^2\rangle$与$\langle\boldsymbol{x}_i^2,\boldsymbol{x}_i^1\rangle$表示的是同一个样本，因此第 1 个视图的有标记标训练集为$D_l^1=\{(\boldsymbol{x}_1^1,y_1),\ldots,(\boldsymbol{x}_l^1,y_l)\}$，第 2 个视图的有标记标训练集为$D_l^2=\{(\boldsymbol{x}_1^2,y_1),\ldots,(\boldsymbol{x}_l^2,y_l)\}$；

第 9 行到第 11 行是根据第j个视图的对缓冲池未标记样本预测结果置信度赋予伪标记，准备交给第$3-j$个视图使用；

第 14 行到第 16 行是不是应该从第 2 轮迭代开始检测，各视图样本集加入伪标记的未标记数据后训练结果是否和原先的一样？or 两个分类器h_1,h_2一样（即无分歧）？

13.6 半监督聚类

本节内容简单易懂，几乎不需要什么注解；当然，需要先理解 9.4.1 节的k-means。

1、图 13.7 的解释

注意算法第 4 行到第 21 行是依次对每个样本进行处理，其中第 8 行到第 21 行是尝试将样本$\boldsymbol{x}_i$到底应该划入哪个簇，具体来说是按样本$\boldsymbol{x}_i$到各均值向量的距离从小到大依次尝试，若最小的不违背$\mathcal{M}$和$\mathcal{C}$中的约束，则将样本$\boldsymbol{x}_i$划入该簇并置$\mathrm{is_merged}=\mathrm{true}$，此时第 8 行的$\mathrm{while}$循环条件为假不再继续循环，若从小到大依次尝试各簇后均违背$\mathcal{M}$和$\mathcal{C}$中的约束则第 16 行的if条件为真，算法报错结束；依次对每个样本进行处理后第 22 行到第 24 行更新均值向量，重新开始新一轮迭代，直到均值向量均未更新。

2、图 13.9 的解释

算法第 6 行到第 10 行即在聚类簇迭代更新过程中不改变种子样本的簇隶属关系；第 11 行到第 15 行即对非种子样本进行普通的k-means聚类过程；第 16 行到第 18 行更新均值向量，反复迭代，直到均值向量均未更新。

13.7 本章小结

监督学习和无监督学习是机器学习的两个学习范型(paradigm)。对于监督学习来说，要求训练样本具有标记信息；而对于无监督学习来说，则不需要训练样本具有标记信息。若在监督学习中没有足够的有标记样本，而是具有大量的无标记样本（在现实世界中正是如此，详见课本 13.1 节的叙述），此时半监督学习闪亮登场；若在无监督学习（本章特指聚类）中具有额外的监督信息（某些约束或少量有标记样本），此时半监督聚类闪亮登场。总之，少量的有标记样本和大量的无标记样本才是更经常要面对的问题，单纯的监督学习和无监督学习都有其局限性。

本章 13.2 节到 13.5 节先后分别介绍了半监督学习四大范型，即生成式半监督学习方法、半监督 SVM、图半监督学习、基于分歧的方法，均是利用了未标记样本的监督学习；而 13.6 节介绍的半监督聚类则是利用了额外监督信息的无监督学习。

附录：9.4.3 高斯混合聚类

1、式(9.28)的解释

该式就是多元高斯分布概率密度函数的定义式：

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}\left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

对应到我们常见的一元高斯分布概率密度函数的定义式：

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

其中$\sqrt{2\pi} = (2\pi)^{\frac{1}{2}}$对应$(2\pi)^{\frac{n}{2}}$，$\sigma$对应$\left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}$，指数项中分母中的方差$\sigma^2$对应协方差矩阵$\boldsymbol{\Sigma}$，$\frac{(x-\mu)^2}{\sigma^2}$对应$(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$。

概率密度函数$p(\boldsymbol{x})$是$\boldsymbol{x}$的函数。其中对于某个特定的$\boldsymbol{x}$来说，函数值$p(\boldsymbol{x})$就是一个数，若$\boldsymbol{x}$的维度为 2，则可以将函数$p(\boldsymbol{x})$的图像可视化，是三维空间的一个曲面。类似于一元高斯分布$p(x)$与横轴$p(x)=0$之间的面积等于 1（即$\int p(x)\mathrm{d}x=1$），$p(\boldsymbol{x})$曲面与平面$p(\boldsymbol{x})=0$之间的体积等于 1（即$\int p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=1$）。

注意，西瓜书中后面将$p(\boldsymbol{x})$记为$p(\boldsymbol{x}\mid\boldsymbol{\mu},\boldsymbol{\Sigma})$。

2、式(9.29)的解释

对于该式表达的高斯混合分布概率密度函数$p_{\mathcal{M}}(\boldsymbol{x})$，与式(9.28)中的$p(\boldsymbol{x})$不同的是，它由$k$个不同的多元高斯分布加权而来。具体来说，$p(\boldsymbol{x})$仅由参数$\boldsymbol{\mu},\boldsymbol{\Sigma}$确定，而$p_{\mathcal{M}}(\boldsymbol{x})$由$k$个“混合系数”$\alpha_i$以及$k$组参数$\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i$确定。

在西瓜书中该式下方(P207 最后一段)中介绍了样本的生成过程，实际也反应了“混合系数”α_i的含义，即α_i为选择第i个混合成分的概率，或者反过来说，α_i为样本属于第i个混合成分的概率。重新描述一下样本生成过程，根据先验分布$\alpha_1,\alpha_2,\ldots,\alpha_k$选择其中一个高斯混合成分（即第$i$个高斯混合成分被选到的概率为$\alpha_i$），假设选到了第$i$个高斯混合成分，其参数为$\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i$; 然后根据概率密度函数$p(\boldsymbol{x}\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$（即将式(9.28)中的$\boldsymbol{\mu},\boldsymbol{\Sigma}$替换为$\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i$）进行采样生成样本$\boldsymbol{x}$。两个步骤的区别在于第 1 步选择高斯混合成分时是从k个之中选其一（相当于概率密度函数是离散的），而第 2 步生成样本时是从$\boldsymbol{x}$定义域中根据$p(\boldsymbol{x}\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$选择其中一个样本，样本$\boldsymbol{x}$被选中的概率即为$p(\boldsymbol{x}\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$。即第 1 步对应于离散型随机变量，第 2 步对应于连续型随机变量。

3、式(9.30)的解释

若由上述样本生成方式得到训练集$D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$，现在的问题是对于给定样本$\boldsymbol{x}_j$，它是由哪个高斯混合成分生成的呢？该问题即求后验概率$p_{\mathcal{M}}(z_j\mid\boldsymbol{x}_j)$，其中$z_j\in\{1,2,\ldots,k\}$。下面对式(9.30)进行推导。

对于任意样本，再不考虑样本本身之前（即先验），若瞎猜一下它由第i个高斯混合成分生成的概率$P(z_j=i)$，那么肯定按先验概率$\alpha_1,\alpha_2,\ldots,\alpha_k$进行猜测，即$P(z_j=i)=\alpha_i$。若考虑样本本身带来的信息（即后验），此时再猜一下它由第i个高斯混合成分生成的概率$p_{\mathcal{M}}(z_j=i\mid\boldsymbol{x}_j)$，根据贝叶斯公式，后验概率$p_{\mathcal{M}}(z_j=i\mid\boldsymbol{x}_j)$可写为

$$p_{\mathcal{M}}(z_j=i\mid\boldsymbol{x}_j)=\frac{P(z_j=i)\cdot p_{\mathcal{M}}(\boldsymbol{x}_j\mid z_j=i)}{p_{\mathcal{M}}(\boldsymbol{x}_j)}$$

分子第 1 项$P(z_j=i)=\alpha_i$; 第 2 项即第i个高斯混合成分生成样本$\boldsymbol{x}_j$的概率$p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$，根据式(9.28)将$\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma}$替换为$\boldsymbol{x}_j,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i$即得；分母$p_{\mathcal{M}}(\boldsymbol{x}_j)$即为将$\boldsymbol{x}_j$代入式(9.29)即得。

注意，西瓜书中后面将$p_{\mathcal{M}}(z_j=i\mid\boldsymbol{x}_j)$记为$\gamma_{ji}$，其中$1\leq j\leq m$，$1\leq i\leq k$。

4、式(9.31)的解释

若将所有γ_{ji}组成一个矩阵Γ，其中γ_{ji}为第j行第i列的元素，矩阵Γ大小为$m\times k$，即

$$\Gamma=\begin{bmatrix}\gamma_{11}&\gamma_{12}&\cdots&\gamma_{1k}\\\gamma_{21}&\gamma_{22}&\cdots&\gamma_{2k}\\\vdots&\vdots&\ddots&\vdots\\\gamma_{m1}&\gamma_{m2}&\cdots&\gamma_{mk}\end{bmatrix}_{m\times k}$$

其中m为训练集样本个数，k为高斯混合模型包含的混合模型个数。

可以看出，式(9.31)就是找出矩阵Γ第j行的所有k个元素中最大的那个元素的位置。维基百科中有符号“arg max”的解释：https://en.wikipedia.org/wiki/Arg_max，可以学习一下。

进一步说，式(9.31)就是最大后验概率。

5、式(9.32)的解释

对于训练集$D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_m\}$，现在要把$m$个样本划分为$k$个簇，即认为训练集$D$的样本是根据$k$个不同的多元高斯分布加权而得的高斯混合模型生成的。

现在的问题是，k个不同的多元高斯分布的参数$\{(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\mid 1\leqslant i\leqslant k\}$及它们各自的权重$\alpha_1,\alpha_2,\ldots,\alpha_k$不知道，$m$个样本归到底属于哪个簇也不知道，该怎么办呢？

其实这跟k均值算法类似，开始时既不知道k个簇的均值向量，也不知道m个样本归到底属于哪个簇，最后我们采用了贪心策略，通过迭代优化来近似求解式(9.24)。

本节的高斯混合聚类求解方法与k均值算法，只是具体问题具体解法不同，从整体上来说，它们都应用了 7.6 节的期望最大化算法(EM 算法)。

具体来说，现假设已知式(9.30)的后验概率，此时即可通过式(9.31)知道m个样本归到底属于哪个簇，再来求解参数$\{(\alpha_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\mid 1\leqslant i\leqslant k\}$，怎么求解呢？对于每个样本$\boldsymbol{x}_j$来说，它出现的概率是$p_{\mathcal{M}}(\boldsymbol{x}_j)$，既然现在训练集$D$中确实出现了$\boldsymbol{x}_j$，我们当然希望待求解的参数$\{(\alpha_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\mid 1\leqslant i\leqslant k\}$能够使这种可能性$p_{\mathcal{M}}(\boldsymbol{x}_j)$最大；又因为我们假设$m$个样本是独立的，因此它们恰好一起出现的概率就是$\prod_{j=1}^{m}p_{\mathcal{M}}(\boldsymbol{x}_j)$，即所谓的似然函数；一般来说，连续容易造成下溢（$m$个大于 0 小于 1 的数相乘，当$m$较大时会非常非常小，以致于计算机无法表达这么小的数，产生下溢），所以常用对数似然替代，即式(9.32)。

6、式(9.33)的推导

该式等号左侧即偏导$\frac{\partial LL(D)}{\partial\boldsymbol{\mu}_i}$，下面先推导$\frac{\partial LL(D)}{\partial\boldsymbol{\mu}_i}$的表达式。重写式(9.32)如下：

$$LL(D)=\sum_{j=1}^{m}\ln\left(\sum_{l=1}^{k}\alpha_l\cdot p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)\right)$$

这里将第 2 个求和号的求和变量由式(9.32)的i改为了l，这是为了避免与$p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$中的变量$i$相混淆；再结合式(9.28)，重写$p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$的表达式如下：

$$p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)=\frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^{\top}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)}$$

接下来开始推导$\frac{\partial LL(D)}{\partial\boldsymbol{\mu}_i}$的表达式。根据链接求导规则

$$\frac{\partial LL(D)}{\partial\boldsymbol{\mu}_i}=\frac{\partial LL(D)}{\partial p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}\cdot\frac{\partial p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{\partial\boldsymbol{\mu}_i}$$

第 1 部分很容易进行求导：

$$\begin{aligned}\frac{\partial LL(D)}{\partial p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} &= \frac{\partial \sum_{j=1}^{m} \ln\left(\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)\right)}{\partial p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \\ &= \sum_{j=1}^{m} \frac{\partial \ln\left(\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)\right)}{\partial p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \\ &= \sum_{j=1}^{m} \frac{\alpha_i}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}\end{aligned}$$

第 2 部分求导略显复杂：

$$\begin{aligned}\frac{\partial p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\partial \boldsymbol{\mu}_i} &= \frac{\partial \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)}}{\partial \boldsymbol{\mu}_i} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \cdot \frac{\partial e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)}}{\partial \boldsymbol{\mu}_i}\end{aligned}$$

上面仅把常数项拿出来，使求导形式看起来更直观一些；剩下的求导部分又是复合函数求导：

$$\frac{\partial e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)}}{\partial \boldsymbol{\mu}_i} = e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)} \cdot -\frac{1}{2} \frac{\partial (\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}_i}$$

又因为协方差矩阵的逆矩阵$\boldsymbol{\Sigma}_i^{-1}$是对称阵，因此：

$$\begin{aligned}-\frac{1}{2} \frac{\partial (\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)}{\partial \boldsymbol{\mu}_i} &= -\frac{1}{2} \cdot 2\boldsymbol{\Sigma}_i^{-1} (\boldsymbol{\mu}_i - \boldsymbol{x}_j) \\ &= \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_j)\end{aligned}$$

上式有关矩阵求导可以搜索《The Matrix Cookbook(Version: November 15, 2012)》（式 86）：
http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf
因此第 2 部分求导结果为

$$\begin{aligned}\frac{\partial p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\partial \boldsymbol{\mu}_i} &= \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)} \cdot \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) \\ &= p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \cdot \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)\end{aligned}$$

综上所述

$$\frac{\partial LL(D)}{\partial \boldsymbol{\mu}_i} = \sum_{j=1}^{m} \frac{\alpha_i}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \cdot \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)$$

注意$\boldsymbol{\Sigma}_i^{-1}$对于求和变量$j$来说是常量，因此可以提到求和号外面，当令$\frac{\partial LL(D)}{\partial \boldsymbol{\mu}_i} = 0$时可以将该常量略掉（即等号两边同时左乘以$\boldsymbol{\Sigma}_i$），即得式(9.33)：

$$\frac{\partial LL(D)}{\partial \boldsymbol{\mu}_i} = \sum_{j=1}^{m} \frac{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} (\boldsymbol{x}_j - \boldsymbol{\mu}_i) = 0$$

7、式(9.34)的推导

根据式(9.30)可知：

$$\gamma_{ji} = p_{\mathcal{M}}(z_j = i \mid \boldsymbol{x}_j) = \frac{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}$$

则式(9.33)可重写为

$$\sum_{j=1}^{m} \gamma_{ji}(\boldsymbol{x}_j - \boldsymbol{\mu}_i) = 0$$

移项，得

$$\sum_{j=1}^{m} \gamma_{ji}\boldsymbol{x}_j = \sum_{j=1}^{m} \gamma_{ji}\boldsymbol{\mu}_i = \boldsymbol{\mu}_i \cdot \sum_{j=1}^{m} \gamma_{ji}$$

第二个等号是因为$\boldsymbol{\mu}_i$对于求和变量j来说是常量，因此可以提到求和号外面；因此

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{m} \gamma_{ji}\boldsymbol{x}_j}{\sum_{j=1}^{m} \gamma_{ji}}$$

8、式(9.35)的推导

该式推导过程与(9.33)(9.34)推导过程基本相同，根据链接求导规则

$$\frac{\partial LL(D)}{\partial \boldsymbol{\Sigma}_i} = \frac{\partial LL(D)}{\partial p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \cdot \frac{\partial p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\partial \boldsymbol{\Sigma}_i}$$

第 1 部分与式(9.33)推导过程一样，第 2 部分与式(9.33)的区别较大且较为复杂：

$$\begin{aligned}
\frac{\partial p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\partial \boldsymbol{\Sigma}_i} &= \frac{\partial \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)}}{\partial \boldsymbol{\Sigma}_i} \\
&= \frac{\frac{\partial e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)}}{\partial \boldsymbol{\Sigma}_i} \cdot (2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}} - e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)} \cdot \frac{\partial (2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}}{\partial \boldsymbol{\Sigma}_i}}{\left((2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}_i|^{\frac{1}{2}}\right)^2}
\end{aligned}$$

上式看起来复杂，实际就是函数求导规则$\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2}$；下面先求分子中的两项求导：

$$\begin{aligned}
\frac{\partial e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)}}{\partial \boldsymbol{\Sigma}_i} &= e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)} \cdot -\frac{1}{2} \frac{\partial (\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)}{\partial \boldsymbol{\Sigma}_i} \\
&= e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)} \cdot \frac{1}{2} \boldsymbol{\Sigma}_i^{-\top} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-\top} \\
&= e^{-\frac{1}{2}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_i)} \cdot \frac{1}{2} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}
\end{aligned}$$

其中第一个等号就是复合函数求导；第三个等号是因为$\boldsymbol{\Sigma}_i^{-1}$为对称阵，其中$\boldsymbol{\Sigma}_i^{-\top}$表示$\boldsymbol{\Sigma}_i^{-1}$的转置；第二个等号参见《The Matrix Cookbook(Version: November 15, 2012)》（式 61）。

$$\begin{aligned}\frac{\partial(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}}{\partial\boldsymbol{\Sigma}_i} &= \frac{(2\pi)^{\frac{n}{2}}}{2}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}\cdot\frac{\partial|\boldsymbol{\Sigma}_i|}{\partial\boldsymbol{\Sigma}_i}\\ &= \frac{(2\pi)^{\frac{n}{2}}}{2}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}\cdot|\boldsymbol{\Sigma}_i|\cdot\boldsymbol{\Sigma}_i^{-\top}\\ &= \frac{(2\pi)^{\frac{n}{2}}}{2}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}\cdot\boldsymbol{\Sigma}_i^{-1}\end{aligned}$$

上式推导中，第一个等号就是运用了复合函数求导规则，其中$\frac{\partial(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}}{\partial|\boldsymbol{\Sigma}_i|}=\frac{(2\pi)^{\frac{n}{2}}}{2}|\boldsymbol{\Sigma}_i|^{-\frac{1}{2}}$；第二个等号中的$\frac{\partial|\boldsymbol{\Sigma}_i|}{\partial\boldsymbol{\Sigma}_i}=|\boldsymbol{\Sigma}_i|\cdot\boldsymbol{\Sigma}_i^{-\top}$为行列式求导，参见《The Matrix Cookbook(Version: November 15, 2012)》（式 49），第三个等号是由于$\boldsymbol{\Sigma}_i^{-1}$为对称阵。

将分子中的两项求导结果代入

$$\begin{aligned}\frac{\partial p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{\partial\boldsymbol{\Sigma}_i} &= \frac{\frac{\partial e^{-\frac{1}{2}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)}}{\partial\boldsymbol{\Sigma}_i}\cdot(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}-e^{-\frac{1}{2}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)}\cdot\frac{\partial(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}}{\partial\boldsymbol{\Sigma}_i}}{\left((2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}\right)^2}\\ &= e^{-\frac{1}{2}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)}\cdot\frac{\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}_i^{-1}\cdot(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}-\frac{(2\pi)^{\frac{n}{2}}}{2}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}\cdot\boldsymbol{\Sigma}_i^{-1}}{\left((2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}\right)^2}\\ &= e^{-\frac{1}{2}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)}\cdot\frac{\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top-\boldsymbol{I}}{\left((2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}\right)^2}\cdot\frac{(2\pi)^{\frac{n}{2}}}{2}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}\cdot\boldsymbol{\Sigma}_i^{-1}\\ &= \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}}e^{-\frac{1}{2}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)}\cdot\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top-\boldsymbol{I}\right)\cdot\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\\ &= p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)\cdot\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top-\boldsymbol{I}\right)\cdot\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\end{aligned}$$

注意$\frac{(2\pi)^{\frac{n}{2}}}{2}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}$为一个数，矩阵$\boldsymbol{I}$为大小与协方差矩阵$\boldsymbol{\Sigma}_i$相同的单位阵。

综上所述

$$\begin{aligned}\frac{\partial LL(D)}{\partial\boldsymbol{\Sigma}_i} &= \sum_{j=1}^{m}\frac{\alpha_i\cdot p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k}\alpha_l\cdot p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)}\cdot\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top-\boldsymbol{I}\right)\cdot\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\\ &= \sum_{j=1}^{m}\gamma_{ji}\cdot\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top-\boldsymbol{I}\right)\cdot\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}\end{aligned}$$

当令$\frac{\partial LL(D)}{\partial\boldsymbol{\Sigma}_i}=0$时可以将该常量略掉（即等号两边同时右乘以$2\boldsymbol{\Sigma}_i$）：

$$\frac{\partial LL(D)}{\partial\boldsymbol{\Sigma}_i}=\sum_{j=1}^{m}\gamma_{ji}\cdot\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top-\boldsymbol{I}\right)=0$$

移项，得

$$\sum_{j=1}^{m}\gamma_{ji}\cdot\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top=\sum_{j=1}^{m}\gamma_{ji}\boldsymbol{I}$$

两边同时左乘以$\boldsymbol{\Sigma}_i$，得

$$\sum_{j=1}^{m}\gamma_{ji}\cdot(\boldsymbol{x}_j-\boldsymbol{\mu}_i)(\boldsymbol{x}_j-\boldsymbol{\mu}_i)^\top=\sum_{j=1}^{m}\gamma_{ji}\boldsymbol{\Sigma}_i=\boldsymbol{\Sigma}_i\sum_{j=1}^{m}\gamma_{ji}$$

第二个等号是因为$\boldsymbol{\Sigma}_i$对于求和变量j来说是常量，因此可以提到求和号外面；因此

$$\boldsymbol{\Sigma}_i = \frac{\sum_{j=1}^{m} \gamma_{ji} \cdot (\boldsymbol{x}_j - \boldsymbol{\mu}_i)(\boldsymbol{x}_j - \boldsymbol{\mu}_i)^\top}{\sum_{j=1}^{m} \gamma_{ji}}$$

9、式(9.36)的解释

该式就是$LL(D)$添加了等式约束$\sum_{i=1}^{k} \alpha_i = 1$的拉格朗日形式。

有个疑问是除了等式约束，还有不等式约束$\alpha_i \geqslant 0$，该式如何保证呢？

10、式(9.37)的推导

重写式(9.32)如下：

$$LL(D) = \sum_{j=1}^{m} \ln \left(\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) \right)$$

这里将第 2 个求和号的求和变量由式(9.32)的i改为了l，这是为了避免对α_i求导时与变量i相混淆。将式(9.36)中的两项分别对α_i求导，得

$$\begin{aligned}
\frac{\partial LL(D)}{\partial \alpha_i} &= \frac{\partial \sum_{j=1}^{m} \ln \left(\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l) \right)}{\partial \alpha_i} \\
&= \sum_{j=1}^{m} \frac{1}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \cdot \frac{\partial \sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}{\partial \alpha_i} \\
&= \sum_{j=1}^{m} \frac{1}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)
\end{aligned}$$

$$\frac{\partial \left(\sum_{l=1}^{k} \alpha_l - 1 \right)}{\partial \alpha_i} = \frac{\partial \left(\alpha_1 + \alpha_2 + \ldots + \alpha_i + \ldots + \alpha_k - 1 \right)}{\partial \alpha_i} = 1$$

综合两项求导结果，并令导数等于零即得式(9.37)。

11、式(9.38)的推导

注意，在西瓜书第 14 次印刷中式(9.38)上方的一行话进行了勘误：“两边同乘以α_i，对所有**混合成分**求和可知$\lambda = -m$”，将原来的“样本”修改为“混合成分”。

对式(9.37)两边同乘以α_i，得

$$\sum_{j=1}^{m} \frac{\alpha_i p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)} + \lambda \alpha_i = 0$$

结合式(9.30)发现，求和号内即为后验概率γ_{ji}，即

$$\sum_{j=1}^{m} \gamma_{ji} + \lambda \alpha_i = 0$$

对所有**混合成分**求和，得

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\gamma_{ji}+\sum_{i=1}^{k}\lambda\alpha_i=0$$

注意$\sum_{i=1}^{k}\alpha_i=1$，因此$\sum_{i=1}^{k}\lambda\alpha_i=\lambda\sum_{i=1}^{k}\alpha_i=\lambda$，根据式(9.30)中$\gamma_{ji}$表达式可知

$$\begin{aligned}\sum_{i=1}^{k}\gamma_{ji}&=\sum_{i=1}^{k}\frac{\alpha_i\cdot p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k}\alpha_l\cdot p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)}\\&=\frac{\sum_{i=1}^{k}\alpha_i\cdot p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k}\alpha_l\cdot p(\boldsymbol{x}_j\mid\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)}\\&=1\end{aligned}$$

再结合加法满足交换律，所以

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\gamma_{ji}=\sum_{j=1}^{m}\sum_{i=1}^{k}\gamma_{ji}=\sum_{j=1}^{m}1=m$$

将以上分析结果代入$\sum_{i=1}^{k}\sum_{j=1}^{m}\gamma_{ji}+\sum_{i=1}^{k}\lambda\alpha_i=0$，移项即得$\lambda=-m$。将此结果代入$\sum_{j=1}^{m}\gamma_{ji}+\lambda\alpha_i=0$，即$\sum_{j=1}^{m}\gamma_{ji}-m\alpha_i=0$，整理即得式(9.38)。

12、图 9.6 的解释

第 1 行初始化参数，本页接下来的例子是按如下策略初始化的：混合系数$\alpha_i=\frac{1}{k}$；任选训练集中的k个样本分别初始化k个均值向量$\boldsymbol{\mu}_i(1\leqslant i\leqslant k)$；使用对角元素为0.1的对角阵初始化$k$个协方差矩阵$\boldsymbol{\Sigma}_i(1\leqslant i\leqslant k)$。

第 3~5 行根据式(9.30)计算共$m\times k$个γ_{ji}。

第 6~10 行分别根据式(9.34)、式(9.35)、式(9.38)使用刚刚计算得到的γ_{ji}更新均值向量、协方差矩阵、混合系数；注意第 8 行计算协方差矩阵时使用的是第 7 行计算得到的均值向量，这并没错，因为协方差矩阵$\boldsymbol{\Sigma}_i'$与均值向量$\boldsymbol{\mu}_i'$是对应的，而非$\boldsymbol{\mu}_i$；第 7 行的$\boldsymbol{\mu}_i'$在第 8 行使用之后会在下一轮迭代中第 4 行计算γ_{ji}再次使用。

整体来说，第 2~12 行就是一个 EM 算法的具体使用例子，学习完 7.6 节 EM 算法可能根本无法理解其思想。此例中有两组变量，分别是γ_{ji}和$(\alpha_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$，它们之间相互影响，但都是未知的，因此 EM 算法就有了用武之地：初始化其中一组变量$(\alpha_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$，然后计算$\gamma_{ji}$；再根据$\gamma_{ji}$根据最大似然推导出的公式更新$(\alpha_i,\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$，反复迭代，直到满足停止条件。