CSLT

现代机器学习技术导论

2017 年 6 月 29 日

Springer

Contents

Chapter 1
线性模型

线性模型是机器学习中最简单，也是最常用的模型。所谓线性，在不同领域和不同应用场景下有不同的含义。我们所讨论的线性，是指变量间具有如下简单形式的模型：

$$\mathbf{t} = \mathbf{W}\mathbf{x} \tag{1.1}$$

其中，$\mathbf{x}$和$\mathbf{t}$是两个多维变量，$\mathbf{W}$ 为模型参数。

线性模型虽然简单，但在机器学习中具有重要意义。首先，线性模型简单高效，容易实现；其次，很多实际问题都具有粗略的线性性，特别是在选择了合理的特征提取方式之后，这一线性性会更为明显，使得线性模型已经足以胜任工作；最后，线性模型参数较少，适应性强，具有很强的泛化能力。因此，当我们面对一个机器学习问题的时候，首先要考虑的就是线性模型。

本章我们将讨论几种简单的线性模型：一种是线性预测模型，包括线性回归和Logistic回归，在讨论它们概率意义的基础上引入贝叶斯方法；另一种是线性概率模型，基于隐变量和观察变量间的线性假设来推理数据的内在结构，如PCA、LDA、PLDA等。前者是有监督学习，后者一般用于无监督学习。

1.1 线性预测模型

如果$\mathbf{x}$和$\mathbf{t}$都是可见变量，则公式 1.1表示一种基于$\mathbf{x}$对$\mathbf{t}$的预测。这一模型称为“线性预测模型”。即使$\mathbf{x}$和$\mathbf{t}$之间不存在线性关系，也可以通过某种变换$\boldsymbol{\phi}(\cdot)$建立这种关系，即：

$$\mathbf{t} = \mathbf{W}\boldsymbol{\phi}(\mathbf{x}) \tag{1.2}$$

从模型角度来说，不论是基于原始数据还是基于变换后的数据，模型性质和优化方法都不会受到影响。我们在后面章节会看到，变换$\boldsymbol{\phi}$ 具有重要意义，引入这一变换使得很多在原始变量空间里无法用线性模型解决的问题在变换空间中得以合理解决。下面我们从简单的多项式拟合问题开始讨论。

1.1.1 从多项式拟合说起

假设一个包括N个样本的数据集$D = [(x_1,t_1),\ldots,(x_N,t_N)]$，其中每个输入变量$x_i$ 对应的观测值为t_i，我们的任务是学习一个预测函数$y = f(x)$，使其对数据集D中x_i预测的结果接近t_i。如果限定该预测函数为M次多阶式，则得到预测公式为：

$$y(x;\mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{1.3}$$

其中$\mathbf{w} = [w_0, w_1, \ldots, w_M]^T$是对应每阶的预测系数。图 1.1给出一个$M = 1$的预测函数，事实上是一条以$w_0$为截距，以$w_1$为斜率的直线。

给定了预测模型（M确定），我们需要对预测系数$\mathbf{w}$ 进行优化。为此，需定义误差函数，最小化该误差函数即可得到优化的$\mathbf{w}$。一般定义训练集D上的平方误差为误差函数，公式如下：

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{y(x_i;\mathbf{w}) - t_i\}^2 \tag{1.4}$$

注意该误差函数是$\mathbf{w}$的函数。对上式根据$\mathbf{w}$进行优化。首先将模型代入该函数，有:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}(\sum_{j=0}^{M} w_j x_i^j - t_i)^2$$

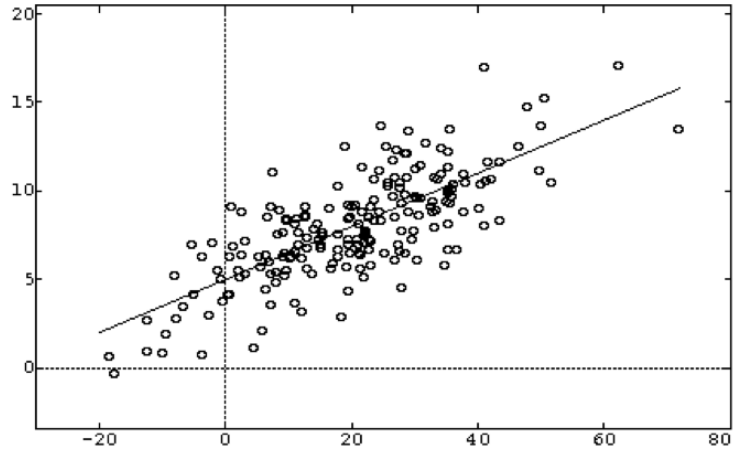

Fig. 1.1 多项式线性预测函数在$M=1$时，$y(x;\mathbf{w})=w_0+w_1x$是一条直线。

对每个参数w_k求偏导数并令其等于零，可得：

$$\sum_{i=1}^{N} x_i^k \sum_{j=0}^{M} w_j x_i^j = \sum_{i=1}^{N} t_i x_i^k \quad j=0,1,2,...,M$$

整理得：

$$\sum_{j=0}^{M} w_j \sum_{i=1}^{N} x_i^{k+j} = \sum_{i=1}^{N} t_i x_i^k \quad j=0,1,2,...,M$$

展开写成：

$$\begin{bmatrix} w_0\sum_{i=1}^{N}x_i^0 + & w_1\sum_{i=1}^{N}x_i^1 + ... + & w_M\sum_{i=1}^{N}x_i^M &= \sum_{i=1}^{N}t_i x_i^0 \\ w_0\sum_{i=1}^{N}x_i^1 + & w_1\sum_{i=1}^{N}x_i^2 + ... + & w_M\sum_{i=1}^{N}x_i^{M+1} &= \sum_{i=1}^{N}t_i x_i^1 \\ \cdots & \cdots & \cdots & \cdots \\ w_0\sum_{i=1}^{N}x_i^M + & w_1\sum_{i=1}^{N}x_i^{M+1} + ... + & w_M\sum_{i=1}^{N}x_i^{2M} &= \sum_{i=1}^{N}t_i x_i^M \end{bmatrix}$$

写成矩阵格式有：

$$\begin{bmatrix} \sum_{i=1}^{N} x_i^0 & \sum_{i=1}^{N} x_i^1 & \cdots & \sum_{i=1}^{N} x_i^M \\ \sum_{i=1}^{N} x_i^1 & \sum_{i=1}^{N} x_i^2 & \cdots & \sum_{i=1}^{N} x_i^{M+1} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{i=1}^{N} x_i^M & \sum_{i=1}^{N} x_i^{M+1} & \cdots & \sum_{i=1}^{N} x_i^{2M} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \cdots \\ w_M \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} t_i x_i^0 \\ \sum_{i=1}^{N} t_i x_i^1 \\ \cdots \\ \sum_{i=1}^{N} t_i x_i^M \end{bmatrix}$$

可以证明，上式中左边矩阵是可逆的，因此$\mathbf{w}$有唯一解，即为最优预测拟数。上述以多项式为预测函数的学习方法一般称为多项式拟合，意思是拟合训练样本中的输入变量x与观测变量t之间的对应关系。

如果将多项式拟合中的每一阶x^j看作一个非线性映射$\phi_j(x) = x^j$，并将一元变量x扩展到多元变量$\mathbf{x}$，则上述多项式拟合可以扩展到通用的线性拟合方法。设非线性映射个数为M，将映射后的变量写成向量格式:

$$\boldsymbol{\phi}(\mathbf{x}) = [\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), ...\phi_M(\mathbf{x})]^T$$

同样采用平方误差作为误差函数：

$$E(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\{\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) - t_i\}^2$$

对上式取$\mathbf{w}$的导数并使之为零，即有：

$$\begin{aligned} 0 &= \sum_{i=1}^{N}\{\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) - t_i\}\boldsymbol{\phi}(\mathbf{x}_i)^T \\ &= \mathbf{w}^T\sum_{i=1}^{N}\boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)^T - \sum_{i=1}^{N} t_i\boldsymbol{\phi}(\mathbf{x}_i)^T \end{aligned}$$

写成矩阵形式，有：

$$\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{w}^T = \boldsymbol{\Phi}^T\mathbf{t}$$

其中$\boldsymbol{\Phi}$为数据矩阵，每一行代表一个样本，每一列代表一个非线性映射，即：

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x_1}) & \phi_1(\mathbf{x_1}) & \cdots & \phi_M(\mathbf{x_1}) \\ \phi_0(\mathbf{x_2}) & \phi_1(\mathbf{x_2}) & \cdots & \phi_M(\mathbf{x_2}) \\ \cdots & \cdots & \cdots & \cdots \\ \phi_0(\mathbf{x_N}) & \phi_1(\mathbf{x_N}) & \cdots & \phi_M(\mathbf{x_N}) \end{pmatrix}$$

由此可得线性拟合的最优预测参数为：

$$\mathbf{w} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}\mathbf{t}$$

到现在为止，我们似乎已经完美解决了多项式拟合问题，连包含非线性变换的多元变量线性拟合问题也得到了确定解。然而，如果我们仔细想一想，就会发现事情并不是那么简单。一个有意思的问题是：我们为什么要使用平方误差作为误差函数而不是其他的函数，比如绝对值误差函数？下一节我们将引入了概率这一工具来解释这些选择。我们会发现，平方误差事实上对应的是训练数据中的一种高斯不确定性。

1.1.2 线性回归

几乎在所有实际问题中，都存在随机性或不确定性。这些不确定性可能来自于观测手段的不精确，但更多来自我们对数据本身理解上的局限性。例如当我们考察一架从北京飞往上海的飞机时，可以看到其飞行轨迹总体上是一条平稳的曲线，但细致观察，就会发现很多不确定性。这些不确定性有些来自驾驶员的自主操作，有些来自发动机转动时的不稳定，有些来自飞行过程中遇到的气流、云朵等的影响，还有的来自机舱内乘客的活动，等等。假设我们考虑到这所有的细节，依动力学列出一个庞大的方程组，则飞行过程中的绝大部分不确定性可以得到解释。然而，在实际应用中，考虑到问题的复杂度，不可能对这些动力学细节一一建模；即便我们想这么做，也不可能穷尽所有细节，总会有些因素超出了我们的考虑范围和理解能力。如果我们忽略这些细节，在宏观上的表现就是观测数据的不确定性。这意味着不确定性在几乎所有实际问题中都是不可避免的。

那么我们该如何处理这些不确定性呢？在这一节中我们将引入概率模型来帮助我们描述这些不确定性，并由此得到基于概率的最优解。引入概率模型是机器学习历史上里程碑式的贡献，本书中几乎所有章节都围绕概率展开。就本章而言，概率方法可让我们对线性模型的建模和优化有更深刻的理解。这一理解有助于我们将线性模型和未来要讨论的复杂模型联系起来，整理成为一个统一的思维，并基于此对模型和算法进行系统化的设计、优化和扩展。

回到我们之前的线性拟合任务$y = \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x})$。我们希望拟合得到的$y$和数据的原始标注$t$越相似越好，因此提出了基于最小均方误差方法。现在我们假设：y与t之所以存在差别，是因为观察值t由于种种原因是随机的，不确

定的。不论产生这种随机性的原因是什么，我们假设这一随机性符合一个以0为均值，以β^{-1}为方差的高斯分布。引入一个随机变量ε 来表示这一随机性，则有：

$$\begin{aligned} t &= y(\mathbf{x};\mathbf{w}) + \varepsilon \\ &= \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + \varepsilon \end{aligned} \tag{1.5}$$

其中$\varepsilon \sim N(0, \beta^{-1})$。

仔细观察公式 1.5可以看到，事实上我们构造了一个由$\mathbf{x}$ 到观察值t的生成模型：首先，由输入变量$\mathbf{x}$经过非线性映射生成特征向量$\boldsymbol{\phi}(\mathbf{x})$，再经过线性映射生成目标$y$，最后加入一个高斯噪声得到目标的观察值$t$。这一模型称为线性回归模型（Linear Regression)。所谓回归，一般指对输入变量$\mathbf{x}$和目标变量y之间相互依赖的定量关系的统计分析。图 1.2给出了在一维输入变量情况下，线性回归模型生成目标变量的示意图。

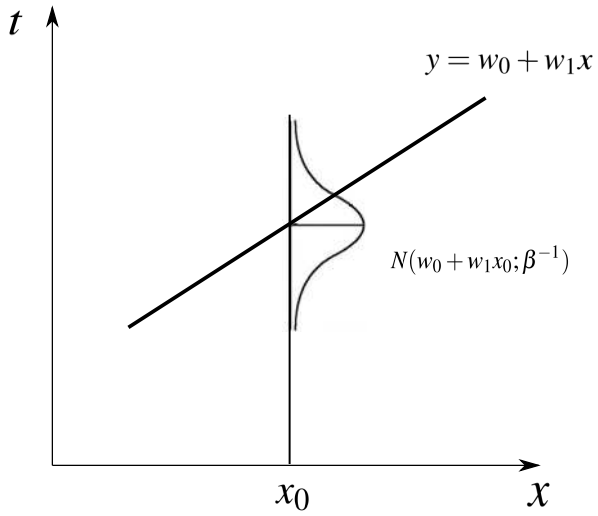

Fig. 1.2 设生成函数为$y(x;\mathbf{w}) = w_0 + w_1 x$，目标值$t$服从以$y(x;\mathbf{w})$ 为中心的高斯分布。

给定一个输入变量$\mathbf{x}$，则可以基于上述线性回归模型度量对应的目标观察值t的生成概率：

$$p(t|\mathbf{x};\mathbf{w}\beta) = N(t|y(\mathbf{x};\mathbf{w}), \beta^{-1})$$

如果我们将$(\mathbf{x},t)$作为一个整体，则上式也是该二元组在这一模型下的生成概率。对给定的训练数据集$D=\{\mathbf{x_i}t_i\}_{i=1}^N$，该模型生成这一数据集的总概率为：

$$p(D;\mathbf{w},\beta)=\prod_{i=1}^{N}N(t_i|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x_i}),\beta^{-1}) \tag{1.6}$$

上式是$\mathbf{w}$和β的函数，一般称为似然函数。显然，模型对某一数据集的建模能力越强，则该模型生成这一数据集的概率越大，似然函数的值也越大。因此，如果我们能找到一组参数使得似然函数的值最大化，则可实现该模型在生成概率意义上的最优化。这一优化准则称为最大似然（Maximum Likelihood, ML）准则，相应的优化方法称为最大似然估计。

对线性回规模型的最大似然估计可通过将式 1.6所示的似然函数取偏导数的零点得到。为方便，对似然函数取对数，并带入高斯分布的概率公式，有：

$$\begin{aligned}\ln p(D;\mathbf{w},\beta) &= \sum_{n=1}^{N} ln\{N(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x_n}),\beta^{-1})\}\\ &= \frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)-\beta E(\mathbf{w})\end{aligned}$$

其中前两项与$\mathbf{w}$无关，第三项为:

$$E(\mathbf{w})=\frac{1}{2}\sum_{n=1}^{N}\{t_n-\mathbf{w}^T\boldsymbol{\phi}(x_n)\}^2$$

因此，对$\mathbf{w}$的似然估计等价于对$E(\mathbf{w})$的最小化。读者想必已经对$E(\mathbf{w})$很熟悉了，这就是上一节我们在线性拟合方法中定义的平方误差。这意味着最大似然估计事实上等价于线性拟合。

现在让我们看看这一等价性是如何产生的。仔细观察似然函数，可以看到平方项其实来自高斯分布中的指数项$e^{\frac{\beta}{2}(t-\mathbf{w}^T\boldsymbol{\phi}(x))^2}$，这说明线性拟合中的平方误差事实上假设了目标观察值的高斯噪声分布。一方面，这说明平方误差是合理的，因为基于中心极限定理，高斯噪声是最简单也是最现实的噪声分布；另一方面，也说明概率模型是一种更通用、更直观的建模方式。在没有引入概率模型时，我们并不清楚平方误差对应的数据分布情况，因而也不知道这一误差是否合理；一旦我们引入了概率工具，将这一误差和高斯假设联系起来，就立刻理解了这一误差的合理性与局限性，并马上想到应对不同的数据分布设计不同的误差函数。事实上，我们在下面几节会反复看到这样的情况：当引入概率模型后，很多传统方法才挖掘出了隐藏在经验设计后面

的假设和含义，意识到自身的优点和不足，并激发出更多设计灵感和改进思路。

然而，现在让我们先把对线性回归模型的讨论完成。优化$\mathbf{w}$的过程和线性回归方法完全相同，我们仅写出其推导形式。首先计算上述平方误差的梯度：

$$\nabla_{\mathbf{w}} \ln p(D|\mathbf{w},\beta) = \sum_{n=1}^{N} \{t_n - \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x_n})\} \boldsymbol{\phi}(\mathbf{x_n}) \tag{1.7}$$

上式说明该误差函数对$\mathbf{w}$的梯度由预测值和观察值之间的差异导致，以这些差异为权重对输入变量取均值即得模型在当前参数下的梯度。

取梯度为零，解得$\mathbf{w}$如下：

$$\mathbf{w}_{ML} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{t}$$

其中下标ML表明该解基于最大似然准则得到。同理，基于β求导并令其等于零：

$$\nabla_{\beta} \ln p(D|\mathbf{w},\beta) = \frac{N}{2\beta} - E(w)$$

代入$\mathbf{w}$的最大似然估计$\mathbf{w}_{ML}$，得到：

$$\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^{N} \{t_n - \mathbf{w}_{ML}^T \boldsymbol{\phi}(\mathbf{x_n})\}^2 \tag{1.8}$$

注意到$\frac{1}{\beta}$事实上是高斯分布的方差，上式说明，该方差的估计事实上是基于最大似然估计的回归模型对集中的数据进行预测而得到的预测残差。

假设我们已经通过上述最大似然估计得到了一个线性模型如下：

$$t = \mathbf{w}\boldsymbol{\phi}(\mathbf{x}) + \varepsilon$$

如何基于该模型预测一个新的输入变量$\mathbf{x}'$ 的目标t 呢？首先注意到我们得到的模型包含一个随机变量ε，因此是一个随机生成模型，该模型对一个确定的输入$\mathbf{x}'$并没有一个确定的预测t，而是给出某一预测t的概率$p(t|\mathbf{x})$。后面我们会看到，这种基于概率的预测对完整描述一个随机系统具有重要意义，但现在我们只希望像通常的预测模型一样有个确定的预测值。一种方法是求预测的期望，即：

$$\mathbb{E}[t|\mathbf{x}] = \int t\, p(t|\mathbf{x}) \mathrm{d}t = y(\mathbf{x};\mathbf{w}_{ML}) = \mathbf{w}_{ML}^T \boldsymbol{\phi}(\mathbf{x})$$

可见，这种预测的结果和线性拟合的结果是一致的，恰是预测的中值$y(\mathbf{x};\mathbf{w_{MT}})$，同时也是预测最大概率所处的位置。这些结果并不是巧合，而是高斯分布的特性决定的。

一个有意思的问题是，高斯分布是最优选择吗？回答这个问题需要结合数据的实际分布情况。比如当数据具有很强的长尾特性时，就可能要考虑像Student-t分布或拉普拉斯分布。一般来说，如果我们对数据特性了解并不清楚，高斯分布总是安全的选择。然而有一种情况需要我们必须考虑非高斯分布：当目标t是离散的，则高斯分布显然是不适合的。下节要讨论的Logistic 回归模型即属于这种情况。

上述讨论中我们假设目标变量t是单变量。同样的方法很容易扩展到多变量情况；同时，ε事实上可以认为是变量$\mathbf{x}$ 的一部分，因此线性回归模型事实上是标准线性模型$\mathbf{t} = W\boldsymbol{\phi}(\mathbf{x})$的一种特殊形式，其中$\mathbf{x}$中一部分是可观察变量，另一部分则是不可见的，称为隐变量（Latent Variable）。在线性概率模型一节中我们会看到$\mathbf{x}$全是隐变量的情况。

1.1.3 Fisher准则与线性分类

在分类问题中，给定输入向量$\mathbf{x}$的特征$\boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x})$，我们希望分类器能预测该输入所属的类别$t$。总体来说，分类问题有三种可能的求解方法：

- 区分函数法：设计一个区分函数$f(\boldsymbol{\phi};\mathbf{w})$，基于某种准则对该函数进行优化，进而得到分类面。代表性方法如Fisher 线性分类函数。与线性拟合类似，这一方法基于人为定义的准则，没有考虑概率意义，但直观简洁。
- 生成性概率模型法：对每个类C_i建立一个统计模型$P(\boldsymbol{\phi}|C_i;\mathbf{w})$，在分类时考察测试样本在每个模型上的概率，再基于贝叶斯公式得到属于某一类的后验概率。这一方法依赖模型假设与实际数据的契合程度，假设越合理，分类性能越好。
- 区分性概率模型法：直接对后验概率$P(C_i|\boldsymbol{\phi};\mathbf{w})$建模。这一方法不对数据分布做假设，只关注分类面，在分类面比较复杂的任务中更有优势。

本节我们从著名的Fisher区分函数开始讨论，这几乎是最容易想到的分类方法。为简便，我们只讨论二分类问题，但相关结论可以扩展到多分类问题中。

设N个采样点$\{\boldsymbol{\phi}_i\}_{i=1}^N$ 分为两类C_1和C_2，这些点通过一个线性映射投影到一维空间y,

$$y = \mathbf{w}^T \boldsymbol{\phi} \tag{1.9}$$

其中$\mathbf{w}$是映射参数。如果基于该训练数据能学习一个优化的$\mathbf{w}$，使得不同类的训练样本在映射空间里的区分性最大，则基于式 1.9即可得到一个简单的分类函数。Fisher准则定义了一个区分性度量：

$$J(\mathbf{w}) = \frac{m_2 - m_1}{s_1^2 + s_2^2}$$

其中m_1, m_2是C_1和C_2的样本点在映射空间里的均值，s_1和s_2是相应的方差。代入公式 1.9，有：

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \tag{1.10}$$

其中S_W是类间协方差矩阵：

$$S_W = \sum_{\mathbf{x}_n \in C_1} (\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^T + \sum_{\mathbf{x}_n \in C_2} (\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^T$$

其中$\mathbf{m}_1$和$\mathbf{m}_2$是两类样本在原空间中的时均值。S_B是类内协方差矩阵，定义如下：

$$S_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T$$

式 1.10表明类间距离越大，类内的分散程度越小，则依Fisher准则这两类的区分性越强。这显然是符合直觉的。基于Fisher准则对$\mathbf{w}$ 进行优化，有:

$$(\mathbf{w}^T S_B \mathbf{w}) S_W \mathbf{w} = (\mathbf{w}^T S_W \mathbf{w}) S_B \mathbf{w}$$

注意到：$(\mathbf{w}^T S_B \mathbf{w})$是标量，且:

$$S_B \mathbf{w} = (\mathbf{m}_2 - \mathbf{m}_1)\{(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w}\} \tag{1.11}$$

$$\propto \mathbf{m}_2 - \mathbf{m}_1 \tag{1.12}$$

则有：

$$S_W \mathbf{w} \propto \mathbf{m_2} - \mathbf{m_1}$$

如果S_w满秩，则有

$$\mathbf{w} \propto S_w^T (\mathbf{m_2} - \mathbf{m_1}) \tag{1.13}$$

上式意味着最有区分性的方向$\mathbf{w}$应该是这两类采样点中心连心的方向，同时被类内方差矩阵调制。上述基于Fisher准则的线性模型也被称为线性判别分

析（Linear Discriminate Analysis, LDA）。通过设计合理的Fisher准则，LDA很容易扩展到多分类问题。

Fisher准则的合理性是显然的，但我们碰到了与线性拟合同样的问题：为什么要选择这一准则，其它准则不好吗？为回答这一问题，让我们换一种思路，用线性拟合来求解区分函数。

设类C_1中的样本数为N_1，C_2中的样本数为N_2，二者加起来一共N个训练样本点。对C_1中的样本点，设其目标$t = N/N_1$，对C_2中的样本点，目标为$t = -N/N_2$。则线性拟合的误差函数为：

$$E(\mathbf{w}) = \frac{1}{2}\sum_{\boldsymbol{\phi}_n \in C_1}(\mathbf{w}^T\boldsymbol{\phi}_n - N/N_1)^2 + \frac{1}{2}\sum_{\boldsymbol{\phi}_n \in C_2}(\mathbf{w}^T\boldsymbol{\phi}_n + N/N_2)^2$$

取对$\mathbf{w}$的梯度为零，有:

$$0 = \sum_{\boldsymbol{\phi}_n \in C_1}(\mathbf{w}^T\boldsymbol{\phi}_n - N/N_1)\boldsymbol{\phi}_n + \sum_{\boldsymbol{\phi}_n \in C_2}(\mathbf{w}^T\mathbf{x}_n + N/N_2)\boldsymbol{\phi}_n$$

整理可得：

$$(S_W + \frac{N_1 N_2}{N}S_B)\mathbf{w} = N(\mathbf{m}_1 - \mathbf{m}_2)$$

由式 1.12可知，$S_B\mathbf{w}$ 与$\mathbf{m}_1 - \mathbf{m}_2$ 同向，这说明：

$$\mathbf{w} \propto S_w^T(\mathbf{m_2} - \mathbf{m_1})$$

而这正是公式 1.13所示的Fisher区分函数的解!

让我们梳理一下上面的推理逻辑：我们希望做一个分类函数，使得不同类之间的区分性最大，为此我们定义了一个以$\mathbf{w}$为参数的线性映射，并定义了Fisher准则来优化$\mathbf{w}$。在这一过程中，我们并没有过多考虑为什么选择Fisher 准则，只是直觉上觉得这一准则定义了一个合理的区分性标准（这本身没什么错误，事实上机器学习里的不少任务中的目标函数都是这么通过‘ 感觉’定义出来的）。然而，经过推导，我们发现在上述二分类问题中，依Fisher准则得到的映射函数和基于最小方差准则的线性拟合方法得到的拟合函数是等价的。从上节的讨论我们已经知道，线性拟合等价于目标t 为高斯分布的线性回归，这说明Fisher方法事实上假设了分类任务中的类别标记是高斯的。这显然是不太合理的。

这一高斯分布假设带来分类界面上的偏差。如图 1.3，其中两类数据点分别标为1和0，均值分别在m_1和m_2。正常情况下，每一类在各自均值附近分布。依线性拟合（等价于Fisher准则）优化拟合函数时，学习目标是各样本

点到拟合直线的距离平方和最小，得到的分类面如绿线所示。当出现一些奇异点时（标为红色），为了照顾这些分类点，拟合直线必须非常接近水平（红线）才能得到要求，这使得y对数据变动不敏感，导致分类面失去意义。

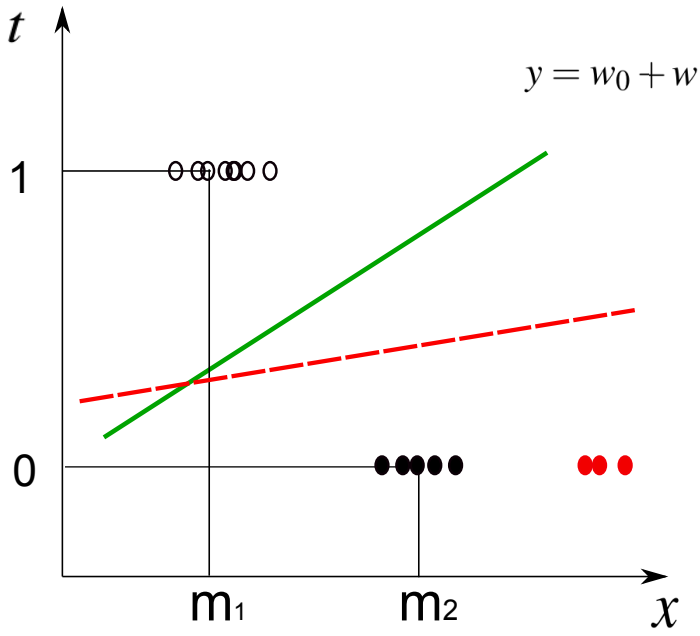

Fig. 1.3 Fisher准则假设目标值的高斯分布。当

1.1.4 Logistic 回归

基于线性回归的分类方法之所以产生对奇异点（Outlier）的敏感，一个直观的原因是奇异点离拟合直线太远，基于最小平方误差准则使得这些奇异点的影响过大。一个可能的解决方法是利用非线性函数对距离进行压缩。同时，由于同类样本点的目标值都是固定的，也就是说样本目标值的取值离散，高斯分布的假设显然不合理。Logistic回归即从这两方面对线性回归进行修正，使之适应分类问题。

同样以二分类问题展开讨论。给定一个包括N个样本的训练集$D = \{(\boldsymbol{\phi}_n, t_n); \boldsymbol{\phi}_n = \boldsymbol{\phi}(\mathbf{x}_n), t_n \in \{0,1\}\}_{n=1}^{N}$，其中$t_n$取不同值代表不同类。不失一般

性，可假设1代表C_1 类、0 代表C_2 类。Logistic 回归假设t符合如下伯努利分布：

$$P(t|\boldsymbol{\phi};\mathbf{w}) = y(\boldsymbol{\phi};\mathbf{w})^t(1-y(\boldsymbol{\phi};\mathbf{w}))^{1-t} \tag{1.14}$$

其中$y(\boldsymbol{\phi};\mathbf{w})$是对$\boldsymbol{\phi}$属于$C_1$的预测函数，定义为：

$$y(\boldsymbol{\phi};\mathbf{w}) = p(C_1|\boldsymbol{\phi};\mathbf{w}) = \sigma(\mathbf{w}^T\boldsymbol{\phi}) \tag{1.15}$$

其中$\sigma(\cdot)$称为Sigmoid函数，定义为：

$$\sigma(a) = \frac{1}{1+e^{-a}}$$

注意$\sigma(\cdot)$将实数域映射到$(0,1)$，可起到非线性压缩作用。Sgimoid函数在机器学习里应用非常广泛，我们在后续章节会陆续看到。在继续讨论之前，我们观察一下预测函数 1.15。我们发现这一形式非常接近线性，只不过在线性预测结果后加入了一个非线性压缩函数。基于Sigmoid 函数的简单性，我们会看到这一非线性并未对学习和推理产生过多阻障，因此我们‘近似’认为这一模型依然是一种线性模型。

和线性回归一样，公式 1.15定义了一个生成过程：首先对输入$\mathbf{x}$经过一个非线性映射$\boldsymbol{\phi}(\cdot)$生成特征向量，经由一个线性函数$\mathbf{w}^T\boldsymbol{\phi}$投影到一个标量空间，再经过$\sigma(\cdot)$压缩到$(0,1)$之间，最后把该压缩值作为伯努利分布生成目标$t$。这一模型称为Logistic回归模型。比较Logistic模型和线性回归模型，可见二者具有相似性，差别只是一个非线性映射函数$\sigma(\cdot)$和不同的分布假设。

基于上述回归模型，我们可以用最大似然估计来优化模型参数$\mathbf{w}$。首先，定义y_n为：

$$y_n = \sigma(\mathbf{w}^T\boldsymbol{\phi}_n) = p(C_1|\boldsymbol{\phi}_n)$$

依公式 1.14，一个样本点$(\mathbf{x}_n, t_n)$ 的概率可形式化写成如下形式：

$$p(t_n|\boldsymbol{\phi}_n,\mathbf{w}) = y_n^{t_n}\{1-y_n\}^{1-t_n}$$

则在数据集D上的似然函数可以表示为：

$$p(D;\mathbf{w}) = \prod_{n=1}^{N} y_n^{t_n}\{1-y_n\}^{1-t_n}$$

其中，$\mathbf{t} = (t_1, \ldots, t_N)^T$。为计算方便，取上述似然函数的负对数作为优化目标：

$$E(\mathbf{w}) = -\ln p(D;\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$

这一目标函数称为交叉熵。

对上述交叉熵函数取梯度为零，并利用关系$\sigma'(a) = \sigma(a)(1-\sigma(a))$，整理后可得:

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\boldsymbol{\phi}_n \tag{1.16}$$

这意味着交叉熵函数对$\mathbf{w}$的梯度取决于预测值和目标值之间的误差。类似的形式在线性回归里也出现过，见式 1.7。注意的是，将上述梯度取零并不能得到$\mathbf{w}$，因为其中y_n是$\mathbf{w}$的函数。但式 1.16中给出的梯度计算方法已经足够我们采用梯度下降法来对$\mathbf{w}$逐步求精了。

梯度下降法（Gradient Descend, GD）是一种通用的函数优化方法。设有函数$f(\mathbf{x})$，优化的目标是找到一个$\mathbf{x}^*$使得$\mathbf{x}^*$最小化。梯度下降法从一个随机的$\mathbf{x}$开始进行迭代优化，每一步t选择一个使$f(\mathbf{x})$下降最大方向，并往该方向前进步长α_t。因为使$f(\mathbf{x}_t)$下降最大的方向即是$f(\mathbf{x})$在$\mathbf{x}_t$ 点的梯度方向，因此该方法称为梯度下降法。如果步长α_t选的合理，梯度下降法可以保证收敛到局部最优。

利用GD对Logistic回归中的交叉熵函数$E(\mathbf{w})$进行优化，相应迭代更新公式如下：

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \alpha_t \nabla_{\mathbf{w}} E(\mathbf{w})$$

一般来说，α_t的取值最初比较大，随着迭代的进行可以逐渐减小，直到收敛到局部最优。

图 1.4给出分别基于线性回归（Fisher 区分函数）和Logistic 回归的在两组二维数据上生成的分类面。当数据分布比较正常，没有奇异值时，可以看到这两种分类方法效果很相似，但当某一类中出现奇异数据时，Logistic 回归受到的影响要小的多。

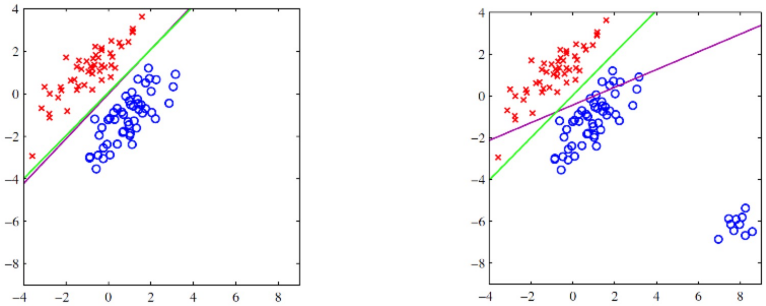

Fig. 1.4 绿色直线和紫色直线分别是Logistic回归和线性回归模型对两类二维数据生成的分类面。左图中的两组数据的分布都比较集中，这两个模型给出的结果相差不大；右图中出现了一些离分类面较远的奇异样本，此时线性回归模型的表现明显变差，而Logistic回归模型受到的影响要小的多。

1.1.5 Softmax回归

同样的方法可应用到多分类问题上，称为Softmax回归。基于多分类问题的性质，Softmax回归将目标t由二分类问题中的伯努利分布扩展到Nomial分布，相应的表示方法也由0-1表示扩展为One-hot表示。设类别为K，则将$\boldsymbol{\phi}_n$的目标写成K 维二值向量$\mathbf{t}_n$，其中$\boldsymbol{\phi}_n$所属类别对应的维度为1，其余维度为0。基于这一表示，$(\boldsymbol{\phi}_n, \mathbf{t}_n)$的概率可写成：

$$p(\boldsymbol{\phi}_n, \mathbf{t}_n) = \prod_i y_k(\boldsymbol{\phi}_n; \mathbf{w})^{t_{nk}} \tag{1.17}$$

其中$y_k(\boldsymbol{\phi}_n; \mathbf{w})$是$\boldsymbol{\phi}_n$属于第$k$类的概率, t_{nk} 是$\mathbf{t}_n$在第k维的取值。注意上式是以$(y_1, y_2, ..., y_k)$为参数的Nomial分布。Softmax 回归定义一个近似线性的概率预测模型如下：

$$y_k(\boldsymbol{\phi}_n; \mathbf{w}) = p(C_k|\boldsymbol{\phi}_n; \mathbf{w}) = \frac{e^{\mathbf{w}_k^T \boldsymbol{\phi}_n}}{\sum_j e^{\mathbf{w}_j^T \boldsymbol{\phi}_n}} \quad k = 1, 2, ..., K$$

其中$\mathbf{w}_k$是对第k类数据进行预测的参数。上式最右侧公式称为Softmax函数。和Logistic函数类似，Softmax函数也是一种非线性归一函数，可将多个类的线性预测输出归一化为概率输出。

下面我们用最大似然估计求解参数$\mathbf{w}$。由单样本概率公式1.17可知在全部训练集上的概率为：

$$p(D;\mathbf{w}) = \prod_{n=1}^{N}\prod_{k=1}^{K} p(C_k|\boldsymbol{\phi}_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

其中$y_{nk} = y_k(\boldsymbol{\phi}_n)$。对上式取负对数得到多分类问题的交叉熵误差函数：

$$E(\mathbf{w}) = -\ln p(D;\mathbf{w}) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln y_{nk} \tag{1.18}$$

取上述误差函数对$\mathbf{w}$的梯度。注意到softmax函数的导数有如下形式：

$$\frac{\partial y_k}{\partial a_j} = \begin{cases} y_k(1-y_j) & j = k \\ -y_k y_j & j \neq k \end{cases}$$

其中$a_j = (w)_j^T\boldsymbol{\phi}$为对$\boldsymbol{\phi}$目标的线性预测，可得到$E(\mathbf{w})$ 对$\mathbf{w}$的梯度如下：

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}) = \sum_{n=1}^{N} (y_{nj} - t_{nj})\boldsymbol{\phi}_n$$

这一梯度公式和线性回归及Logistic回归的梯度公式具有同一格式，都意味着这样一个事实：误差函数的梯度之所以产生，是因为训练样本的预测值和目标值之间存在误差，并且这些误差不能通过数据平均完全抵消。和Logistic回归一样，我们要依赖梯度下降算法实现对交叉熵误差函数的最小化。

至此我们已经介绍了最基本的线性模型。这些模型虽然简单，却是我们分析更复杂模型的基础。我们特别强调模型的概率意义。传统基于朴素优化准则（如平方误差和Fisher准则）的模型方法是直观的，但在一定程度上却也是浅显的。引入概率意义以后，我们才发现隐藏在各种转换函数和优化准则背后的分布假设。认识这些假设对我们理解某种方法的优点和劣势具有重要意义。例如在图 1.4所示的分类例子中，Fisher准则之所以在非规则数据分布上性能下降，本质上是因为这种方法对数据做了高斯假设，这种假设与实际数据偏离的越远，模型的性能就下降的越明显。当我们及时将高斯假设换成了伯努利假设时，这种偏差就被纠正了。如果我们不理解线性回归背后的概率模型，可能很难发现Fisher准则背后的这些问题，更难发现对其进行修正的办法。从另一个角度讲，这也提示我们在学习某种模型和算法时，理解其基础假设要远比推导优化公式重要– 幸运的是，也容易的多。

我们还要特别强调一下本节中介绍的最大似然（ML）准则。上面我们谈到各种对应关系、数据分布假设等，到目前为止都是基于最大似然准则，即模型对训练数据的概率最大化。最大似然显然是一种非常理性的选择，但并不完美。例如，我们经常希望将知识和数据结合在一起，这时就不仅要考

虑训练数据概率最大，还要考虑先验知识在目标函数中的比重，这时的准则就不再是最大似然，而是最大后验（Maximum A Posterior，MAP）。事实上，选择哪种学习准则本身也许并不重要，重要的是面对不同任务时选择恰当的准则，并根据数据的分布情况选择恰当的模型。

1.2 线性概率模型

前一节讨论的线性预测模型假设输入变量和输出变量是可见的，基于输入变量对目标变量进行预测。这事实上是建立一个条件概率模型$p(t|\mathbf{x})$。现在我们考虑一个反向问题：如果给我们一个t，根据$t=\mathbf{w}^T\mathbf{x}+\varepsilon$是否可以推理出$\mathbf{x}$呢？这一问题并不简单。假设$\mathbf{x}$是可见变量，则可基于$\mathbf{w}^T\mathbf{x}=t+\varepsilon$ 对$\mathbf{x}$进行估计。然而，在多数推理任务中$\mathbf{x}$是不可见的，即隐变量（Latent Variable），这时不论是对$\mathbf{w}$ 的估计还是对$\mathbf{x}$ 的推理都不再直观。为解决这类推理问题，我们通常假设$\mathbf{x}$符合某一先验概率，再基于最大后验概率$p(\mathbf{x}|t)$对$\mathbf{x}$其进行推理。

如果将上述问题形式化，将一元观察变量t扩展到多元变量$\mathbf{t}$，将隐变量$\boldsymbol{\varepsilon}$和$\mathbf{x}$写在一起，则可定义$\mathbf{x}$和$\mathbf{t}$之间具有如下线性关系：

$$\mathbf{t}=W\mathbf{x}$$

其中W是参数矩阵。注意，这里我们没有考虑$\mathbf{x}$上的变换$\boldsymbol{\phi}(\cdot)$，因为$\mathbf{x}$是被推理的隐含变量，加入变换将改变模型的性质。这一模型称为线性概率模型。注意线性概率模型和线性预测模型具有类似的形式，唯一的区别是在该模型中$\mathbf{x}$是不可见的。然而，这一区别并不是微不足道的：$\mathbf{x}$变成隐变量以后，我们能观察到的数据只有$\mathbf{t}$，因此学习方式由有监督学习变成了无监督学习。

线性概率模型建立了一种对观察变量$\mathbf{t}$的概率描述框架，在这一框架下，每当观察到一个$\mathbf{t}$时，我们可以依定义好的线性关系去推理数据背后的隐藏原因$\mathbf{x}$。因此，该模型被广泛用在因子分析和结构发现等领域。在后面章节我们会看到，线性概率模型是概率图模型的简单形式。我们将这一模型放到本章讨论，目的是让读者从本书开始即建立这样一种概念：很多看似很不相同的方法事实上具有天然的内在联系，即使是线性拟合和因子分析这两种很不相同的方法，从某种角度看也只是一点点模型假设上的改变。

1.2.1 主成分分析

本节我们从大家所熟知的主成分分析（Principal Component Analysis，PCA）开始讨论。通过讨论，我们将发现PCA是一种简单的线性概率模型，由此可更深入理解这一方法背后对数据分布、目标函数的假设，并讨论将这一传统方法进行扩展的可能。

传统的PCA希望通过找到若干相互相交的方向，使得原数据在这些方向上的映射最大可能地代表原数据的分布性质。每个方向称为一个主成分（Principle Component），依对数据的代表能力排序，称为第一主成份，第二主成份...一般来说，我们希望找到最具有代表性的几个主成份来代表数据，主成份的个数一般远小于数据维度，因此PCA一般多用在数据降维中。

有哪两种方法衡量主成份对原数据的代表性呢：一是使数据在主成分这个映射空间的方差尽可能大，既能够更好地对原数据作出区分；二是由映射恢复成原始数据时的损失最小（事实上这两个准则是等价的）。下面我们从映射空间方差最大这一准则推导出PCA基本公式，该准则意味着我们希望经过主成份投影后尽量保持原数据的分散性。

我们从寻找第一主成份开始。设包括N个样本的训练数据集为$\mathscr{D} = \{\mathbf{x}_i \in R^D\}_{i=1}^N$，我们的目的是要寻找一个方向$\mathbf{v}_1 \in R^D$，使得数据集$\mathscr{D}$在这一方向上的映射方差最大。不失一般性，我们令$\mathbf{v}_1$是单位向量，即$\mathbf{v}_1^T\mathbf{v}_1 = 1$。数据集在映射空间的方差计算如下：

$$\begin{aligned} L(\mathbf{v}_1) &= \frac{1}{N}\sum_{i=1}^{N}(\mathbf{v}_1^T\mathbf{x}_i - \mathbf{v}_1^T\overline{\mathbf{x}})^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}\mathbf{v}_1^T(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T\mathbf{v}_1 \\ &= \mathbf{v}_1^T\{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T\}\mathbf{v}_1 \\ &= \mathbf{v}_1^T S\mathbf{v}_1 \end{aligned} \tag{1.19}$$

其中$\overline{\mathbf{x}}$为训练样本集的均值：

$$\overline{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{n}\mathbf{x}_i$$

S是数据集的协方差矩阵。我们的目的是求使$L(\mathbf{v}_1)$最小化的$\mathbf{v}_1$, 且满足约束条件$\mathbf{v}_1^T\mathbf{v}_1 = 1$。基于拉格朗日乘子法，上述带限制优化任务等价于如下无限制优化任务：

$$\Psi(\mathbf{v}_1) = L(\mathbf{v}_1) + \lambda_1(1 - \mathbf{v}_1^T\mathbf{v}_1) = \mathbf{v}_1^T S\mathbf{v}_1 + \lambda(1 - \mathbf{v}_1^T\mathbf{v}_1) \tag{1.20}$$

求上式对v_1的梯度：

$$\nabla\Psi(\mathbf{v}_1) = 2S\mathbf{v}_1 - 2\lambda_1\mathbf{v}_1 = 0$$

整理可得下式:

$$S\mathbf{v}_1 = \lambda\mathbf{v}_1$$

上式说明满足优化条件的主成分方向$\mathbf{v}_1$必然是协方差矩阵S的特征向量。将上式代入优化目标函数，有：

$$L(\mathbf{v}_1) = \mathbf{v}_1^T S\mathbf{v}_1 = \lambda_1$$

上式说明数据集在$\mathbf{v}_1$上映射的方差等于$\mathbf{v}_1$ 对应的特征向量λ的大小。因此，为取使$L(\mathbf{v}_1)$最大的主成份，只需取协方差矩阵的最大特征值所对应的特征向量即可。

取得第一主成份后，依类似步骤可取后续各主成份。设已经得到前$k-1$个主成份，我们来寻找第k个主成份$\mathbf{v}_k$，使得数据在$\mathbf{v}_k$上的映射方差最大，且$\mathbf{v}_k$与前$k-1$个主成份正交。基于这些目标和条件，得到类似式 1.20的目标函数：

$$\Psi(\mathbf{v}_k) = \mathbf{v}_k^T S\mathbf{v}_k + \lambda_k(1 - \mathbf{v}_k^T\mathbf{v}_k) + \sum_{i=1}^{k-1}\lambda_i\mathbf{v}_k^T v_i$$

取对$\mathbf{v}_k$的梯度：

$$\nabla\Psi(\mathbf{v}) = 2S\mathbf{v}_k - 2\lambda_k\mathbf{v}_k + \sum_{i=1}^{k-1}\lambda_i\mathbf{v}_i = 0$$

将上式左右各左乘$\mathbf{v}_i^T(\forall i < k)$，依正交性可得$\lambda_i = 0(\forall i < k)$，因此得到和求第一主成份完全一样的形式，即$\mathbf{v}_k$为协方差的特征向量，对应的特征向量为$\lambda_k$。由此可知，所有主成份都是协方差的特征向量，且由第一个主成份起，都会选择第k大的特征值对应的特征向量。因此，想要求前K 个主成份，只需选择最大K个特征值对应的特征向量即可。

1.2.2 概率主成分分析

PCA是非常经典的无监督学习方法，被广泛应用在降维、正规化、流形学习中。然而，PCA的优化函数（映射空间方差最大或恢复误差最小）和主成分之间的正交限制很大程度上是种人为定义，这使得PCA的适用性缺少明确解释。本节我们将寻求PCA的概率意义，用线性概率模型来解释PCA，这一方法称为概率主成分分析（probabilistic PCA，PPCA）。

我们考虑如下简单的线性概率模型：

$$\mathbf{t} = \mu + W\mathbf{x} + \varepsilon \tag{1.21}$$

其中$\mathbf{t} \in R^D$是D维观察变量，$\mu \in R^D$是固定偏移量，W是模型参数。$\mathbf{x} \in R^M$是符合正态分布的M维隐含变量:

$$p(\mathbf{x}) = N(\mathbf{x}|\mathbf{0}, \mathbf{I})$$

$\varepsilon \in R^D$ 是D维高斯变量：

$$p(\varepsilon) = N(\mathbf{x}|\mathbf{0}, \sigma^2 I)$$

上式定义了一个数据$\mathbf{t}$的生成模型：首先基于先验概率$p(\mathbf{x})$ 生成采样点$\mathbf{x}$，之后将该采样通过线性变换变为$W\mathbf{x}$，第三步将变换后的采样点位移μ，最后加入一个高斯噪音ε。这一过程如图 1.5所示。值得一提的是，这一模型中所有变量都是高斯的：变量$\mathbf{x}$是一个高斯变量，其线性变换$\mu + W\mathbf{x}$也是高斯的，再加入一个高斯噪音依然是高斯的。因此，实际观察到的数据$\mathbf{t}$是一个高斯分布，如图 1.5中图（c）所示。这种基于高斯分布的线性概率模型通常称为线性高斯模型。我们本节所述的模型都属于这一类型。

仔细观察PPCA的模型形式 1.21，可以看到这一方程和线性拟合十分相似，但$mathbfx$不论在训练还是推理是都是隐藏且随机的，我们所能知道的只是一个假设的先验概率$p(\mathbf{x})$，这一差别带来某些根本性的改变，不论是训练方法、推理方法和应用场合。另外，注意隐变量不一定是随机的。例如模型中的μ也是不可见的，但这一变量本身不是随机的，因此只是模型参数，并不会使模型更复杂。

下面我们用最大似然估计PPCA的模型参数W，$\boldsymbol{\mu}$和σ^2。为了写出似然函数，我们需要知道观测变量的边缘分布$p(\mathbf{t})$的表达式。基于PPCA的模型假设，有：

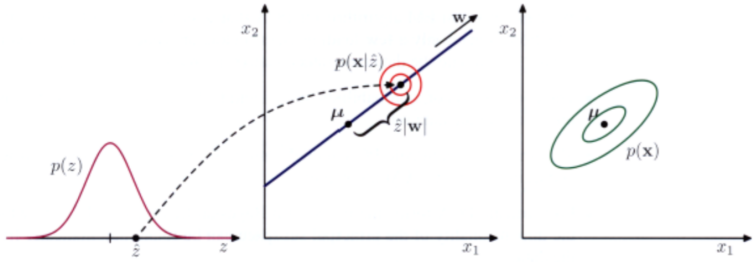

Fig. 1.5 PPCA模型，其中观察数据$\mathbf{t}$ 是二维，隐变量x是一维。首先根据隐变量x的先验分布$p(x)$得到一个样本点$\hat{x}$，之后对该样本点做线性变换$W\hat{x}+\boldsymbol{\mu}$，最后加入均值为0、协方差为$\sigma^2\mathbf{I}$高斯噪声（图中的红色圆圈），得到观察值$\mathbf{t}$的一个取值。绿色的椭圆表示边缘分布$p(\mathbf{t})$的密度轮廓线。

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}$$

因为$\mathbf{x}$和$\boldsymbol{\varepsilon}$都服从高斯分布，$\mathbf{t}$也服从高斯分布，其均值可由期望得到：

$$\mathbb{E}[\mathbf{t}] = \mathbb{E}[\boldsymbol{\mu} + W\mathbf{x} + \boldsymbol{\varepsilon}] = \boldsymbol{\mu} \tag{1.22}$$

其中利用了$\mathbf{x}$和$\boldsymbol{\varepsilon}$的均值为零的事实。类似，方差可由$\mathbf{t}$的协方差矩阵得到：

$$\begin{aligned}\mathrm{cov}(\mathbf{t}) &= \mathbb{E}[(W\mathbf{x}+\boldsymbol{\varepsilon})(W\mathbf{x}+\boldsymbol{\varepsilon})^T] \\ &= \mathbb{E}[W\mathbf{x}\mathbf{x}^T W^T] + \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] \\ &= WW^T + \sigma^2 I\end{aligned} \tag{1.23}$$

由此可得：

$$p(\mathbf{t}) = N(\mathbf{t}|\boldsymbol{\mu}, C) = N(\mathbf{t}|\boldsymbol{\mu}, WW^T + \sigma^2 I)$$

接下来，我们基于最大似然准则来确定模型的参数。给定N个观测点的数据集$D = \{\mathbf{t}_i\}_{i=1}^N$，对应的对数似然函数为：

$$\begin{aligned}\ln p(D|\boldsymbol{\mu}, W, \sigma^2) &= \sum_{i=1}^{N} \ln p(\mathbf{t}_i|\boldsymbol{\mu}, W, \sigma^2) \\ &= -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln(|C|) - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{t}_n - \boldsymbol{\mu})^T C^{-1}(\mathbf{t}_n - \boldsymbol{\mu})\end{aligned}$$

Tipping 和Bishop 证明 [7]，对上述似然函数进行优化，可得：

$$\boldsymbol{\mu}_{ML} = \bar{\mathbf{x}}$$

$$W_{ML} = U_M(L_M - \sigma^2 I)^{1/2} R$$

$$\sigma^2_{ML} = \frac{1}{D-M}\sum_{i=M+1}^{D} \lambda_i$$

其中，$U_M \in R^{D\times M}$由原始数据协方差矩阵S前M个最大特征值对应的特征向量排列而成，$L_M \in R^{M\times M}$是对应的特征值组成的对角矩阵，$R \in R^{M\times M}$是任意正交矩阵。注意正交矩阵R是任意的，不会对数据分布$p(\mathbf{t})$产生影响，这一任意性来源于先验概率$p(\mathbf{x})$的各向同性。

基于上述生成模型，我们来确定从观察变量$\mathbf{t}$到隐变量$\mathbf{x}$ 的映射。在概率模型下，这一映射可表示成后验概率$p(\mathbf{x}|\mathbf{t})$。由于先验概率$p(\mathbf{x})$和条件概率$p(\mathbf{t}|\mathbf{x})$都是高斯的，可以确定该后验概率亦具有高斯分布形式。依贝叶斯定理，可得：

$$p(\mathbf{x}|\mathbf{t}) = \frac{p(\mathbf{t}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{t})} = N(\mathbf{x}|M^{-1}W^T_{ML}(\mathbf{t}-\boldsymbol{\mu}_{ML}), \sigma^{-2}_{ML}M)$$

$$M = W^T_{ML}W_{ML} + \sigma^2_{ML}I$$

基于该后验概率，对给定$\mathbf{t}$，其在隐变量空间映射为该后验概率取大值所对应的$\mathbf{x}$，这一方法称为最大后验估计(MAP)。由于该后验概率是高斯的，MAP估计等同于均值，因而有：

$$\mathbf{x}_{MAP}|\mathbf{t} = M^{-1}W^T_{ML}(\mathbf{t}-\boldsymbol{\mu}) = (W^T_{ML}W_{ML} + \sigma^2_{ML}I)^{-1}W^T_{ML}(\mathbf{t}-\boldsymbol{\mu}_{ML}) \tag{1.24}$$

当$\sigma_{ML} \to 0$时，可推导出$\mathbf{x}$的MAP估计:

$$\begin{aligned}\mathbf{x}_{MAP}|\mathbf{t} &= (W^T_{ML}W_{ML})^{-1}W^T_{ML}(\mathbf{t}-\boldsymbol{\mu}_{ML}) \\ &= R^T L^{-1/2}_{ML}U^T_{ML}(\mathbf{t}-\boldsymbol{\mu}_{ML})\end{aligned}$$

我们已知U_{ML}由协方差矩阵的前M个特征向量组成，这意味着PPCA 在对原始数据$\mathbf{t}$映射过程中，首先将其映射到和传统PCA同样的子空间，再通过对角阵$L^{-1/2}_{ML}$进行尺度调整以便得到的$\mathbf{x}$具有各向同性。这说明一等价性提示我们，PCA事实上假设了一个线性高斯模型，观察数据由一个简单的各向同性的

正态分布经过一个线性变换得到，因此$\mathbf{t}$也是高斯的。当这一条件不满足时，PCA的适用性将会降低。例如，当数据$\mathbf{t}$ 明显不符合高斯分布时，PCA的结果可能会产生较大偏差。

1.2.3 概率线性判别分析

PCA描述数据的分布特性，并不考虑不同类数据的不同分布。我们对该模型做一点扩展，使其可以在对有类别标记的数据进行建模。为了表达的更清楚，我们将PPCA的模型公式 1.21重写如下：

$$\mathbf{t} = \mu + W\mathbf{x} + \varepsilon$$

$$x \sim N(0, I),\ \ \varepsilon \sim N(0, \sigma^2 I)$$

其中每个采样t的生成过程都是一样的：首先从高斯分布中得到x，再经过一个线性映射$\mu + W\mathbf{x}$映射到高维数据空间，再加上一个高斯噪声ε。基于最大似然准则，可以对参数$\{\mu, W, \sigma^2\}$ 进行估计。
注意上述模型中的t没有区分类别，因此是一个标准描述型模型（即对数据的分布属性进行描述）。现在考虑t属于不同类，每一类C_i数据的均值μ_i各不相同，则上述生成模型可写成：

$$\mathbf{t} = \mu_i + W\mathbf{x} + \varepsilon \tag{1.25}$$

$$x \sim N(0, I),\ \ \varepsilon \sim N(0, \sigma^2 I)$$

同样基于最大似然准则，可对参数$\{\mu_i, W, \sigma^2\}$进行估计。注意在上式中，不同类别的数据仅是均值不同，但方差共享。这一模型事实上等价于前一节讨论过的线性区分性分析（LDA） [4]。因此，LDA可以认为是一个多类生成模型，其中每一类数据用一个线性高斯模型表示，且所有类共享一个线性映射矩阵W，意味着所有类的协方差矩阵都是$W^T W + \sigma^2$。

上述模型中每一类的中心μ_i都是一个模型参数，因此模型是和数据相关的，参数数量随类别增多而增长，这种模型一般称为非参模型（No-Parametric Model）。这一模型有两个严重问题：一是如果数据中包含的类很多，则不仅计算开销增加，对数据的利用也不够充分（例如对那些样本很小的类，基于该模型得到的μ_i估计会产生偏差）;二是无法处理在测试时遇到新

类，因为这些类没在训练数据出现过，模型对这些数据的描述能力有限，得到的区分性方向有很大偏差。

概率LDA（Probabilistic LDA，PLDA）将μ_i看作一个随机变量（而非模型参数）完美解决了上述问题。在PLDA中，首先由一个先验概率采样出某一类（设为C_i）的均值μ_i，再由一个线性高斯模$\mu_i + W\mathbf{x} + \varepsilon$生成该类的所有采样。在实际应用中，类中心$\mu_i$通常代表某种物理性质（如人脸、声音等），因此通常限制在一个低维空间上。由此，PLDA可形式化成如下公式：

$$\mathbf{t} = \mu + F\mathbf{u}_i + W\mathbf{x}_{i,j} + \varepsilon \tag{1.26}$$

$$u_i \sim N(0,I),\ x \sim N(0,I),\ \varepsilon \sim N(0,\sigma^2 I).$$

注意上式中的u和x是不同于低维空间的两个随机变量，其中u代表类间差异，x代表类内差异。基于该模型，$\mathbf{u}_i$不再是参数，而是隐变量，模型参数不再和训练样本中的类别数相关，因此是一个严格的生成模型。给定一个测试数据，基于上述生成模型可推理得到后验概率$p(\mu|t)$，该后验概率代表t的类别属性，因而可用作分类的特征向量。注意这一推理与数据的类别没有直接关系，因为在模型中没有和类别相关的参数，所有参数对所有类都是共享的。图 1.6给出PLDA模型的生成过程：首先由一个高斯随机变量依一维正态分布生成一个类中心u_i，经过线性映射$u_i' = Fu_i + \mu$在二维空间上生成类i的中心变量。基于此中心，依PPCA 的生成过程得到该类的所有样本数据。具体来说，首先由一个高斯随机变量依一维正态分布生成随机数x，经过一个线性变换W映射到二维空间，得到Wx，加入高斯随机变量ε，得到的采样点加上类中心u_i'，即可得到该类的一个采样数据。这一采样过程重复进行, 即可得到该类的数据分布。注意上述对类i的数据生成过程基于一个固定的类中心u_i，基于此类中心，每一类的数据分布是一个高斯分布。

上述PLDA模型的参数$\{\mu, F, W, \sigma\}$可基于最大似然准则进行优化。直接求似然函数需要对所有隐变量边缘化（Marginalize），计算上比较困难，可采用期望最大化（Expectation Maximization, EM）算法通过迭代方式求解。该方法是一个迭代优化过程：首先对参数进行初始化，利用当前参数求u,x 的后验概率，再基于该后验概率求似然函数的期望（E步），再基于该期望函数对参数进行优化（M步）。上述E步和M步交替迭代进行，直到收敛。EM算法是解决包含隐变量的概率模型的基本方法，可以证明该算法总会收敛到局部最优。

对比PLDA和PPCA，可以发现他们在形式上非常相似，都是线性高斯模型; 事实上，对于某一个特定的类C_i，PLDA事实上是一个对μ 进行随机

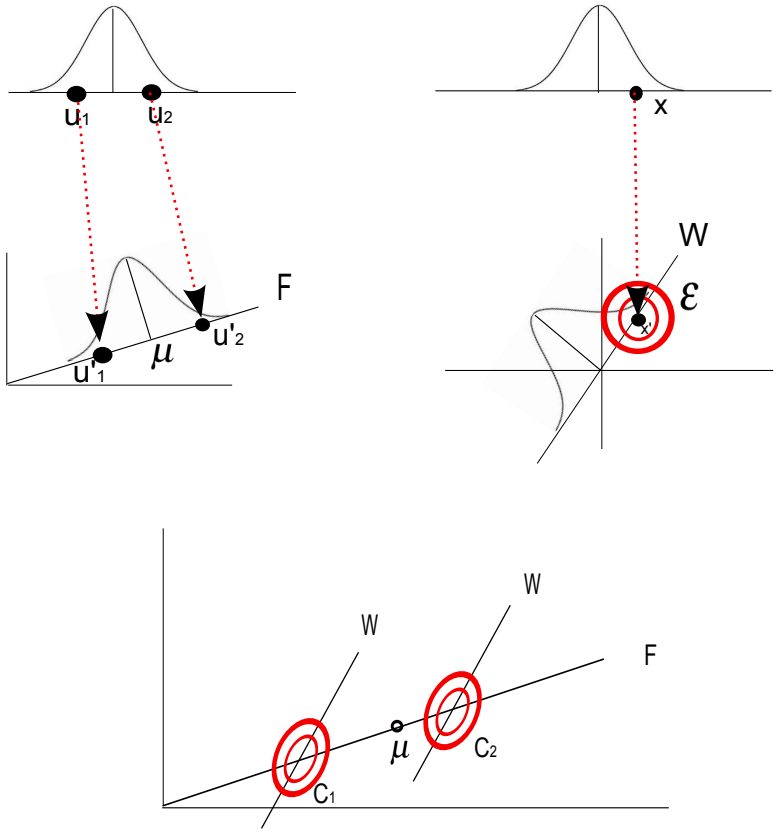

Fig. 1.6 PLDA模型，其中观察数据**t** 是二维，类中心隐变量u和类内隐变量x都是一维。首先根据隐变量u基于正态分布采样得到类i 的中心向量u_i，经过线性变换得到在数据空间中的中心向量u_i'。基于该类中心，对类中每一个数据点，依下述过程采样生成：首先根据正态分布采样得到一维变量x，经过线性变换得到数据空间中的采样x'作为相对类中心u_i'的偏移量，最后加入高斯噪声ε 后，即得到该类的一个采样点。

化的PPCA。从采样角度看，PLDA和PPCA的不同只是在PPCA采样之前，先对μ_i进行一次采样。二者的优化方法是一样的，都基于最大似然准则。上述相似性让我们从另一个角度认识PCA和LDA这两种看似完全不同的机器学习工具：一个是描述任务，一个是分类任务，一个是非监督学习，一个是监督学习。如此不同的两个工具，当我们从概率模型角度去考察时，却发现他们事实上是非常相似的模型。这对我们有很大启发：对一个模型和方法不能只看公式，更不能只看实现步骤，更重要的是对其基本假设、基本思路的理

解，有时甚至需要从更通用的模型角度去理解其内涵。唯有如此，我们才有可能对这些纷纭复杂的模型分条缕析，并把他们用在最恰当的地方。

1.3 贝叶斯方法

前面几节我们介绍了几种基础的线性模型，这些线性模型通过引入若干随机变量，不仅对数据的分布情况描述的更加准确，而且可以对若干重要模型（PCA,LDA）进行概率解释。不论哪种模型，一个基本假设是模型中的参数是确定的，因而可用各种优化方法进行求解。这种确定性参数的一个缺点是我们无法对它们引入先验知识。例如，我们已经知道某些参数的大约取值范围，如果考虑这一知识，过训练的风险就会减小。

本节中，我们介绍将介绍贝叶斯方法，该方法将模型参数看作随机变量，将我们对参数取值范围的知识转化成对参数的先验概率。引入先验概率不仅可以使我们将人为先验引入训练过程中，更重要的是对模型性质的改变：对模型的优化不再是寻找一个最优参数（如最大似然估计），而是对该参数的概率推理。因此，即使引入的先验是一个无信息先验（如大范围的平均分布），贝叶斯方法依然有重要价值。

令模型参数为$\mathbf{w}$，先验概率为$p(\mathbf{w})$。基于某一观察数据集D，依贝叶斯定理可重估$\mathbf{w}$的后验概率：

$$p(\mathbf{w}|D) = \frac{p(D|\mathbf{w})p(\mathbf{w})}{p(D)} \tag{1.27}$$

上式中先验概率（prior）$p(\mathbf{w})$可以认为是先验知识，代表没有任何数据的情况下，我们对$\mathbf{w}$取值的假设；似然函数$p(D|\mathbf{w})$是经验知识，表示在给定一组参数$\mathbf{w}$时，可以生成数据D的概率；后验概率$p(\mathbf{w}|D)$ 是对参数$\mathbf{w}$ 的再估计，表示在观察到数据D后，对参数$\mathbf{w}$的概率重估。$p(D)$是归一化因子，同时也是各种可能的$\mathbf{w}$得到的似然函数的期望,这一期望在模型结构选择中有重要意义。

基于后验概率$p(\mathbf{w}|D)$可以选择最优模型参数:

$$\mathbf{w}_{MAP} = argmax_w p(\mathbf{w}|D)$$

这种参数选择方法称为最大后验（MAP）估计。和最大似然估计相比，最大后验估计考虑了$\mathbf{w}$的先验知识，因此当经验数据较小时可获得更好的估计。

当数据量增大时，先验占的比重越来越小，最大后验估计趋近于最大似然估计。

最大后验估计依然是点估计，即选择某一确定的参数建模再进行预测和推理。事实上，后验概率$p(\mathbf{w}|D)$提供了一种更有价值的预测和推理方式，即在预测或推理时考虑所有可能的参数$\mathbf{w}$，这样得到的结果会更加可靠。这一方法称为贝叶斯方法。以预测任务为例，贝叶斯方法可写成下式：

$$t = \int t p(t|\mathbf{w},\mathbf{x}) p(\mathbf{w}|D) \mathrm{d}\mathbf{w}$$

下面我们以第一节中介绍的线性回归模型为例来说明贝叶斯方法。与传统线性回归模型类似，定义输入$\mathbf{x}$和输出t之间存在线性关系，但对参数$\mathbf{w}$定义高斯先验：

$$t = \mathbf{w}^T \mathbf{x} + \varepsilon$$

$$\mathbf{w} \sim N(0, \alpha^{-1} I); \quad \varepsilon \sim N(0, \beta^{-}1I)$$

上式等价于：

$$p(t|x, mathbfw, \beta) = N(t|\mathbf{w}^T \mathbf{x}, \beta^{-1} I)$$

$$p(\mathbf{w}|\alpha) = N(\mathbf{w}|0, \alpha^{-1} I) \tag{1.28}$$

对于一个包含N个采样点的数据集$D = \{\mathbf{x}_i, t\}_{i=1}^N$，由后验概率公式:

$$p(\mathbf{w}|D) \propto p(D|\mathbf{w}) p(\mathbf{w}|\alpha)$$

可得:

$$\ln p(\mathbf{w}|D) = \ln \Pi_i p(t_i|\mathbf{x}_i, \mathbf{w})| + \ln p(\mathbf{w}|\alpha) + const \tag{1.29}$$

$$= -\frac{\beta}{2} \sum_i t_i - \mathbf{w}^T \mathbf{x}_i{}^2 - \frac{1}{2} \mathbf{w}^T \mathbf{w} + const \tag{1.30}$$

其中$const$是与w无关的常量。最大后验估计即是取使上式最大的$\mathbf{w}$。有意思的是，上式的形式是一个平方误差加上一个二阶量$\mathbf{w}^T\mathbf{w}$，其中平方误差是标准线性回归模型的目标函数，而二阶量来源于先验概率$p(\mathbf{w})$。如果我们将该二阶量看作原目标函数的正则项（Regularization），则引入先验概率等价于

在原目标函数上引入一个约束，该约束使得优化目标倾向于模更小的参数，而这正是先验概率$p(\mathbf{w})$取最大的位置。不同的先验概率等价于不同的正则约束，这就在概率上解释了传统模式识别方法中正则约束的作用。

注意式 1.28定义的先验概率包含一个参数α，这一参数是参数的先验概率的参数，因此也称超参数。超参数可以依经验设定，也可以基于最大似然准则学习。注意到$p(D) = \int p(D|\mathbf{w})p(\mathbf{w};\alpha)$是$\alpha$的函数，因此可以将$p(D)$作为似然函数优化$\alpha$。也可对$\alpha$引入一个先验概率，再基于最大后验准则求解$\alpha$，这种方法有时称为层次性贝叶斯模型（Hierarchical Bayes Model）。

一般来说，先验概率可以自由选择各种形式，但在实际操作时我们希望后验概率计算越简便越好。一种方法是选择一种先验概率，其与似然函数相乘后得到的后验概率具有和先验概率相同的形式。这一先验概率称为似然函数的共轭先验。共轭先验提供了一个简洁的在线学习框架。在这一框架中，依过去的经验D得到的后验$p(\mathbf{w}|D)$ 成为新的先验，再来新的数据D'可依此先验进一步学习得到新的后验.... ..注意这一学习方法的前提是先验分布必须共轭的，否则后验概率无法得到与先验一致的形式，从而无法形成新的先验。

1.4 本章小结

线性模型是机器学习里最简单、也是最重要的模型之一，是学习其它模型的基础。本章讨论了两种线性模型：一种是输入变量可见的预测模型，一种是输入变量不可见的描述模型。预测模型多用于预测任务，描述模型多用于推理任务；预测模型多用监督学习，描述模型多用于非监督学习。不论哪种模型，都可表示成简单的$\mathbf{y} = W\mathbf{x}$ 线性形式（或稍有变化）。这些模型都对观察值或隐变量做出某种概率分布假设，并依最大似然准则估计模型参数。

在线性预测模型中，我们讨论了用于回归问题的线性回归模型和用于分类问题的Logistic回归模型。线性回归模型假设目标变量是一个条件高斯分布，而Logistic回归模型假设目标变量是一个条件伯努利分布。通过讨论发现，线性回归模型通过最大似然估计得到的回归参数与线性拟合得到的拟合参数是一致的，这意味着这两个模型事实上是等价的，这就为线性拟合找到了合理的概率解释。同时，我们发现传统基于Fisher准则的线性分析模型在二分类问题上也等价于线性回归模型，这相当于对数据的类别标签假设了高斯分布，显然是不合理的。我们介绍了Logistic回归模型，该模型将高斯分布假设修正为伯努利分布分布假设，因此更适合分类问题。

在线性描述模型中，我们讨论了基于非监督学习的概率PCA（PPCA）方法和基于监督学习的概率LDA（PLDA）方法。类似线性拟合与线性回归模型的关系，PPCA和PLDA分别是PCA和LDA的概率模型形式。在PPCA方法中，我们假设观察数据是由一个正态分布的隐随机变量通过一个线性变换再加上一个高斯噪声生成，这意味着PPCA（及其低级形式，PCA）事实上假设了观察数据是一个高斯分布，因此任何不符合高斯分布的数据都不适合用PPCA（及PCA）。PLDA 在PCA基础上考虑类间差异。这一模型假设每个类的均值由低维空间的一个正态分布经过线性变换得到；得到类均值后，通过一个共享的PCA模型生成该类的所有数据。因为数据生成模型是共享的，PPCA中各类的协方差矩阵是相同的，唯一的差别是均值上的不同。和LDA相比，PLDA是将原来作为模型参数的类均值统一成一个随机变量。这一修正具有重要意义：它使得模型参数与数据无关，因此具有更强的泛化能力，可以处理训练数据中没有见过的新类别。

最后，我们讨论了对线性模型的贝叶斯扩展，将原来确定性的参数扩展为随机变量，依据先验知识对这些变量设置先验概率。利用贝叶斯公式，这些先验知识可以和经验知识有效结合起来，得到一种更有效的参数估计方法：最大后验估计。更重要的是，贝叶斯方法改变了我们对模型参数的认识：模型参数不一定是一些确定的数值，还可以是一些随机变量。基于这些随机变量的后验概率估计一般可以做出更合理的预测或推理，因为在预测或推理过程中考虑了所有可能的模型参数配置。

1.5 相关资源

- 本章对线性预测部分的讨论参考了Bishop的《Pattern Recognition and Machine Learning》第三章、第四章对线性模型的描述。对PCA部分的讨论参考了该书的第十二章关于连续隐变量模型的讨论 [2]。
- 对PLDA的讨论参考了Prince等人的论文《Probabilistic Linear Discriminant Analysis for Inferences About Identity》 [5]。
- 线性模型在各种机器学习和模式识别经典教材中都是基础内容。如Hastie, Tibshirani 和Friedman的《The Elements of Statistical Learning》一书的第三、四章 [3]，周志华《机器学习》一书的第三章 [1]。
- 关于线性高斯模型，可参考Roweis等人在1999 年的综述文章 [6]。

References

[1] 周志华(2016) 机器学习. 清华大学出版社

[2] Bishop CM (2006) Pattern recognition. Springer

[3] Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning, vol 1. Springer series in statistics Springer, Berlin

[4] Kumar N, Andreou AG (1998) Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. Speech communication 26(4):283–297

[5] Prince SJ, Elder JH (2007) Probabilistic linear discriminant analysis for inferences about identity. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE, pp 1–8

[6] Roweis S, Ghahramani Z (1999) A unifying review of linear gaussian models. Neural computation 11(2):305–345

[7] Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61(3):611–622