《模式识别与机器学习》读书笔记

zYx.Tom

2020-05-03

Contents

全书总评

书本印刷质量: 4 星

PDF 打印，非影印版，字迹清楚，图片清晰。

- 英文版
  - 排版更利于阅读，请到这里 2006 年版 下载
  - 文中错误需要修正文件，否则会影响理解。
- 中文版
  - 排版有点紧凑，错误较少，但是中文版的数学符号不如英文版的规范
  - 译者已经将英文版中的错误进行了修订。但是，译文本身也存在少量错误，建议与英文版结合起来阅读。

著作编写质量: 4 星

模式识别与机器学习的进阶。

- 全书内容自洽，数学方面不算太深，具备大学工科数学功底 (微积分、线性代数、概率统计) 就可以阅读，如果想深入理解还需要补充随机过程、泛函分析、最优化、信息论等，如果还想更深一层还可以补充决策论、测度论、流形几何等理论；再深俺就完全不知道了。
- 作者以讲清楚 Bayesian 方法的来龙去脉为根本目的，所以全书紧紧围绕在 Bayesian 同志的周围，尽可能以 Bayesian 思想来分析各种模式识别与机器学习中的常用算法，对于已经零散地学习了许多种算法的同学大有裨益。
- 全局的结构是点到面的风格，以一个二项式拟合的例子一点点铺开，节奏稍慢但是前后连贯，知识容易迁移理解；

著作翻译质量: 4 星

马春鹏免费将自己的翻译稿贡献出来，翻译的质量值得肯定，用词符合专业认知，只是译文的流畅性稍感不足 (纯属个人感受)。

- 符号的使用不如原文规范。例如：原文都以加粗的罗马字母表示向量，而译文只是以加粗默认的数字字母表示向量，而在印刷中默认的数学字母加粗和未加粗的区别不明显，容易混淆。精读一段时间后，觉得译者提供的符号系统在自己手写公式的时候比较方便，对于印刷的符号系统依然推荐原作者的风格。

读书建议

前四章是重点，建议先从 1~2 读一遍，再从 1~4 读一遍，对前四章心中有数再看后面的章节会方便很多。

需要手写公式，推不动看不懂不害怕，因为理论推导的细节都被作者跳过了。如果想深入了解，只能去找相关文献。但是如果有了前 4 章的基础，又有很好的机器学习的背景知识，用贝叶斯视角读完全书的可能性还是有的。

需要把公式推导中用到前面的公式的部分抄过来，反复推导前面的公式才能流畅地看后面的内容。

笔记目的：记录重点，方便回忆。

重要的数学符号与公式

建议将下面列出的重要的数学符号与公式找张纸列在上面，方便在后面看到时可以查询。

重要的数学符号

- x $= (x_1, x_2, ..., x_D)^T$：表示 D 维向量
- X $= [\mathrm{x}_1, \mathrm{x}, ..., \mathrm{x}_N]^T$：表示 N 个 D 维向量组成的矩阵
- M：表示矩阵；$(w_1, \cdots, w_M)$：表示一个行向量有 M 个元素；w= ($w_1, \cdots, w_M)^T$ 表示对应的列向量。
- I_M：表示 $M \times M$ 单位阵。
- x：表示元素；$y(x)$：表示函数；$f[y]$：表示泛函。
- $g(x) = O(f(x))$：表示复杂度。
- $\mathbb{E}_x[f(x, y)]$：随机变量 x 对于函数 $f(x, y)$ 的期望，符号可以简化为 $\mathbb{E}[x]$
- $\mathbb{E}_x[f(x)|z]$：随机变量 x 基于变量 z 的条件期望。
- var$[f(x)]$：随机变量 x 的方差
- cov[x, y]：协方差，$cov[\mathbb{x}]$ 是 $cov[\mathbb{x}, \mathbb{x}]$ 的缩写

重要的数学公式

(括号内的是公式的编号)

- 概率论
  - 期望：(1.33) (1.34)
    - 期望估计：(1.35)

        * 条件期望：(1.36) (1.37)
    - 方差：(1.38) (1.39) (1.40)
    - 协方差：(1.41) (1.42)
    - 高斯分布：(1.46) (1.52) (2.42) (2.43)
- 信息论
    - 熵：(1.98)
        * 微分熵：(1.103)
        * 条件熵：(1.111)
        * 联合熵：(1.112)
    - 相对熵，KL 散度：(1.113)
    - 互信息：(1.120) (1.121)
- 概率分布
    - Bernoulli 分布：(2.2)
    - 二项分布：(2.9)
    - Beta 分布：(2.13)
    - 多项式分布：(2.34)
    - Dirichlet 分布：(2.38)
    - Gamma 分布：(2.146)
        * Gamma 函数 (1.141)
    - 学生 t 分布：(2.158)
    - 混合高斯分布：(2.188) (2.193)
    - 指数族分布：(2.194)
        * softmax 函数：(2.213)
        * 共轭先验：(2.229)，(2.230)
- 回归的线性模型
    - 线性基函数模型：(3.1) (3.2) (3.3)
        * 高斯基函数：(3.4)
        * sigmoid 基函数：(3.5)
        * logistic sigmoid 基函数：(3.6)
    - 平方损失函数：(1.90) (3.37)
        * 平方和误差函数：(3.12) (3.26)
            · 正则化最小平方：(3.24) (3.27)
        * 设计矩阵：(3.16)

        * 顺序学习：(3.22) (3.23)
    - 预测分布：(3.57)

- 分类的线性模型
    - 推广的线性模型：(4.3)
    - 神经网络
        * 整体的网络函数：(5.9)

书中符号说明

- (P xx) , 代表第 xx 页；
- (Ch xx) , 代表第 xx 章；
- (Sec xx) , 代表第 xx 节；
- (Eq xx) , 代表第 xx 公式；
- (Fig xx) , 代表第 xx 图

Ch 01. 绪论

本书需要机器学习的基本知识，研读后可以加深对贝叶斯学习的理解。

提纲

- 重点

多项式曲线拟合: 这个例子通过不同角度介绍机器学习的常用算法，从而更好地理解贝叶斯估计的思想，也是后面反复讨论的基础。

- 难点
    - 贝叶斯概率论: 最大似然函数、先验概率与后验概率
    - 模型选择: 模型复杂度控制、模型质量评价
    - 分类决策: 决策评价准则
- 要点
    - 概率论: 期望、方差、贝叶斯公式
    - 最优化: 参数估计

  - 机器学习: 模型和算法 [1] [2]
  - 模式识别: 分类决策 [3]

基本知识点

- 训练集 (training set) : 用来通过训练来调节模型的参数。
  - 输入变量 x 的 N 次观测组成，记作 $\mathbf{x} \equiv \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$
  - 目标变量 t 的 N 次观测组成，记作 $\mathbf{t} \equiv \{t_1, \cdots, t_N\}$
- 学习的结果: 表示为一个函数 $y(x)$，它以新的 x 为输入，产生的 y 为输出，结果与 t 的形式相同。
  - y 的具体形式 (参数) 是在训练 (training) 阶段被确定的，也被称为学习 (learning) 阶段。
  - 当训练阶段完成后，可以使用新的数据集去检验训练的结果，这种数据集称为测试集 (test set)。
  - 泛化 (generalization) : 正确分类与训练集不同的新样本的能力。
- 原始输入向量需要被预处理 (pre-processed)，变换到新的变量空间，也称为特征抽取 (feature extraction)，使问题变得更加容易解决。
- 有监督学习 (supervised learning)
  - 离散输出学习称为分类 (classification) 问题
  - 连续输出学习称为回归 (regression) 问题
- 无监督学习 (unsupervised learning)
  - 离散输出学习称为聚类 (clustering) 问题
  - 连续输出学习称为密度估计 (density estimation)
    * 高维空间投影到二维或者三维空间，为了数据可视化 (visualization) 或者降维
- 反馈学习 (强化学习) (reinforcement learning) : 本书不关注

1.1. 例子: 多项式曲线拟合

理论基础

- 概率论提供了数学框架，用来描述不确定性

[1] 李航著。统计学习方法。清华大学出版社。2012.
[2] 周志华著。机器学习清华大学出版社。2018.
[3] Duda R O,Peter E Hart,etc. 李宏东等译。模式分类。机械工业出版社。2003.

- 决策论提供了合适的标准，用来进行最优的预测。

前提条件

- 训练集: 输入数据: 由 x 的 N 次观察组成 x $\equiv (x_1, \cdots, x_N)^T$
- 训练集: 目标数据: 由 t 的 N 次观察组成 t $\equiv (t_1, \cdots, t_N)^T$

多项式函数是线性模型，应用于线性回归 (Ch 03) 和线性分类 (Ch 04)

$$y(x, \mathrm{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

最小化误差函数 (error function) 可以调整多项式函数的参数

- 平方误差函数 (square error function) : 最常用

$$E(\mathrm{w}) = \frac{1}{2} \sum_{n=1}^{N} [y(x_n, \mathrm{w}) - t_n]^2$$

- 根均方 (root-mean-square, RMS) 误差函数: 更方便

$$E_{RMS} = \sqrt{2E(\mathrm{w}^*)/N}$$

多项式的阶数 M 的选择，属于模型对比 (model comparison) 问题或者模型选择 (model selection) 问题。

拟合问题: 模型容量与实际问题不匹配

- 欠拟合 (Under-fitting) : 模型过于简单，模型容量低，不能充分描述问题
- 过拟合 (Over-fitting) : 模型过于复杂，模型容量高，可能描述数据噪声

正则化 (regularization) : 解决过拟合问题，即给误差函数增加惩罚项

- 正则项的 λ 系数控制过拟合的影响
- 统计学: 叫做收缩 (shrinkage) 方法
- 二次正则项: 称为岭回归 (ridge regression)
- 神经网络: 称为权值衰减 (weight decay)

确定模型容量: 验证集 (validation set)，也被称为拿出集 (hold-out set)，缺点是不能充分利用数据

数据集规模: 训练数据的数量应该是模型可调节参数的数量的 $5 \sim 10$ 倍。

最大似然 (maximum likelihood, ML)

- 最小二乘法是最大似然法的特例
- 过拟合问题是 ML 的一种通用属性
- 使用 Bayesian 方法解决过拟合问题，等价于正则化

1.2. 概率论

(建议跟着公式和例子推导)

理解离散随机变量与连续随机变量之间的关系

- 离散随机变量
  - 联合概率: X 取值 x_i，Y 取值 y_j，的联合概率是 $p(X=x_i,Y=y_j)=\frac{n_{ij}}{N}$
  - 边缘概率: X 取值 x_i(与 Y 取值无关) 的边缘概率是 $p(X=x_i)=\frac{c_i}{N}$
  - 加和规则推导: $p(X=x_i)=\sum_{j=1}^{L}p(X=x_i,Y=y_j)$
  - 条件概率: 在 X 取值 x_i 中 Y 取值 y_j 的条件概率是 $p(Y=y_j|X=x_i)=\frac{n_{ij}}{c_i}$
  - 乘积规则推导: $p(X=x_i,Y=y_j)=\frac{n_{ij}}{N}=\frac{n_{ij}}{c_i}\cdot\frac{c_i}{N}=p(Y=y_j|X=x_i)p(X=x_i)$
  - 贝叶斯定理: $p(Y|X)=\frac{p(X|Y)p(Y)}{p(X)}=\frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$
  - X 和 Y 相互独立: $p(X,Y)=p(X)p(Y)$

概率论的两种基本规则 (后面会有大量的应用，需要充分理解才能正确的使用)

- 加和规则 (sum rule) : $p(X)=\sum_Y p(X,Y)$
- 乘积规则 (product rule) : $p(X,Y)=p(Y|X)p(X)$

1.2.1. 概率密度函数

离散随机变量 概率质量函数 (probability mass function) : $p(X=x_i)=\sum_j n_{ij}/N$

条件概率 (conditional probability) : $p(Y=y_j|X=x_i)=n_{ij}/\sum_j n_{ij}$

- 利用乘积规则得到联合概率 (conditional probability) : $p(X=x_i,Y=y_j)=p(Y=y_j|X=x_i)p(X=x_i)$

联合概率 (joint probability) : $p(X=x_i,Y=y_j)=n_{ij}/N$

- 利用加和规则得到边缘概率 (marginal probability) : $p(X=x_i)=\sum_{j=1}^{L}p(X=x_i,Y=y_j)$

连续随机变量 概率密度函数 (probability density function) : $p(x \in (a,b)) = \int_a^b p(x)dx$

累积分布函数 (cumulative distribution function) : $p(z) = \int_{-\infty}^{z} p(x)dx$

1.2.2. 期望与协方差

期望 (expectation) : 函数的平均值

- 离散随机变量的期望: $\mathbb{E}[f] = \sum_x p(x)f(x)$
- 连续随机变量的期望: $\mathbb{E}[f] = \int p(x)f(x)dx$
- 有限数量的数据集，数据集中的点满足某个概率分布函数或者概率密度函数，那么 **期望**可以用 **求和**的方式来估计: $\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$
- 多变量函数的期望，使用下标 $\mathbb{E}_x[f(x,y)]$ 表明关于 x 的分布的平均，是 y 的一个函数。
  - $\mathbb{E}_x[f(x,y)] = \sum_x p(x)f(x,y)$
- 条件分布的条件期望 (conditional expectation) : $\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x)$

方差 (variance) : 度量了函数 $f(x)$ 在均值 $\mathbb{E}[f(x)]$ 附近的变化性

- $\text{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$
- $\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$

协方差 (covariance) : 度量两个随机变量之间的关系

- 两个值随机变量: $\text{cov}[x,y] = \mathbb{E}_{x,y}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])] = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]$
  - 如果两个随机变量相互独立，则它们的协方差为 0
- 两个向量随机变量: $\text{cov}[\text{x},\text{y}] = \mathbb{E}_{\text{x},\text{y}}[(\text{x} - \mathbb{E}[\text{x}])(\text{y} - \mathbb{E}[\text{y}])] = \mathbb{E}_{\text{x},\text{y}}[\text{xy}^T] - \mathbb{E}[\text{x}]\mathbb{E}[\text{y}^T]$
- 一个向量随机变量: $\text{cov}[x] \equiv \text{cov}[x,x]$

1.2.3. 经典概率论与贝叶斯概率论

经典概率论

- 概率: 随机重复事件发生的频率。
- 似然函数
  - 参数 w 是固定的值，可以通过某种形式的 “估计” 来确定
    - 估计的误差通过考察可能的数据集 $\mathcal{D}$ 获得

        * 确定误差的方法: Bootstrap[4]。通过从原始数据集中随机抽取来创建多个数据集，再对多个数据集的估计值求得方差确定参数误差。
    - 似然函数的负对数被叫做误差函数（error function），最大化似然函数等价于最小化误差函数
    - 广泛使用最大似然估计（maximum likelihood estimator, MLE）[5]（P 68），
    - 无信息先验概率就是最大似然估计

Bayesian 概率

- 概率: 定量描述不确定性的工具。
- Bayesian 定理: $p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$
    - 先验概率: $p(w)$ 表达参数 w 的假设
        * 先验概率可以方便地将先验知识包含在模型中
    - 似然函数: 使用条件概率 $p(\mathcal{D}|w)$ 表达观测数据的效果
        * 由观测数据集 D 来估计，可以被看成参数向量 w 的函数。
            · 参数的不确定性通过概率分布来表达。
        * 不是 w 的概率分布，因为它关于 w 的积分并不（一定）等于 1。因此它只是 w 的似然函数，不是概率。
    - 后验概率: $p(w|\mathcal{D})$ 表达观测到 $\mathcal{D}$ 之后估计参数 w 的不确定性。
        * $p(w|mathcalD) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})}$
    - 后验概率 $\propto$ 似然函数 x 先验概率
    - 取样方法: MCMC（Ch 11）[6]
    - 判别方法: 变分贝叶斯（Variational Bayes）和期望传播（Expectation Propagation）

1.2.4. 高斯分布

（主要概率分布 Ch 02）

高斯分布（Gaussian distribution），也叫正态分布（normal distribution）

一元实值随机变量 x 的高斯分布

- 定义

$$\mathcal{N}(\mathrm{x}|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp[-\frac{1}{2\sigma^2}(\mathrm{x} - \mu)^2]$$

[4]Geof H. Givens, Jennifer A. Hoeting. 计算统计人民邮件出版社。2009.
[5]Duda R O,Peter E Hart,etc. 李宏东等译。模式分类。机械工业出版社。2003.
[6]Geof H. Givens, Jennifer A. Hoeting. 计算统计人民邮件出版社。2009.

- 控制参数
  - 均值 μ
  - 方差 σ^2，标准差 σ
  - 精度 (precision) $\beta = 1/\sigma^2$
- 性质
  - 归一化: $\int_{-\infty}^{+\infty} \mathcal{N}(x|\mu, \sigma^2)dx = 1$
  - 期望: $\mathbb{E}[x] = \int_{-\infty}^{+\infty} \mathcal{N}(x|\mu, \sigma^2)xdx = \mu$
  - 二阶矩: $\mathbb{E}[x^2] = \int_{-\infty}^{+\infty} \mathcal{N}(x|\mu, \sigma^2)x^2dx = \mu^2 + \sigma^2$
  - 方差: $\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$
  - 分布的最大值叫众数，与均值恰好相等。

D 维随机向量 x 的高斯分布

- 定义

$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\}$$

- 控制参数
  - 均值向量: $\mu \in \mathcal{R}^D$
  - 协方差矩阵: $\Sigma \in \mathcal{R}^{D \times D}$

一元随机实值变量的例子

- 观测数据集 $\mathbf{x} = (x_1, \cdots, x_N)^T$ 表示实值变量 x 的 N 次观测
  - 观测数据集独立地从高斯分布 $\mathcal{N}(\mu, \sigma^2)$ 中抽取出来
    * μ 和 σ^2 未知
    * 独立同分布 (independent and identically distributed, i.i.d.)：独立地从相同数据点集合中抽取的数据
- 观测数据集的概率

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$$

- 对数似然函数:

$$\ln p(\mathbb{x}|\mu,\sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

- 基于最大似然函数估计参数
  - 均值的最大似然解等于样本均值: $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$
  - 方差的最大似然解等于样本均值的样本方差: $\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$
  - 理论上需要同时关于 μ 和 σ^2 最大化函数，实际上 μ 的解与 σ^2 的无关，因此可以先求解 μ，再求解 σ^2
- 最大似然的偏移问题是多项式曲线拟合问题中遇到的过拟合问题的核心。
  - 最大似然估计的均值等于模型中输入的真实均值
    - (Eq 1.55) : $\mathbb{E}[\mu_{ML}] = \mu$
  - 最大似然估计的方差小于模型中输入的真实方差
    - (Eq 1.56) : $\mathbb{E}[\sigma^2_{ML}] = \frac{N-1}{N}\sigma^2$
  - 无偏的方差估计
    - (Eq 1.59) : $\tilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$
    - 在 $N\to\infty$ 时，最大似然解的偏移影响不大，方差的最大似然解与真实方差相等
- 最大似然估计 vs 最大后验估计
  - 最大似然估计: 在给定的 “数据集” 下最大化 “参数” 的概率
  - 最大后验估计: 在给定 “参数” 下最大化 “数据集” 的概率

1.2.5. 概率建模: 多项式曲线拟合

- 概率建模
  - 一对数据点 (x,y) 的模型: $p(t|x,\mathsf{w},\beta) = \mathcal{N}(t|y(x,\mathsf{w}),\beta^{-1})$
  - N 对数据点 (数据集) 的模型: $p(\mathsf{t}|\mathsf{x},\mathsf{w},\beta) = \prod_{n=1}^{N}\mathcal{N}(t_n|y(x_n,\mathsf{w}),\beta^{-1})$
  - 对数似然函数: $\ln p(\mathsf{t}|\mathsf{x},\mathsf{w},\beta) = -\frac{\beta}{2}\sum_{n=1}^{N}(y(x_n,\mathsf{w})-t_n)^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$
  - 模型参数
    - $\beta = 1/\sigma^2$
    - $y(x,\mathsf{w}) = \sum_{j=0}^{M} w_j x^j$

最大似然解 (最大化似然函数等价于最小化平方和误差函数)

- 模型参数: w_{ML}
- 精度参数: $\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}[y(x_n,\mathsf{w}_{ML})-t_n]^2$

- 预测分布 (Eq 1.64) : $p(t|x, \mathsf{w}_{ML}, \beta_{ML}) = \mathcal{N}(t|y(x, \mathsf{w}_{ML}, \beta_{ML}^{-1}))$

最大后验解 (最大化后验概率等价于最小化正则化的平方和误差函数)

- 先验概率: $p(\mathsf{w}|\alpha) = \mathcal{N}(\mathsf{w}|0, \alpha^{-1}\mathsf{I}) = (\frac{\alpha}{2\pi})^{(M+1)/2} \exp\left(-\frac{\alpha}{2}\mathsf{w}^T\mathsf{w}\right)$
- 后验概率: $p(\mathsf{w}|\mathsf{x}, \mathsf{t}, \alpha, \beta) \propto p(\mathsf{t}|\mathsf{x}, \mathsf{w}, \beta)p(\mathsf{w}|\alpha)$
- 等价公式: $\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathsf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathsf{w}^T\mathsf{w}$

1.2.6. 贝叶斯曲线拟合 (Sec 3.3.)

(依然用于理解贝叶斯观点在模型中的应用，无法理解可以暂时放弃)

使用最大后验估计，依然属于点估计，不属于贝叶斯的观点。在模式识别中，对所有的模型参数 w 进行积分才是贝叶斯方法的核心。

贝叶斯模型

- 前提条件
  - 输入数据 x
  - 目标数据 t
  - 新的测试点 x
  - 新的预测目标 t
- 预测概率 (Eq 1.69) : $p(t|x, \mathsf{x}, \mathsf{t}) = \int p(t|x, \mathsf{w}, \beta)p(\mathsf{w}|\mathsf{x}, \mathsf{t}, \alpha, \beta)d\mathsf{w} = \mathcal{N}(t|m(x), s^2(x))$
  - 模型参数
    * $m(x) = \beta\phi(x)^T\mathsf{S}\sum_{n=1}^{N}\phi(x_n)t_n$: 表示预测值 t 的不确定性，受目标变量上的噪声影响，在最大似然的预测分布 (1.64) 中通过 β_{ML}^{-1} 表达
    * $s^2(x) = \beta^{-1} + \phi(x)^T\mathsf{S}\phi(x)$: 对参数 w 的不确定性有影响
    * $\mathsf{S}^{-1} = \alpha\mathsf{I} + \beta\sum_{n=1}^{N}\phi(x_n)\phi(x_n)^T$
    * $\phi(x) \equiv \phi_i(x) = x^i, (i = 0, \cdots, M)$

1.3. 模型选择

- 模型复杂度的控制:
  - 多项式的阶数控制了模型的自由参数的个数，即控制了模型的复杂度
  - 正则化系数 λ 影响着模型的自由参数的取值，也控制了模型的复杂度
- 交叉验证 (cross validation)
  - 可以解决验证模型时面临的数据不足问题

  - 会增加训练成本
  - 留一法 (leave-one-out) 是特例

- 信息准则: 度量模型的质量，避免交叉验证的训练成本，使模型选择完全依赖于训练数据
  - 修正最大似然的偏差的方法: 增加惩罚项来补偿过于复杂的模型造成的过拟合
  - 赤池信息准则 (Akaike information criterion, AIC) : $\ln p(\mathcal{D}|\mathrm{w}_{ML}) - M$
  - 贝叶斯信息准则 (Bayesian information criterion, BIC) : (Sec 4.4.1.)

1.4. 维度灾难

在高维空间中，一个球体的大部分的体积在哪里？(可以推导公式理解)

- D 维空间的 $r = 1$ 的球体的体积: $V_D(r) = K_D r^D$
- $r = 1 - \epsilon$ 和 $r = 1$ 之间的球壳体积与 $r = 1$ 的球体的体积比:
  - $(V_D(1) - V_D(1 - \epsilon))/V_D(1) = 1 - (1 - \epsilon)^D$
  - D 的维数越大，体积比趋近于 1
  - 在高维空间中，一个球体的大部分体积都聚焦在表面附近的薄球壳上。

维度灾难 (curse of dimensionality) : 在高维空间中，大部分的数据都集中在薄球壳上导致数据无法有效区分

- 维度灾难的解决方案
  - 真实数据经常被限制在有着较低的有效维度的空间区域中；(通过特征选择降维)
  - 真实数据通常比较光滑 (至少局部上比较光滑) (通过局部流形降维)

1.5. 决策论 (分类问题、模式识别)

问题原型：输入向量 x 和目标向量 t

- 回归问题：目标向量是连续性变量组成
- 分类问题：目标向量是离散性变量组成，即类别标签

决策论: 保证在不确定性的情况下做出最优的决策。

- 联合概率分布 $p(\mathrm{x}, \mathrm{t})$ 总结了所有的不确定性
- 从训练数据集中确定输入向量 x 和目标向量 t 的联合分布 $p(\mathrm{x},\mathrm{t})$ 是推断 (inference) 问题
- 从测试数据集中确定输入向量 x 和目标分类 $\mathcal{C}_k$ 的条件概率 $p(\mathcal{C}_k|\mathrm{x})$ 是决策 (inference) 问题

1.5.1. 最小化错误分类率

基本概念

- 输入空间根据规则切分成不同的区域，这些区域被称为决策区域。
- 每个类别都有一个决策区域，区域 $\mathcal{R}_k$ 中的点都被分到对应类别 $\mathcal{C}_k$ 中
- 决策区域间的边界叫做决策边界 (decision boundary) 或者决策面 (decision surface)。

决策: $p(\mathcal{C}_k|\mathsf{x}) = \frac{p(\mathsf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathsf{x})}$

- $p(\mathcal{C}_k)$ 称为类 $\mathcal{C}_k$ 的先验概率
- $p(\mathcal{C}_k|\mathsf{x})$ 称为类 $\mathcal{C}_k$ 的后验概率
- (Fig 1.24) 最小化错误分类率，将 x 分配到具有最大后验概率的类别中

最小化错误分类率: 将 x 分配到后验概率 $p(\mathcal{C}_k|\mathsf{x})$ 最大的类别中，分类错误的概率最小

$$\begin{aligned} p(\text{mistake}) &= p(\mathsf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathsf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathsf{x}, \mathcal{C}_2)\mathsf{dx} + \int_{\mathcal{R}_2} p(\mathsf{x}, \mathcal{C}_1)\mathsf{dx} \end{aligned}$$

最大化正确分类率: 将 x 分配到联合概率 $p(\mathcal{C}_k, \mathsf{x})$ 最大的类别中，分类正确的概率最大

$$p(\text{correct}) = \sum_{k=1}^{K} p(\mathsf{x} \in \mathcal{R}_k, \mathcal{C}_k) = \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathsf{x}, \mathcal{C}_k)\mathsf{dx}$$

1.5.2. 最小化期望损失 (加先验)

损失函数 (loss function) 也称为代价函数 (cost function)，是对所有可能的决策产生的损失的一种整体度量，目标是最小化整体的损失。

效用函数 (utility function)，目标是最大化整体的效用，如果效用函数等于损失函数的相反数，则两者等价。

最小化损失的期望: $\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathsf{x}, \mathcal{C}_k) dx$

- L_{kj} 表示损失矩阵的第 k, j 个元素，即把类别 $\mathcal{C}_k$ 的数据 x 错分为 C_j 的损失

最小化损失的期望的决策规则: 对于每个新的 x，分到的类别 C_j 保证 $\sum_k L_{kj} p(\mathcal{C}_k|\mathsf{x})$ 最小化

1.5.3. 拒绝选项

拒绝选项 (reject option) : 为了避免做出错误的决策，系统拒绝从做出类别选择，从而使模型的分类错误率降低。

1.5.4. 推断与决策

分类问题的两个阶段

- 推断 (inference) : 使用训练数据学习后验概率 $p(\mathcal{C}_k|x)$ 的模型
- 决策 (decision) : 使用后验概率模型进行最优的分类

解决分类问题的三种方法:

- 生成式模型 (generative model) : 显式或者隐式地对输入和输出进行建模的方法
  - 推断阶段:
    * 首先基于每个类别 $\mathcal{C}_k$ 推断类条件密度 $p(|\mathcal{C}_k)$
    * 再推断类别的先验概率密度 $p(\mathcal{C}_k)$
    * 再基于贝叶斯定理 $p(\mathcal{C}_k|\mathrm{x}) = \frac{p(\mathrm{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathrm{x})} = \frac{p(\mathrm{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_k p(\mathrm{x}||\mathcal{C}_k)p(\mathcal{C}_k)}$ 学习类别的后验概率密度
  - 决策阶段: 使用决策论来确定每个新的输入 x 的类别。
  - 优点：能够计算数据的边缘概率密度 $p(\mathrm{x})$，方便计算具有低概率的新数据点，即离群点检测。
  - 缺点：需要计算的工作量最大
- 判别式模型 (Discriminative Models) : 直接对后验概率密度建模的方法
  - 推断阶段: 首先推断类别的后验概率密度 $p(\mathcal{C}_k|\mathrm{x})$
  - 决策阶段: 使用决策论来确定每个新的输入 x 的类别。
- 同时解决两个阶段的问题，即把输入直接映射为决策的函数称为判别函数 (discriminant function)

使用后验概率进行决策的理由

- 最小化风险: 修改最小风险准则就可以适应损失矩阵中元素改变
- 拒绝选项:
  - 确定最小化误分类率的拒绝标准
  - 确定最小化期望损失的拒绝标准
- 补偿类别先验概率: 解决不平衡的数据集存在的问题

- 组合模型 (Ch 14) : 通过特征的独立性假设，将大问题分解成小问题，例如: 朴素贝叶斯模型
  - $p(\mathsf{x}_I, \mathsf{x}_B|\mathcal{C}_k) = p(\mathsf{x}_I|\mathcal{C}_k)p(\mathsf{x}_B|\mathcal{C}_k)$
  - $p(\mathcal{C}_k|\mathsf{x}_I, \mathsf{x}_B) \propto p(\mathsf{x}_I, \mathsf{x}_B|\mathcal{C}_k)p(\mathcal{C}_k) \propto p(\mathsf{x}_I|\mathcal{C}_k)p(\mathsf{x}_B|\mathcal{C}_k)p(\mathcal{C}_k) \propto \frac{p(\mathcal{C}_k|\mathsf{x}_I)p(\mathcal{C}_k|\mathsf{x}_B)}{p(\mathcal{C}_k)}$

1.5.5. 回归问题的损失函数

回归问题的决策阶段：对于每个输入向量 x，输出目标变量 t 的特定估计 $y(\mathsf{x})$，造成损失 $L(t, y(\mathsf{x}))$，则整个问题的平均损失，即损失函数的期望: $\mathbb{E}[L] = \int\int L(t, y(\mathsf{x}))p(\mathsf{x}, t)\mathsf{d}\mathsf{x}\,\mathsf{d}t$

- 变分法最小化 $\mathbb{E}[L]$: $\frac{\delta\mathbb{E}[L]}{\delta y(\mathsf{x})} = 2\int[y(\mathsf{x}) - t]p(\mathsf{x}, t)dt = 0$
- 求解 $y(\mathsf{x})$ 的最优解: $y(\mathsf{x}) = \frac{\int tp(\mathsf{x},t)\mathsf{d}t}{p(\mathsf{x})} = \int tp(t|\mathsf{x})\mathsf{d}t = \mathbb{E}_t[t|\mathsf{x}]$

损失函数的具体选择

- 平方损失: $L(t, y(\mathsf{x})) = \{y(\mathsf{x}) - t\}^2$
  - $\mathbb{E}[L] = \int\int\{y(\mathsf{x}) - t\}^2p(\mathsf{x}, t)\mathsf{d}\mathsf{x}\,\mathsf{d}t$
  - $\mathbb{E}[L] = \int\int\{y(\mathsf{x}) - \mathbb{E}_t[t|\mathsf{x}] + \mathbb{E}_t[t|\mathsf{x}] - t\}^2p(\mathsf{x}, t)\mathsf{d}\mathsf{x}\,\mathsf{d}t$
  - $\mathbb{E}[L] = \int\int[\{y(\mathsf{x}) - \mathbb{E}_t[t|\mathsf{x}]\}^2 + 2\{y(\mathsf{x}) - \mathbb{E}_t[t|\mathsf{x}]\}\{\mathbb{E}_t[t|\mathsf{x}] - t\} + \{\mathbb{E}_t[t|\mathsf{x}] - t\}^2]p(\mathsf{x}, t)\mathsf{d}\mathsf{x}\,\mathsf{d}t$
  - $\mathbb{E}[L] = \int\{y(\mathsf{x}) - \mathbb{E}_t[t|\mathsf{x}]\}^2p(\mathsf{x})\mathsf{d}\mathsf{x} + \int \mathsf{var}[t|\mathsf{x}]p(\mathsf{x})\mathsf{d}\mathsf{x}$
  - 当 $y(\mathsf{x}) = \mathbb{E}_t[t|\mathsf{x}]$ 时第一项取得最小值，即 $\mathbb{E}[L]$ 最小化。
    - 说明最小平方预测由条件均值给出。
  - 第二项是 t 的分布的方差在 x 上进行了平均，表示目标数据内在的变化性，即噪声。
    - 与 $y(\mathsf{x})$ 无关，表示损失函数的不可减小的最小值。
- 闵可夫斯基 (Minkowski) 损失函数 (平方损失函数的一种推广) $L_q(t, y(\mathsf{x})) = |y(\mathsf{x}) - t|^q$
  - $\mathbb{E}[L_q] = \int\int|y(\mathsf{x}) - t|^qp(\mathsf{x}, t)\mathsf{d}\mathsf{x}\,\mathsf{d}t$

解决回归问题的三种思路 (建议从机器学习参考书中熟悉这三种方法)

- 函数模型: 直接从训练数据中寻找一个回归函数
- 概率建模: 最大似然: 先解决条件概率密度的推断问题，再求条件均值；
- 概率建模: 最大后验: 先求联合概率密度，再求条件概率密度，最后求条件均值；

1.6. 信息论

(推导过程值得理解，没有深入讨论的部分可以跳过)

离散随机变量 x

- 信息量: 学习 x 的值的时候的「惊讶」程度，
  - $h(x) = -\log_2 p(x)$
  - 负号确保信息一定是正数或者为零
  - 对数确保低概率事件 x 对应高的信息量
    - 使用 2 作为对数的底，$h(x)$ 的单位是比特 (bit, binary digit)
    - 使用 e 作为对数的底，$h(x)$ 的单位是奈特 (nat, Napierian digit)
- 平均信息量: 发送者想传输一个随机变量的值给接收者，在这个传输过程中，他们传输的平均信息量就是随机变量 x 的 **熵**。
  - $H[x] = -\sum_x p(x)\log_2 p(x)$
  - 无噪声编码定理 (noiseless coding theorem) 表明：**熵**是传输一个随机变量状态值所需要的比特位的下界。
  - 在统计力学中，熵用来描述无序程度的度量。

离散随机变量的熵 (entropy)

- 熵：$H[p(x)] = -\sum_i p(x_i)\ln p(x_i)$
- 离散随机变量服从均匀分布时，其熵值最大

连续随机变量的熵

- H_Δ 的定义
  - $\sum_i p(x_i)\Delta = 1$
  - 因为 $\Delta \to 0$ 时，熵的连续形式与离散形式的差 $\ln\Delta \to \infty$，即具体化一个连续变量需要大量的比特位，因此省略 $-\ln\Delta$ 得连续随机变量的熵 $\lim_{\Delta\to 0}\{-\sum_i p(x_i)\Delta\ln p(x_i)\} = -\int p(\mathrm{x})\ln p(\mathrm{x})\mathrm{dx}$

$$
\begin{aligned}
H_\Delta &= -\sum_i p(x_i)\Delta\ln(p(x_i)\Delta) \\
&= -\sum_i p(x_i)\Delta\ln p(x_i) - \sum_i p(x_i)\Delta\ln\Delta \\
&= -\sum_i p(x_i)\Delta\ln p(x_i) - \ln\Delta
\end{aligned}
$$

- 微分熵：$H[\mathrm{x}] = -\int p(\mathrm{x})\ln p(\mathrm{x})\mathrm{dx}$

  - 微分熵可以为负

- 最大化微分熵需要遵循的三个限制
  - $\int_{-\infty}^{+\infty} p(x)\mathrm{d}x = 1$
  - $\int_{-\infty}^{+\infty} xp(x)\mathrm{d}x = \mu$
  - $\int_{-\infty}^{+\infty} (x-\mu)^2 p(x)\mathrm{d}x = \sigma^2$
- 使用拉格朗日乘数法求解带有限制条件的最大化问题
  - 使用变分法求解这个函数，令其导数等于零，得 $p(x) = \exp\{-1 + \lambda_1 + \lambda_2 x + \lambda_3(x-\mu)^2\}$
  - 将结果带入限制方程，得 $p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$

$$-\int_{-\infty}^{+\infty} p(x)\ln p(x)\mathrm{d}x + \lambda_1\left(\int_{-\infty}^{+\infty} p(x)\mathrm{d}x - 1\right) + \lambda_2\left(\int_{-\infty}^{+\infty} xp(x)\mathrm{d}x - \mu\right) + \lambda_3\left(\int_{-\infty}^{+\infty} (x-\mu)^2 p(x)\mathrm{d}x - \sigma^2\right)$$

- 连续随机变量服从高斯分布时，其熵值最大。
  - 高斯分布的微分熵 $H[x] = \frac{1}{2}\{1 + \ln(2\pi\sigma^2)\}$
  - 熵随着分布宽度（σ^2）的增加而增加
  - 当 $\sigma^2 < \frac{1}{2\pi e}$ 时，$H[x] < 0$
- 在给定 x 的情况下，y 的条件熵
  - $H[\mathrm{y}|\mathrm{x}] = -\int\int p(\mathrm{y},\mathrm{x}) \ln p(\mathrm{y}|\mathrm{x})\ \mathrm{dy}\ \mathrm{dx}$

联合熵

- $H[\mathrm{x},\mathrm{y}] = H[\mathrm{y}|\mathrm{x}] + H[\mathrm{x}]$
- $H[\mathrm{x},\mathrm{y}]$ 是 $p(\mathrm{x},\mathrm{y})$ 的微分熵
- $H[\mathrm{y}|\mathrm{x}]$ 是 $p(\mathrm{y}|\mathrm{x})$ 的微分熵
- $H[\mathrm{x}]$ 是 $p(\mathrm{x})$ 的微分熵

1.6.1. 相对熵和互信息

凸函数: 每条弦都位于函数的弧或者弧的上面。$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda)f(b)$

- 凸函数的二阶导数处处为正。

- 严格凸函数 (strictly convex function) : 等号只在 $\lambda = 0$ 和 $\lambda = 1$ 处取得。

凹函数: 如果 $f(x)$ 是凸函数，则 $-f(x)$ 是凹函数

- 严格凹函数 (strictly concave function)

Jensen 不等式

- $f(\sum_{i=1}^{M} \lambda_i x_i \leq \sum_{i=1}^{M} \lambda_i f(x_i)$
  - 将 λ_i 看作取值为 $\{x_i\}$ 的离散变量 x 的概率分布，则公式为 $f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$
- $f(\int \mathrm{x} p(\mathrm{x})\,\mathrm{dx}) \leq \int f(\mathrm{x}) p(\mathrm{x})\mathrm{dx}$

KL 散度: 是两个分布 $p(\mathrm{x})$ 和 $q(\mathrm{x})$ 之间不相似程度的度量。

- $-\ln x$ 是严格凸函数
- 归一化条件：$\int q(\mathrm{x})\mathrm{dx} = 1$

$$\mathrm{KL}(p||q) = -\int p(\mathrm{x}) \ln q(\mathrm{x})\,\mathrm{dx} - (-\int p(\mathrm{x}) \ln p(\mathrm{x})\,\mathrm{dx})$$

$$\mathrm{KL}(p||q) = -\int p(\mathrm{x}) \ln(\frac{q(\mathrm{x})}{p(\mathrm{x})})\mathrm{dx} \geq -\ln \int q(\mathrm{x})\,\mathrm{dx} = 0$$

Laplace 近似: 是用方便计算的分布去逼近不方便计算的分布，解决 KL 散度不易计算的问题 (Sec 4.4)

- 假设数据通过未知分布 $p(\mathrm{x})$ 生成，为了对 $p(\mathrm{x})$ 建模可以使用 $q(\mathrm{x}|\theta)$ 来近似，$q(\mathrm{x}|\theta)$ 可以是一个已知分布 (例如：θ 是控制多元高斯分布的参数)，通过最小化 $p(\mathrm{x})$ 和 $q(\mathrm{x}|\theta)$ 之间关于 θ 的 KL 散度，就可以得到 $p(\mathrm{x})$ 的近似分布。
  - $\mathrm{KL}(p||q) \simeq \frac{1}{N} \sum_{n=1}^{N} [-\ln q(\mathrm{x}_n|\theta) + \ln p(\mathrm{x}_n)]$
    - $-\ln q(\mathrm{x}_n|\theta)$ 是使用训练集估计的分布 $q(\mathrm{x}|\theta)$ 下的 θ 的负对数似然函数
    - $\ln p(\mathrm{x}_n)$ 与 θ 无关
    - 最小化 KL 散度等价于最大化似然函数
- 互信息 (mutual information) : 表示一个新的观测 y 造成的 x 的不确定性的减小 (反之亦然)
  - 两个随机变量 $p(\mathrm{x})$ 和 $p(\mathrm{y})$ 组成的数据集由 $p(\mathrm{x},\mathrm{y})$ 给出
    - 如果两个变量相互独立，则联合分布 $p(\mathrm{x},\mathrm{y}) = p(\mathrm{x})p(\mathrm{y})$，互信息 $I[\mathrm{x},\mathrm{y}] = 0$
    - 如果两个变量没有相互独立，可以通过 KL 散度判断它们之间的「独立」程度，也称为随机变量 $p(\mathrm{x})$ 和 $p(\mathrm{y})$ 之间的互信息 $I[\mathrm{x},\mathrm{y}] > 0$

$$
\begin{aligned}
I[\mathrm{x},\mathrm{y}] &\equiv \mathrm{KL}(p(\mathrm{x},\mathrm{y})||p(\mathrm{x})p(\mathrm{y})) \\
&= -\int\int p(\mathrm{x},\mathrm{y}) \ln(\frac{p(\mathrm{x})p(\mathrm{y})}{p(\mathrm{x},\mathrm{y})}) \, \mathrm{dx} \, \mathrm{dy} \\
I[\mathrm{x},\mathrm{y}] &= H[\mathrm{x}] - H[\mathrm{x}|\mathrm{y}] \\
&= H[\mathrm{y}] - H[\mathrm{y}|\mathrm{x}]
\end{aligned}
$$

相对熵 (relative entropy) 或者 KL 散度 (Kullback-Leibler divergence) (推导过程建议理解)

小结

- 本章对于后面需要的重要概念都进行了说明和推导，方便深入理解后面提及的各种算法。
- 如果看完本章后感觉充斥着许多新的概念，那么建议先放下，找本更加基础的模式识别与机器学习的书，例如: 李航的《统计学习方法》和周志华的《机器学习》等。

Ch 02. 概率分布

提纲

重点

- 密度估计
  - 充分统计量
- 高斯分布 (建议充分熟悉)

难点

- 贝叶斯估计
- 多元高斯分布
- 指数族分布
- 共轭分布
  - 共轭先验分布
  - 超参数

学习要点

密度估计（density estimation）：在给定有限次观测 $x_1, \cdots, x_N$ 的前提下，对随机变量 x 的概率分布 $p(x)$ 建模。

- 参数密度估计: 对控制概率分布的参数进行估计。
  - 最大似然估计: 最优化似然函数在确定参数的具体值。
  - 顺序估计: 利用充分统计量，在线进行密度估计。
    * 充分统计量（sufficient statistic）：最大似然估计的解只通过一个统计量就可以满足对数据的依赖，这个统计量就是充分统计量。
  - 贝叶斯估计: 引入参数的先验分布，再来计算对应后验概率分布。
- 非参数（nonparametric）密度估计: 分布的形式依赖于数据集的规模。
  - 直方图: 最基本的估计方法，但是在高维度问题中无法应用。
  - 核密度估计: 固定区域 V 的大小，统计区域内的数据点个数 K，利用平滑的核函数，从而得到光滑的概率分布模型。
  - K 近邻估计: 固定区域内的数据点个数 K，计算区域 V 的大小。不是真实的概率密度模型，因为它在整个空间的积分是发散的。
    * $K = 1$ 时就是最近邻规则。
- 对比
  - 参数密度估计需要假设准备估计的概率分布是什么，如果估计错误就没办法得到正确的结果；
  - 非参数估计则不需要进行这种假设。
  - 详情参考（[7] Ch 03, [8] Ch 04）

学习方式

- 离线学习，也叫批量学习。
  - 所有数据一次性采集完成后，对所有数据进行学习。
- 在线学习，也叫顺序学习。
  - 数据无法一次性采集得到，需要随着时间片按顺序得到，学习的结果也需要随着时间发生改变。

参数分布（parametric distribution）：少量可调节的参数控制了整个概率分布。

[7] Andrew R. Webb. 统计模式识别。电子工业出版社。2004.
[8] Duda R O,Peter E Hart,etc. 李宏东等译。模式分类。机械工业出版社。2003.

- 先验分布（prior）：设定的参数的先验分布，参数可以看成先验分布中假想观测的有效观测数。
  - 共轭先验（conjugate prior）：使得后验分布的函数形式与先验概率相同，从而使贝叶斯分析得到简化。
    * Gamma 函数: （P 48, Ex 1.17）
    * 超参数: 控制着参数的概率分布。
  - 无信息先验（non-informative prior）：对先验分布几乎无知，需要寻找一个先验分布，能够对后验分布产生尽可能少的影响。
- 后验分布（posterior）：加入先验分布信息的概率分布，用于贝叶斯估计需要。

指数族分布（exponential family）：具有指定的指数形式的概率分布的集合。

共轭分布: 使得后验分布的函数形式与先验概率相同，从而使贝叶斯分析得到简化。

- 二项分布与 Beta 分布共轭
- 多项式分布与 Dirichlet 分布共轭
- 高斯分布与高斯分布共轭
  - 条件高斯: 如果两组变量是联合高斯分布，那么以一组变量为条件，另一组变量同样也是高斯分布。
  - 边缘高斯: 如果两组变量是联合高斯分布，那么任何一个变量的边缘分布也是高斯分布。
- 混合分布（mixture distribution）模型: 将多个基本概率分布进行线性组合形成的分布。
  - 混合高斯（mixture of Gaussian）：将多个高斯分布进行线性组合形成的分布。每一个高斯概率密度称为混合分布中的一个成分。
  - 学生 t 分布: 表现为无限多个同均值不同精度的高斯分布的叠加形成的无限高斯混合模型。比普通高斯分布具有更好的「鲁棒性」。
- 周期概率分布: 用于描述具有周期性质的随机变量。
  - Von Mises 分布: 也称为环形正态分布（circular normal）。是高斯分布对于周期变量的推广。

2.1. 二元变量: 离散分布

二元变量: 用于描述只能取两种可能值中的某一种这样的量。

Bernoulli 分布:

- Bern$(x|\mu) = \mu^x(1-\mu)^{1-x}$

  - 均值: $\mathbb{E}[x] \equiv \sum_{x=0}^{1} x\text{Bern}(x|\mu) = \mu$
  - 方差: $\text{var}[x] \equiv \sum_{x=0}^{1}(x - \mathbb{E}[x])^2\text{Bern}(x|\mu) = \mu(1-\mu)$
- 假设数据集 $\mathcal{D} = \{x_1, \cdots, x_N\}$ 中的观测都是独立地从 $p(x|\mu) = \text{Bern}(x|\mu)$ 中抽取的
  - 似然函数: $p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$
  - 对数似然: $\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N}[x_n \ln \mu + (1-x_n)\ln(1-\mu)]$
    * $\sum_n x_n$ 是 Bernoulli 分布的充分统计量 (sufficient statistic)，因为对数似然函数只通过 $\sum_n x_n$ 对数据产生依赖
  - 样本均值: $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{m}{N}$，$m$ 为 $x=1$ 在数据集里面的数量

二项分布 (binomial distribution) : 给定数据集规模 N，在数据集里面 $x=1$ 的观测数量为 m 的概率分布为二项分布

- $\text{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}, \binom{N}{m} \equiv \frac{N!}{(N-m)!m!}$
  - 均值 (Eq 2.11) : $\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\text{Bin}(m|N,\mu) = N\mu$
  - 方差 (Eq 2.12) : $\text{var}[m] \equiv \sum_{m=0}^{N}(m - \mathbb{E}[m])^2\text{Bin}(m|N,\mu) = N\mu(1-\mu)$

2.2.1. Beta 分布: Bin 分布的共轭分布

Beta 分布:

- $\text{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}$
- Gamma 函数 (Ex 1.17) (Eq 1.141)
  - $\Gamma(x) \equiv \int_0^\infty u^{x-1}e^{-u}du$
  - 用于保证 Beta 分布的归一化性质 $\int_0^1 \text{Beta}(\mu|a,b)\text{d}\mu = 1$
- 均值: $\mathbb{E}[\mu] = \frac{a}{a+b}$
- 方差: $\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$
- 超参数 (hyperparameter) : 用于控制参数的参数
  - 例如: 参数 a, b 用于控制参数 μ 的概率分布。

共轭性: 后验概率 ∝ 先验概率 × 似然函数，有着与先验概率分布相同的函数形式

- Beta 分布是二项分布的共轭分布。用于引入 μ 的先验概率分布 $p(\mu)$。
- $p(\mu|m,l,a,b) \propto \mu^{m+a-1}(1-\mu)^{l+b-1} = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}$
  - m : 代表硬币「正面朝上」的样本数量

  - $l = N - m$: 代表硬币「反面朝上」的样本数量
  - $\frac{\Gamma(N+a+b)}{\Gamma(m+a)\Gamma(n+b)}$: 归一化系数，满足后验分布归一化的需要
  - 从先验概率到后验概率，a 的值增加了原始 m 的值，b 的值增加了原始 l 的值，因此先验概率就是利用硬币曾经的数据为后验概率提供信息，详情参考 (Fig 2.2)

贝叶斯观点: 学习过程中的顺序方法与先验和似然函数的选择无关，只取决于数据独立同分布的假设

- 如果数据集有限，后验均值总是位于先验均值和最大似然估计之间
- 如果数据集无限大，先验概率对结果的影响几乎为零，贝叶斯估计和最大似然估计的结果将趋于一致

贝叶斯推断问题推导 - 前提条件 : 基于观测数据集 $\mathcal{D}$，使用联合概率分布 $p(\theta, \mathcal{D})$ 描述的参数 θ 的贝叶斯推断问题 - $\mathbb{E}_\theta[\theta] = \mathbb{E}_\mathcal{D}[\mathcal{E}_\theta[\theta|\mathcal{D}]]$ - $\mathbb{E}_\theta[\theta] \equiv \int p(\theta)\theta \mathrm{d}\theta$ - $\mathbb{E}_\mathcal{D}[\mathcal{E}_\theta[\theta|\mathcal{D}]] \equiv \int\{\int \theta p(\theta|\mathcal{D})\mathrm{d}\theta\}p(\mathcal{D})\mathrm{d}\mathcal{D}$ - θ 的后验均值在产生数据集的整个分布上求平均等于 θ 的先验均值 - $\mathrm{var}_\theta[\theta] = \mathbb{E}_\mathcal{D}[\mathrm{var}_\theta[\theta|\mathcal{D}]] + \mathrm{var}_\mathcal{D}[\mathbb{E}_\theta[\theta|\mathcal{D}]]$ - θ 的先验方差: $\mathrm{var}_\theta[\theta]$ - θ 的平均后验方差: $\mathbb{E}_\mathcal{D}[\mathrm{var}_\theta[\theta|\mathcal{D}]]$ - θ 的后验均值方差: $\mathrm{var}_\mathcal{D}[\mathbb{E}_\theta[\theta|\mathcal{D}]]$

2.2. 多项式变量: 离散分布

多项式变量: 用于描述只能取 K 种可能值中某一种的量。

「1-of-K」表示法: 也称为「One-Hot 编码」。

- 变量被表示成一个 K 维向量 $\mathrm{x} = (0, 0, 1, 0, 0, 0)^T$，向量中的一个元素 x_k 等于 1，剩余的元素等于 0。
  - 向量满足 $\sum_{k=1}^K x_k = 1$

x 的分布: $p(\mathrm{x}|\mu) = \prod_{k=1}^K \mu_k^{x_k}$ 是 Bernoulli 分布对于多个输出的推广。

- 向量 $\mu = (\mu_1, \cdots, \mu_K)^T$
  - 参数 μ_k 表示 $x_k = 1$ 的概率
  - $\sum_\mathrm{x} p(\mathrm{x}|\mu) = \sum_{k=1}^K \mu_k = 1, \mu_k \geq 0$
- 期望: $\mathbb{E}[\mathrm{x}|\mu] = \sum_\mathrm{x} p(\mathrm{x}|\mu)\mathrm{x} = (\mu_1, \cdots, \mu_K)^T = \mu$

有 N 个独立观测值的数据集 $\mathcal{D} = \{\mathrm{x}_1, \cdots, \mathrm{x}_N\}$

- 似然函数: $p(\mathcal{D}|\mu) = \prod_{n=1}^N \prod_{k=1}^K \mu_k^{x_{nk}} = \prod_{k=1}^K \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^K \mu_k^{m_k}$
  - $m_k = \sum_n x_{nk}$: 表示观测到 $x_k = 1$ 的次数，是这个分布的充分统计量 (Sufficient Statistics)。

- 最大似然解
  - 拉格朗日乘数法: $\sum_{k=1}^{K} m_k \ln \mu_k + \lambda(\sum_{k=1}^{K} \mu_k - 1)$
  - 关于 μ_k 的导数等于零，得: $\mu_{k_{ML}} = \frac{m_k}{N}$

多项式分布 (Multinomial Distribution) : 给定数据集规模 N，在数据集里面 $x_k = 1$ 的观测数量为 m_k 的概率分布为

- $\text{Multi}(m_1, m_2, \cdots, m_K|\mu, N) = \binom{N}{m_1 m_2 \cdots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$
  - 归一化系数: $\binom{N}{m_1 m_2 \cdots m_K} = \frac{N!}{m_1! m_2! \cdots m_K!}$
    - m_k 满足的限制: $\sum_{k=1}^{K} m_k = N$
  - 多项式分布 vs 二项分布

二项分布 (binomial distribution) : 给定数据集规模 N，在数据集里面 $x = 1$ 的观测数量为 m 的概率分布为二项分布

- $\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}, \binom{N}{m} \equiv \frac{N!}{(N-m)!m!}$
  - 均值: $\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\text{Bin}(m|N, \mu) = N\mu$
  - 方差: $\text{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \text{Bin}(m|N, \mu) = N\mu(1-\mu)$

2.2.1. Dirichlet 分布: 多项式分布的共轭分布

Dirichlet 分布

- 共轭先验: $p(\mu|\alpha) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$
  - 限制条件: $0 \le \mu_k \le 1, \sum_k \mu_k = 1$
  - $\alpha = (\alpha_1, \cdots, \alpha_K)^T$ 是分布的参数 (Fig 2.4)
- 归一化形式: $\text{Dir}(\mu|\alpha) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$
  - $\alpha_0 = \sum_{k=1}^{K} \alpha_k$
- Gamma 函数 (Ex 1.17)
  - $\Gamma(x) \equiv \int_0^\infty u^{x-1} e^{-u} du$
  - 用于保证 Dirichlet 分布的归一化性质 $\int_0^1 \text{Dir}(\mu|\alpha) \text{d}\mu = 1$
- 后验分布: 依然是 Dirichlet 分布的形式

$$
\begin{aligned}
p(\mu|\alpha, \mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|\alpha) \\
&\propto \text{Multi}(m_1, m_2, \cdots, m_K|\mu, N)\text{Dir}(\mu|\alpha) \\
&\propto \prod_{k=1}^{K} \mu_k^{m_k}\mu_k^{\alpha_k-1} \\
&\propto \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1}
\end{aligned}
$$

- 归一化的后验分布: 其中 α_k 可以看作先验概率中 $x_k = 1$ 的有效观测数。

$$
\begin{aligned}
p(\mu|\alpha, \mathcal{D}) &= \text{Dir}(\mu|\alpha + m) \\
&= \frac{\Gamma(\sum_{k=1}^{K} \alpha_k + N)}{\Gamma(\alpha_1 + m_1)\cdots\Gamma(\alpha_K + m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1}
\end{aligned}
$$

2.3. 高斯分布: 连续分布

高斯分布，也称为正态分布。

一元高斯分布:

$$
\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]
$$

- 均值: μ
- 方差: σ^2
- 精度: $\lambda = 1/\sigma$

多元高斯分布:

$$
\mathcal{N}(\mathsf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathsf{x}-\mu)^T\Sigma^{-1}(\mathsf{x}-\mu)\right]
$$

- 均值向量: $\mu \in \mathcal{R}^D$
- 协方差矩阵: $\Sigma \in \mathcal{R}^{D\times D}$
  - 因为 Σ 控制了高斯分布下 x 的协方差
- 行列式: $|\Sigma|$

高斯分布的性质

- 一元高斯分布使熵取得最大值；多元高斯分布也使熵取得最大值。
- 中心极限定理 (central limit theorem) : 一组随机变量之和的概率分布随着求和公式中变量计算项的数量的增加而逐渐趋向高斯分布。
- 高斯分布的几何形式
  - 二次型: $\Delta^2 = (\mathsf{x} - \mu)^T \Sigma^{-1} (\mathsf{x} - \mu)$
    * 这个二次型出现在高斯分布的指数位置
    * 在 x 空间的曲面上，如果二次型是常数，则高斯分布也是常数
  - Δ 是 x 和 μ 之间的马氏距离 (Mahalanobis Distance)。
    * 当 Σ 是单位矩阵时，Δ 为欧氏距离。
- 两个高斯分布的的卷积
  - 卷积的均值是两个高斯分布的均值的和
  - 卷积的协方差是两个高斯分布的协方差的和

多元高斯分布的性质

- 前提条件: 两组变量是联合高斯分布
- 条件概率分布: 以一组变量为条件，另一组变量还是高斯分布
- 边缘概率分布: 任何一个变量的边缘分布也是高斯分布

多元高斯分布的分解

- 协方差矩阵的特征向量方程: $\Sigma \mathsf{u}_i = \lambda_i \mathsf{u}_i$, $i = 1, \cdots, D$
- $\Sigma = \sum_{i=1}^{D} \lambda_i \mathsf{u}_i \mathsf{u}_i^T$ 是实对称矩阵，得: $\mathsf{u}_i^T \mathsf{u}_i = I_{ij}$, $\lambda_i \in \mathcal{R}$
  - I_{ij} 是指示函数

$$I_{ij} = \begin{cases} 1, & i = j \\ 0, & \text{others} \end{cases}$$

- $\Sigma^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathsf{u}_i \mathsf{u}_i^T$
- $\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$
  - $y_i = \mathsf{u}_i^T (\mathsf{x} - \mathsf{u})$
  - $\{y_i\}$ 是一个新的坐标系统 (坐标轴旋转) [^Aapo,2007]
    * 新的坐标系统是由原始的 x_i 坐标经过平移和旋转后形成的单位正交向量 u_i 定义的
    * 定义新的坐标系统的向量 $\mathsf{y} = (y_1, \cdots, y_D)^T$
- $\mathsf{y} =$ U (x - μ)

  - U 是正交矩阵，即 $\mathsf{U}\mathsf{U}^T = I = \mathsf{U}^T\mathsf{U}$

- 正定矩阵: 特征值 λ_i 严格大于零的矩阵。
  - 如果特征矩阵不是正定矩阵，则定义的概率分布无法被归一化
  - 半正定矩阵: 是一种奇异矩阵，处理方式参考 (Ch 12)

- 由 y_j 定义的新坐标系下的高斯分布
  - $J_{ij} = \frac{\partial x_i}{\partial y_j} = \mathsf{U}_{ji}$
  - $|J|^2 = |\mathsf{U}^T|^2 = |\mathsf{U}^T||\mathsf{U}| = |\mathsf{U}^T\mathsf{U}| = |I| = 1$
  - $|\Sigma|^{1/2} = \prod_{j=1}^{D} \lambda_j^{1/2}$
  - D 元高斯分布分解为 D 个一元高斯分布的乘积
    * 特征向量定义了一个新的旋转、平移后的坐标系
    * 坐标系中的联合概率分布分解为独立分布的乘积

$$p(\mathsf{y}) = p(\mathsf{x})|J| = \prod_{j=1}^{D} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\{\frac{y_j^2}{2\lambda_j}\}$$

$$\int p(\mathsf{y})\mathsf{dy} = \prod_{j=1}^{D} \int_{-\infty}^{+\infty} \frac{1}{(2\pi\lambda_j)^{1/2}} \exp\{\frac{y_j^2}{2\lambda_j}\}\mathsf{dy}_j = 1$$

$$\mathbb{E}[\mathsf{x}] = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \int \exp\left\{-\frac{1}{2}(\mathsf{x}-\mu)^T\Sigma^{-1}(\mathsf{x}-\mu)\right\}\mathsf{x}\ \mathsf{dx}$$

$$\mathbb{E}[\mathsf{x}\mathsf{x}^T] = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \int \exp\left\{-\frac{1}{2}(\mathsf{x}-\mu)^T\Sigma^{-1}(\mathsf{x}-\mu)\right\}\mathsf{x}\mathsf{x}^T\ \mathsf{dx}$$

$$\mathsf{z} = \mathsf{x} - \mu = \sum_{j=1}^{D} y_j \mathsf{u}_j = \sum_{j=1}^{D} \mathsf{u}_j^T\mathsf{z}\mathsf{u}_j$$

$$\mathbb{E}[\mathsf{x}] = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \int \exp\left\{-\frac{1}{2}\mathsf{z}^T\Sigma^{-1}\mathsf{z}\right\}(\mathsf{z}+\mu)\mathsf{dz}$$

$$\mathbb{E}[\mathsf{x}\mathsf{x}^T] = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \int \exp\left\{-\frac{1}{2}\mathsf{z}^T\Sigma^{-1}\mathsf{z}\right\}(\mathsf{z}+\mu)(\mathsf{z}+\mu)^T\mathsf{dz}$$

$$\begin{aligned}
&\frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\int \exp\left\{-\frac{1}{2}\mathsf{z}^T\Sigma^{-1}\mathsf{z}\right\}\mathsf{z}\mathsf{z}^T\mathsf{dz} \\
&\qquad = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\Sigma|^{1/2}}\sum_{i=1}^{D}\sum_{j=1}^{D}\mathsf{u}_i\mathsf{u}_j^T\int \exp\left\{-\sum_{k=1}^{D}\frac{y_k^2}{2\lambda_k}\right\}y_iy_j\mathsf{dy} \\
&\qquad = \sum_{i=1}^{D}\mathsf{u}_i\mathsf{u}_j^T\lambda_i \\
&\qquad = \Sigma
\end{aligned}$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{+\infty}\mathcal{N}(x|\mu,\sigma^2)x^2\mathrm{d}x = \mu^2+\sigma^2 \qquad \text{(二阶矩：1.50)}$$

$$\mathbb{E}[\mathsf{x}\mathsf{x}^T] = \mu\mu^T+\Sigma$$

$$\text{var}\,[\mathsf{x}] = \mathbb{E}[(\mathsf{x}-\mathbb{E}[\mathsf{x}])(\mathsf{x}-\mathbb{E}[\mathsf{x}])^T] = \Sigma$$

高斯分布的局限性

- 多元高斯分布中自由参数的数量随着维数的平方增长。
  - μ 有 D 个独立参数
  - Σ 有 $\frac{D(D+1)}{2}$ 个独立参数
  - 通过限协方差矩阵的形式可以简化计算，但是会限制概率密度的形式
    * 对角协方差矩阵: $\Sigma = \text{diag}(\sigma_i^2)$
    * 正比于单位矩阵: $\Sigma = \sigma^2 I$
- 高斯分布是单峰的，不能很好地描述多峰分布。
  - 引入隐变量 (latent variable, hidden variable, unobserved variable)
  - 离散型隐变量: 混合高斯分布描述多峰分布 (Sec 2.3.9)
  - 连续型隐变量: 自由参数可以被设计成与数据空间的维度无关
    * 高斯版本 Markov 随机场: 反映空间中像素组织的结构
    * 线性动态系统: 对时序数据建模
    * 概率图模型: 用来表达复杂的概率分布 (Ch 8)

2.3.1 条件高斯分布

前提条件

- 假设 x 服从高斯分布 $\mathcal{N}(\mathsf{x}|\mu,\Sigma)$ 的 D 维向量
- 将 x 划分为两个不相交的子集 x_a x_b
- $\mathsf{x}=\begin{bmatrix}\mathsf{x}_a\\ \mathsf{x}_b\end{bmatrix}$
- $\mu=\begin{bmatrix}\mu_a\\ \mu_b\end{bmatrix}$
- $\Sigma=\begin{bmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{bmatrix}=\begin{bmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ab}^T & \Sigma_{bb}\end{bmatrix}$
- $\Lambda\equiv\Sigma^{-1}=\begin{bmatrix}\Lambda_{aa} & \Lambda_{ab}\\ \Lambda_{ba} & \Lambda_{bb}\end{bmatrix}=\begin{bmatrix}\Lambda_{aa} & \Lambda_{ab}\\ \Lambda_{ab}^T & \Lambda_{bb}\end{bmatrix}$

求解条件概率分布

$$\begin{aligned}-\frac{1}{2}(\mathsf{x}-\mu)^T\Sigma^{-1}(\mathsf{x}-\mu)=&\\ &-\frac{1}{2}(\mathsf{x}_a-\mu_a)^T\Lambda_{aa}(\mathsf{x}_a-\mu_a)-\frac{1}{2}(\mathsf{x}_a-\mu_a)^T\Lambda_{ab}(\mathsf{x}_b-\mu_b)\\ &-\frac{1}{2}(\mathsf{x}_b-\mu_b)^T\Lambda_{ab}(\mathsf{x}_a-\mu_a)-\frac{1}{2}(\mathsf{x}_b-\mu_b)^T\Lambda_{bb}(\mathsf{x}_b-\mu_b)\end{aligned}\tag{2.70}$$

- 把上面的公式看成 x_a 的函数，这仍然是一个二次型，因此对应的条件分布 $p(\mathsf{x}_a|\mathsf{x}_b)$ 是高斯分布

如何确定这个二次型对应的高斯分布的均值与方差？

前提条件

-

$$-\frac{1}{2}(\mathsf{x}-\mu)^T\Sigma^{-1}(\mathsf{x}-\mu)=-\frac{1}{2}\mathsf{x}^T\Sigma^{-1}\mathsf{x}+\mathsf{x}^T\Sigma^{-1}\mu+\text{const}\tag{2.70}$$

- 基于「配平方法」：与常规的高斯分布 $\mathcal{N}(\mathsf{x}|\mu,\Sigma)$ 的指数项对应
  - const : 表示与 x 无关的项
  - (Eq 2.70) 中的项与 (Eq 2.71) 中的项对应，可以得到所需要的均值和协方差
- (Eq 2.70) 中 x_a 看作函数中的变量，x_b 看作函数中的常数

  - 找出 x_a 的二阶项，得: $\frac{1}{2}\mathsf{x}_a^T\Lambda_{aa}\mathsf{x}_a$
    * 得到 $p(\mathsf{x}_a|\mathsf{x}_b)$ 的协方差: $\Sigma_{a|b}=\Lambda_{aa}^{-1}$
  - 找出 x_a 的一阶项，得: $\mathsf{x}_a^T\{\Lambda_{aa}\mu_a-\Lambda_{ab}(\mathsf{x}_b-\mu_b)\}$
    * 辅助条件: $\Lambda_{ba}^T=\Lambda_{ab}$
  - x_a 的系数 $\Sigma_{a|b}^{-1}\mu_{a|b}$
    * $\mu_{a|b}=\Sigma_{a|b}\{\Lambda_{aa}\mu_a-\Lambda_{ab}(\mathsf{x}_b-\mu_b)\}=\mu_a-\Lambda_{aa}^{-1}\Lambda_{ab}(\mathsf{x}_b-\mu_b)$

- 分块矩阵
  - $M=(A-BD^{-1}C)^{-1}$
  - M^{-1} 是 (Eq 2.76) 的舒尔补 (Schur complement)

$$\begin{bmatrix}A & B\\ C & D\end{bmatrix}^{-1}=\begin{bmatrix}M & -MBD^{-1}\\ -D^{-1}CM & D^{-1}+D^{-1}CMBD^{-1}\end{bmatrix} \tag{2.76}$$

- 定义: $\Sigma^{-1}=\Lambda$ 即 $\begin{bmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{bmatrix}^{-1}=\begin{bmatrix}\Lambda_{aa} & \Lambda_{ab}\\ \Lambda_{ba} & \Lambda_{bb}\end{bmatrix}$
  - $\Lambda_{aa}=(\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}$
  - $\Lambda_{ab}=-(\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}$
- 条件概率分布 $p(\mathsf{x}_a|\mathsf{x}_b)$
  - 均值: $\mu_{a|b}=\mu_a+\Sigma_{ab}\Sigma_{bb}^{-1}(\mathsf{x}_b-\mu_b)$
  - 方差: $\Sigma_{a|b}=\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$

2.3.2 边缘高斯分布

求边缘概率分布的积分公式 (Eq 2.83) : $p(\mathsf{x}_a)=\int p(\mathsf{x}_a,\mathsf{x}_b)\,\mathrm{d}\mathsf{x}_b$

从 (Eq 2.70) 中选出涉及到 x_b 的项得到 (Eq 2.84)

$$\begin{aligned}-\frac{1}{2}(\mathsf{x}-\mu)^T\Sigma^{-1}(\mathsf{x}-\mu)=&\\ &-\frac{1}{2}(\mathsf{x}_a-\mu_a)^T\Lambda_{aa}(\mathsf{x}_a-\mu_a)-\frac{1}{2}(\mathsf{x}_a-\mu_a)^T\Lambda_{ab}(\mathsf{x}_b-\mu_b)\\ &-\frac{1}{2}(\mathsf{x}_b-\mu_b)^T\Lambda_{ab}(\mathsf{x}_a-\mu_a)-\frac{1}{2}(\mathsf{x}_b-\mu_b)^T\Lambda_{bb}(\mathsf{x}_b-\mu_b)\end{aligned} \tag{2.70}$$

$$-\frac{1}{2}\mathsf{x}_b\Lambda_{bb}\mathsf{x}_b+\mathsf{x}_b\mathsf{m}=-\frac{1}{2}(\mathsf{x}_b-\Lambda_{bb}^{-1}\mathsf{m})^T\Lambda_{bb}(\mathsf{x}_b-\Lambda_{bb}^{-1}\mathsf{m})+\frac{1}{2}\mathsf{m}^T\Lambda_{bb}^{-1}\mathsf{m} \tag{2.84}$$

$$\mathsf{m}=\Lambda_{bb}\mu_b-\Lambda_{ba}(\mathsf{x}_a-\mu_a)$$

与 x_b 相关的项转化为高斯分布的标准二次型 (Eq 2.84 第一项) + 只与 x_a 相关的项 (Eq 2.84 第二项)。取二次型项 (Eq 2.84 第一项) 带入求边缘概率分布的积分公式 (Eq 2.83)，得

$$\int\exp\left\{-\frac{1}{2}(\mathsf{x}_b-\Lambda_{bb}^{-1}\mathsf{m})^T\Lambda_{bb}(\mathsf{x}_b-\Lambda_{bb}^{-1}\mathsf{m})\right\}\mathrm{d}\mathsf{x}_b$$

基于标准的多元高斯概率分布公式 (Eq 2.43)，再次基于「配平方法」

将 x_b 积分得：

(Eq 2.84) 和 (Eq 2.70) 中与 x_a 相关的项相加，const 表示与 x_a 无关的项

$$\begin{aligned}&\frac{1}{2}\mathsf{m}^T\Lambda_{bb}^{-1}\mathsf{m}-\frac{1}{2}\mathsf{x}_a\Lambda_{aa}\mathsf{x}_a+\mathsf{x}_a^T(\Lambda_{aa}\mu_a+\Lambda_{ab}\mu_b)+\text{const}\\&\qquad=\frac{1}{2}[\Lambda_{bb}\mu_b-\Lambda_{ba}(\mathsf{x}_a-\mu_a)]^T\Lambda_{bb}^{-1}[\Lambda_{bb}\mu_b-\Lambda_{ba}(\mathsf{x}_a-\mu_a)]\\&\qquad-\frac{1}{2}\mathsf{x}_a\Lambda_{aa}\mathsf{x}_a+\mathsf{x}_a^T(\Lambda_{aa}\mu_a+\Lambda_{ab}\mu_b)+\text{const}\end{aligned}$$

继续基于「配平方法」，得出边缘概率分布 $p(\mathsf{x}_a)$ 的参数

- 协方差矩阵：$\Sigma_a=(\Lambda_{aa}-\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}$
- 均值：$\mu_a=\Sigma_a(\Lambda_{aa}-\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}\mu_a$
- 协方差矩阵与精度矩阵 (Eq 2.69) 的关系：$\begin{bmatrix}\Sigma_{aa}&\Sigma_{ab}\\\Sigma_{ba}&\Sigma_{bb}\end{bmatrix}=\begin{bmatrix}\Lambda_{aa}&\Lambda_{ab}\\\Lambda_{ba}&\Lambda_{bb}\end{bmatrix}^{-1}$
- 根据分块矩阵的逆矩阵的关系 (Eq 2.76) 得：$\Sigma_{aa}=(\Lambda_{aa}-\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})^{-1}$
- 结论：配平方法得出的结果与实际结果相符
  - 均值 (Eq 2.92) ：$\mathbb{E}[\mathsf{x}_a]=\mu_a$
  - 协方差 (Eq 2.93) ：$\mathrm{cov}[\mathsf{x}_a]=\Sigma_{aa}$
  - 使用分块协方差矩阵表示边缘概率分布时，公式的形式可以得到简化
  - 基于条件概率分布时，使用分块精度矩阵表示，公式的形式还能进一步简化

分块高斯的边缘分布和条件分布的总结如下

- 前提条件

  - 给定联合高斯分布：$\mathcal{N}(\mathsf{x}|\mu,\Sigma)$
    * $\Lambda \equiv \Sigma^{-1}$
    * $\mathsf{x} = \begin{bmatrix} \mathsf{x}_a \\ \mathsf{x}_b \end{bmatrix}$
    * $\mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}$
    * $\Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$
    * $\Lambda = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}$
  - 条件概率分布
    * $p(\mathsf{x}_a|\mathsf{x}_b) = \mathcal{N}(\mathsf{x}_a|\mu_{a|b}, \Lambda_{aa}^{-1})$
    * $\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(\mathsf{x}_b - \mu_b)$
  - 边缘概率分布
    * $p(\mathsf{x}_a) = \mathcal{N}(\mathsf{x}_a|\mu_a, \Sigma_{aa})$

2.3.3 高斯变量的贝叶斯定理

前提条件

- 高斯边缘概率分布：$p(\mathsf{x}) = \mathcal{N}(\mathsf{x}|\mu, \Lambda^{-1})$
- 高斯条件概率分布：$p(\mathsf{y}|\mathsf{x}) = \mathcal{N}(\mathsf{y}|\mathsf{Ax + b}, \mathsf{L}^{-1})$
- 控制参数
  - 控制均值的参数：μ, A, b
  - 精度矩阵：Λ, L
  - 如果 $\mathsf{x} \in \mathcal{R}^M, \mathsf{y} \in \mathcal{R}^D$, 则 $\mathsf{A} \in \mathcal{R}^{(D\times M)}$

求解高斯联合概率分布

- 定义 $\mathsf{z} = \begin{bmatrix} \mathsf{x} \\ \mathsf{y} \end{bmatrix}$
- 联合概率分布的对数：

$$\ln p(\mathsf{z}) = \ln p(\mathsf{x}) + \ln p(\mathsf{y}) = -\frac{1}{2}(\mathsf{x}-\mu)^T\Lambda(\mathsf{x}-\mu) - \frac{1}{2}(\mathsf{y}-\mathsf{Ax - b})^T\mathsf{L}(\mathsf{y}-\mathsf{Ax - b}) + const \tag{2.102}$$

- 基于「配平方法」

$$
\begin{aligned}
-\frac{1}{2}\mathsf{x}^T(\Lambda+\mathsf{A}^T\mathsf{L}\mathsf{A})\mathsf{x}-\frac{1}{2}\mathsf{y}^T\mathsf{L}\mathsf{y}+\frac{1}{2}\mathsf{y}^T\mathsf{L}\mathsf{A}\mathsf{x}+\frac{1}{2}\mathsf{x}^T\mathsf{A}^T\mathsf{L}\mathsf{y} \\
&= -\frac{1}{2}\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix}^T\begin{bmatrix}\Lambda+\mathsf{A}^T\mathsf{L}\mathsf{A} & -\mathsf{A}^T\mathsf{L}\\ -\mathsf{L}\mathsf{A} & \mathsf{L}\end{bmatrix}\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix} \\
&= -\frac{1}{2}\mathsf{z}^T\mathsf{R}\mathsf{z}
\end{aligned}
$$

- 精度矩阵：$\mathsf{R}=\begin{bmatrix}\Lambda+\mathsf{A}^T\mathsf{L}\mathsf{A} & -\mathsf{A}^T\mathsf{L}\\ -\mathsf{L}\mathsf{A} & \mathsf{L}\end{bmatrix}$
- 协方差矩阵：$\text{cov}[\mathsf{z}]=\mathsf{R}^{-1}=\begin{bmatrix}\Lambda^{-1} & \Lambda^{-1}\mathsf{A}^T\\ \mathsf{A}\Lambda^{-1} & \mathsf{L}^{-1}+\mathsf{A}\Lambda^{-1}\mathsf{A}^T\end{bmatrix}$
- 基于「配平方法」，寻找 (Eq 2.102) 中的线性项与 (Eq 2.71) 配对，再基于 cov [z] 得

$$
\mathsf{x}^T\Lambda\mu-\mathsf{x}^T\mathsf{A}^T\mathsf{L}\mathsf{b}+\mathsf{y}^T\mathsf{L}\mathsf{b}=\begin{bmatrix}\mathsf{x}\\ \mathsf{y}\end{bmatrix}^T\begin{bmatrix}\Lambda\mu-\mathsf{A}^T\mathsf{L}\mathsf{b}\\ \mathsf{L}\mathsf{b}\end{bmatrix}
$$

- 均值：$\mathbb{E}[\mathsf{z}]=\mathsf{R}^{-1}\begin{bmatrix}\Lambda\mu-\mathsf{A}^T\mathsf{L}\mathsf{b}\\ \mathsf{L}\mathsf{b}\end{bmatrix}=\begin{bmatrix}\mu\\ \mathsf{A}\mu+\mathsf{b}\end{bmatrix}$

求解边缘分布 $p(\mathsf{y})$，通过对 x 求积分得

- 均值：$\mathbb{E}[\mathsf{y}]=\mathsf{A}\mu+\mathsf{b}$
- 协方差：$\text{cov}[\mathsf{y}]=\mathsf{L}^{-1}+\mathsf{A}\Lambda^{-1}\mathsf{A}^T$

求解条件分布 $p(\mathsf{y}|\mathsf{x})$

- 均值：$\mathbb{E}[\mathsf{x}|\mathsf{y}]=(\Lambda+\mathsf{A}^T\mathsf{L}\mathsf{A})^{-1}\{\mathsf{A}^T\mathsf{L}\,(\,\mathsf{y}\text{-}\mathsf{b}\,)+\Lambda\mu\}$
- 协方差：$\text{cov}[\mathsf{x}|\mathsf{y}]=(\Lambda+\mathsf{A}^T\mathsf{L}\mathsf{A})^{-1}$

贝叶斯公式推导

- 前提条件：给定 x 的边缘概率分布和条件概率分布
  - (Eq 2.113) : $p(\mathsf{x})=\mathcal{N}(\mathsf{x}|\mu,\Lambda^{-1})$
  - (Eq 2.114) : $p(\mathsf{y}|\mathsf{x})=\mathcal{N}(\mathsf{y}|\mathsf{A}\mathsf{x}+\mathsf{b},\,\mathsf{L}^{-1})$
- 推导结果：求解 y 的边缘概率分布和条件概率分布
  - (Eq 2.115) : $p([y])=\mathcal{N}(\mathsf{y}|\mathsf{A}\mu+\mathsf{b},\mathsf{L}^{-1}+\mathsf{A}\Lambda^{-1}\mathsf{A}^T)$
  - (Eq 2.116) : $p(\mathsf{x}|\mathsf{y})=\mathcal{N}(\mathsf{x}|\Sigma\{\mathsf{A}^T\mathsf{L}\,(\,\mathsf{y}\text{-}\mathsf{b}\,)+\Lambda\mu\},\Sigma)$
    - $\Sigma=(\Lambda+\mathsf{A}^T\mathsf{L}\mathsf{A})^{-1}$

2.3.4 高斯分布的最大似然估计

前提条件

- 给定数据集 $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^T$
- 观测 $\{\mathbf{x}_n\}$ 是独立地从多元高斯分中 k 抽取的

使用最大似然估计分布的参数

- 对数似然函数
  - $\ln p(\mathbf{X}|\mu, \Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \mu)^T\Sigma^{-1}(\mathbf{x} - \mu)$
- 高斯分布的充分统计量
  - $\sum_{n=1}^{N}\mathbf{x}_n$
  - $\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T$
  - 似然函数对数据集的依赖通过以上两个量体现
- 对数似然函数对于 μ 的导数
  - $\frac{\partial}{\partial\mu}\ln p(\mathbf{X}|\mu, \Sigma) = \sum_{n=1}^{N}\Sigma^{-1}(\mathbf{x}_n - \mu)$
- 令这个导数等于零，得到均值的最大似然估计（即数据点的观测集合的均值)
  - $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$
  - $\Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \mu_{ML})(\mathbf{x}_n - \mu_{ML})^T$
- 估计真实概率分布下最大似然解的期望
  - 均值的估计是无偏的：$\mathbb{E}[\mu_{ML}] = \mu$
  - 协方差的估计是有偏的：$\mathbb{E}[\Sigma_{ML}] = \frac{N-1}{N}\Sigma$
  - 定义一个估计值：$\tilde{\Sigma} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \mu_{ML})(\mathbf{x}_n - \mu_{ML})^T$
  - 修正协方差的估计：$\mathbb{E}[\tilde{\Sigma}] = \Sigma$

2.3.5 最大似然的顺序估计

顺序估计：每次处理一个数据点，然后丢弃这个点，适合在线应用和数据集非常大的情况。

$$\begin{aligned}\mu_{ML}^{(N)} &= \frac{1}{N}\sum_{n=1}^{N}\mathsf{x}_n \\ &= \frac{1}{N}\mathsf{x}_N + \frac{1}{N}\sum_{n=1}^{N-1}\mathsf{x}_n \\ &= \frac{1}{N}\mathsf{x}_N + \frac{N-1}{N}\mu_{ML}^{(N-1)} \\ &= \mu_{ML}^{(N-1)} + \frac{1}{N}(\mathsf{x}_N - \mu_{ML}^{(N-1)})\end{aligned}$$

公式推导只是说明最后一次估计只与最后一个数据点和前一次估计有关。

Robbins-Monro 算法：(通用的顺序学习算法)

前提条件

- 联合概率分布：$p(z,\theta)$ 控制着一对随机变量 z 和 θ
- 回归函数：$f(\theta) \equiv \mathbb{E}[z|\theta] = \int zp(z|\theta)\mathsf{d}z$
- 条件方差是有穷的：$\mathbb{E}[(z-f)^2|\theta] < \infty$
- 当 $\theta > \theta^*$ 时 $f(\theta) > 0$；当 $\theta < \theta^*$ 时 $f(\theta) < 0$

收敛公式

- 根 θ^* 顺序估计的序列：$\theta^{(N)} = \theta^{(N-1)} - a_{N-1}z(\theta^{(N-1)})$
  - $z(\theta^{(N)})$ 是当 θ 取值为 $\theta^{(N)}$ 时 z 的观测值
  - 系数 a_N 是一个正数序列 $\{a_N\}$，并且满足以下条件
    * $\lim_{N\to\infty} a_N = 0$：确保后续的修正的幅度会逐渐变小，保证学习过程收敛于一个极限值
    * $\sum_{N=1}^{\infty} a_N = \infty$：确保算法不会收敛不到根的值
    * $\sum_{N=1}^{\infty} a_N^2 < \infty$：保证累计的噪声具有一个有限的方差，不会导致收敛失败

算法案例：最大似然问题

前提条件

- 最大似然解 θ_{ML} 是负对数似然函数的一个驻点
- 满足条件：$\frac{\partial}{\partial\theta}\{\frac{1}{N}\sum_{n=1}^{N} - \ln p(x_n|\theta)\}|_{\theta_{ML}} = 0$
- 将求导与求和顺序交换，对公式取极限得

$$-\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\ln p(x_n|\theta) = \mathbb{E}_x[-\frac{\partial}{\partial\theta}\ln p(x|\theta)]$$

- 顺序算法的计算公式得

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1} \frac{\partial}{\partial \theta^{(N-1)}} \left[-\ln p(x_N|\theta^{(N-1)}) \right]$$

算法案例：高斯分布的顺序估计公式

前提条件

- 随机变量 z 的形式：$z = -\frac{\partial}{\partial \mu_{ML}} \ln p(x|\mu_{ML}, \sigma^2) = -\frac{1}{\sigma^2}(x - \mu_{ML})$
  - 均值：$-(\mu - \mu_{ML})/sigma^2$

2.3.6 高斯分布的贝叶斯推断

一元高斯随机变量

- 一组 N 次观测 $\mathbf{x} = \{x_1, \cdots, x_N\}$

方差已知，推断均值，均值的先验可以选高斯分布。

- 方差已知：σ^2
- 基于似然函数 $p(\mathbf{x}|\mu)$ 推断均值 μ
  - 注：似然函数不是 μ 的概率密度，没有归一化

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}$$

- 先验分布 (共轭：高斯分布) ：$p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$
- 后验分布：$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu)$
  - 基于「配平方法」得：$p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$
    - $\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{ML}$
    - $\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$
    - $\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$ 是 μ 的最大似然解
  - 分解最后一个数据点的贡献的公式：$p(\mu|\mathbf{x}) \propto [p(\mu)\prod_{n=1}^{N-1} p(x_n|\mu)]p(x_N|\mu)$

均值已知，推断精度，使用精度 λ 更易计算，精度的先验可以选 Gamma 分布。

- 似然函数

$$p(\mathsf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

- 先验分布：$\mathsf{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$
  - 均值：$\mathbb{E}[\lambda] = \frac{a}{b}$
  - 方差：$\mathsf{var}[\lambda] = \frac{a}{b^2}$
  - Gamma 函数 (Eq 1.141) ：$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}\mathsf{d}u$
  - 共轭先验：可以看作有效假想数据点
- 后验分布：

$$p(\lambda|\mathsf{x}) \propto p(\mathsf{x}|\lambda)\mathsf{Gam}(\lambda|a_0,b_0) \propto \lambda^{N/2}\lambda^{a_0-1} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2 - b_0\lambda\right\}$$

- 将后验分布看作 Gamma 分布：$\mathsf{Gam}(\lambda|a_N,b_N)$
  - $a_N = a_0 + \frac{N}{2}$
  - $b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma_{ML}^2$
  - σ_{ML}^2 是方差的最大似然估计

均值未知，精度未知，推断均值和方差，与均值和精度相关的先验分布可以选 Normal-Gamma 分布或 Gauss-Gamma 分布 (Eq 2.154)。

- 似然函数

$$\begin{aligned} p(\mathsf{x}|\mu,\lambda) &= \prod_{n=1}^{N}(\frac{\lambda}{2\pi})^{1/2} \exp\left\{-\frac{\lambda}{2}(x_n-\mu)^2\right\} \\ &\propto \left[\lambda^{1/2}\exp(-\frac{\lambda\mu^2}{2})\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\} \end{aligned}$$

- 假设先验分布形式：(c, d, β 都是常数)

$$\begin{aligned} p(\mu,\lambda) &\propto [\lambda^{1/2}\exp(-\frac{\lambda\mu^2}{2})]^\beta \exp\{c\lambda\mu - d\lambda\} \\ &= \exp\{-\frac{\beta\lambda}{2}(\mu-\frac{c}{\beta})^2\}\lambda^{\beta/2}\exp\{-(d-\frac{c^2}{2\beta})\lambda\} \end{aligned}$$

- 归一化的先验概率的形式：($\mu_0 = c/\beta, a = (1+\beta)/2, b = d - \frac{c^2}{2\beta}$ 是常数)

$$p(\mu, \lambda) = p(\mu|\lambda)p(\mu) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\text{Gam}(\lambda|a, b) \tag{2.154}$$

多元高斯随机变量

- 一组 N 次观测 $\mathsf{X} = \{\mathsf{x}_1, \cdots, \mathsf{x}_N\}$

方差已知，推断均值，均值的先验可以选多元高斯分布。

- 先验分布：$\mathcal{N}(\mathsf{x}|\mu, \Lambda^{-1})$

均值已知，推断精度，精度的先验可以选 Wishart 分布。

- 先验分布：$\mathcal{W}(\Lambda|\mathsf{W}, \nu) = B|\Lambda|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\text{Tr}(\mathsf{W}^{-1}\Lambda)\right)$
  - 分布的自由度的数量：μ
  - $\mathsf{W} \in \mathcal{R}^{D\times D}$
  - $\text{Tr}(\cdot)$ 表示矩阵的迹
  - 归一化系数：$B(\mathsf{W}, \nu) = |\mathsf{W}^{-\nu/2}\left(2^{\nu D/2}\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma(\frac{\nu+1-i}{2})\right)^{-1}$

均值未知，精度未知，推断均值和精度，与均值和精度相关的先验分布可以选 Normal-Wishart 分布或 Gauss-Wishart 分布。

- 先验分布：$p(\mu, \Lambda|\mu_0, \beta, \mathsf{W}, \nu) = \mathcal{N}(\mu|\mu_0, (\beta\Lambda)^{-1})\mathcal{W}(\Lambda|\mathsf{W}, \nu)$

2.3.7 学生 t 分布（ Student's t-distribution ）

前提条件

- Gauss 分布的精度的共轭先验是 Gamma 分布
  - 一元高斯分布：$\mathcal{N}(x|\mu, \tau^{-1})$
  - Gamma 先验分布：$\text{Gam}(\tau|a, b)$

边缘分布：对精度积分

$$
\begin{aligned}
p(x|\mu,a,b) &= \int_0^\infty \mathcal{N}(x|\mu,\tau^{-1})\mathsf{Gam}(\tau|a,b)\mathsf{d}\tau \\
&= \int_0^\infty (\frac{\tau}{2\pi})^{\frac{1}{2}} \exp\left\{-\tau\frac{(x-\mu)^2}{2}\right\}\frac{1}{\Gamma(a)}b^a\tau^{a-1}\exp(-b\tau)\mathsf{d}\tau \\
&= \int_0^\infty (\frac{\tau}{2\pi})^{\frac{1}{2}} \exp\left\{-\tau(b+\frac{(x-\mu)^2}{2})\right\}\frac{1}{\Gamma(a)}b^a\tau^{a-1}\mathsf{d}\tau \\
&= \Gamma(a+\frac{1}{2})\frac{b^a}{\Gamma(a)}(\frac{1}{2\pi})\left[b+\frac{(x-\mu)^2}{2}\right]^{(-a-\frac{1}{2})}
\end{aligned}
$$

使用变量替换技术对指数积分 $z=\tau(b+\frac{(x-\mu)^2}{2})$

定义新的参数：$\nu=2a, \lambda=a/b$，得到学生 t 分布

$$
\mathsf{St}(x|\mu,\lambda,\nu)=\frac{\Gamma(\frac{\nu}{2}+\frac{1}{2})}{\Gamma(\frac{\nu}{2})}(\frac{\lambda}{\pi\nu})^{1/2}[1+\lambda\frac{(x-\mu)^2}{\nu}]^{-\frac{\nu}{2}-\frac{1}{2}}
$$

- 当自由度 $\nu=1$ 时，t 分布就变成为柯西分布（Cauchy distribution）；
- 当自由度 $\nu\to\infty$ 时，t 分布就变成为高斯分布。
- 学生 t 分布可以通过将无限多个同均值不同精度的高斯分布相加得到，即可以表示成无限的高斯混合模型
- 学生 t 分布有着比高斯分布更长的「尾巴」，因此具有很好的鲁棒性（robustness），对于离群点（outliers）的出现不像高斯分布那么敏感。

2.3.8 周期变量

前提条件

- 周期变量的观测数据集 $\mathcal{D}=\{\theta_1,\cdots,\theta_N\}$
- 观测 θ 的单位是弧度
- 将观测描述为单位圆上的点，即一个二维单位向量 $\mathcal{D}=\{\mathsf{x}_1,\cdots,\mathsf{x}_N\}$
  - $||\mathsf{x}_n||=1, n=1,\cdots,N$

均值求解

- $\mathbb{E}[\theta]=(\theta_1+\cdots+\theta_N)/N$ 需要依赖于选择的坐标系
- $\mathbb{E}[\mathsf{x}]=\frac{1}{N}\sum_{n=1}^N \mathsf{x}_n$
  - $x_1=\bar{r}\cos\bar{\theta}=\frac{1}{N}\sum_{n=1}^N\cos\theta_n$

- $x_2 = \bar{r}\sin\bar{\theta} = \frac{1}{N}\sum_{n=1}^{N}\sin\theta_n$

$$\mathbb{E}[\theta] = \tan^{-1}\left\{\frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n}\right\}$$

Von Mises 分布: 也叫环形正态分布 (circular normal distribution)，是高斯分布对于周期变量的推广。

- 优点：类似于高斯分布，方便计算
- 缺点：类似于高斯分布，是单峰分布。可以将多个 Von Mises 分布混合来处理多峰问题。

一元 Von Mises 分布 $p(\theta)$

- 满足下面三个条件
  - $p(\theta) \geq 0$
  - $\int_0^{2\pi} p(\theta)\mathrm{d}\theta = 1$
  - $p(\theta + 2\pi) = p(\theta)$

构建过程

- 两个变量 $\mathbf{x} = (x_1, x_2)$ 的高斯分布
  - 均值 : $\mu = (\mu_1, \mu_2)$
  - 协方差 : $\Sigma = \sigma^2\mathbf{I}, \mathbf{I} \in \mathcal{R}^{(2\times 2)}$

$$p(x_1, x_2) = \frac{1}{2\pi\sigma^2}\exp\{-\frac{(x_1-\mu_1)^2 + (x_2-\mu_2)^2}{2\sigma^2}\}$$

- 将笛卡尔坐标 (x_1, x_2) 转化为极坐标 (r, θ)
  - $x_1 = r\cos\theta, x_2 = r\sin\theta$
  - $\mu_1 = r_0\cos\theta_0, \mu_2 = r_0\sin\theta_0$
  - 高斯分布的指数项为 (const 表示与 θ 无关的项)

$$\begin{aligned}
&-\frac{1}{2\sigma^2}\{(r\cos\theta - r_0\cos\theta_0)^2 + (r\sin\theta - r_0\sin\theta_0)^2\} \\
&\quad = -\frac{1}{2\sigma^2}\{1 + r_0^2 - 2r_0\cos\theta\cos\theta_0 - 2r_0\sin\theta\sin\theta_0\} \\
&\quad = \frac{r_0}{\cos(\theta - \theta_0)} + \text{const}
\end{aligned}$$

计算过程中使用的三角恒等式

$$\cos^2 A + \sin^2 A = 1 \cos A \cos B + \sin A \sin B = \cos(A - B)$$

在单位圆 $r = 1$ 上的概率分布 $p(\theta)$，即 Von Mises 分布

- 均值：θ_0
- 集中度 (concentration) 参数：m，类似于高斯分布的方差的倒数 (精度)。
  - 当 m 值足够大时，分布逼近高斯分布
- 零阶修正的第一类 Bessel 函数：$I_0(m) = \frac{1}{2\pi} \int_0^{2\pi} \exp\{m \cos\theta\} \mathrm{d}\theta$

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\{m \cos(\theta - \theta_0)\}$$

Von Mises 分布关于参数 θ_0 和 m 的最大似然估计

- 对数似然函数

$$\ln p(\mathcal{D}|\theta_0, m) = -N \ln(2\pi) - N \ln \mathrm{I}_0(m) + m \sum_{n=1}^{N} \cos(\theta_n - \theta_0)$$

- 令对数似然函数关于 θ_0 的导数为零，得：$\sum_{n=1}^{N} \sin(\theta_n - \theta_0) = 0$
  - 基于三角恒等式 $\sin(A - B) = \cos B \sin A - \cos A \sin B$，得

$$\theta_0^{ML} = \tan^{-1}\left\{ \frac{\sum_n \sin\theta_n}{\sum_n \cos\theta_n} \right\}$$

- 令对数似然函数关于 m 的导数为零，得：$A(m_{ML}) = \frac{1}{N} \sum_{n=1}^{N} \cos(\theta_n - \theta_0^{ML})$
  - $A(m) = I_1(m)/I_0(m)$
  - $I_1(m) = I_0'(m)$

$$A(m_{ML}) = (\frac{1}{N} \sum_{n=1}^{N} \cos\theta_n) \cos\theta_0^{ML}) + (\frac{1}{N} \sum_{n=1}^{N} \sin\theta_n) \sin\theta_0^{ML})$$

- 其他建立周期概率分布的方法
  - 直方图: 极坐标被划分成大小固定的箱子。

    * 优点是简洁并且灵活。
    * 局限 (Sec 2.5)
  - 类似于 Von Mises 分布: 考察欧几里得空间的高斯分布，在单位圆上做积分，使概率分布的形式相比直方图要复杂。
  - 持续地把宽度为 2π 的区间映射为周期变量 $(0, 2\pi)$
    * 优点：可以将实数轴上的任何合法的分布都可以转化成周期分布
    * 缺点：概率分布的形式相比于 Von Mises 分布更加复杂

2.3.9 混合高斯模型

混合模型 (mixture distribution) : 通过将基本的概率分布进行线性组合叠加形成概率模型。

一元混合高斯 (mixture of Gaussian) : K 个高斯概率密度的叠加形成混合高斯模型

$$p(\mathsf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathsf{x}|\mu_k, \Sigma_k)$$

- 每个高斯概率密度 $\mathcal{N}(\mathsf{x}|\mu_k, \Sigma_k)$ 是混合分布的一个成分
  - 参数 μ_k, Σ_k 是每个成分的均值和协方差
  - 参数 π_k 是混合系数，并且
    * $\sum_{k=1}^{K} \pi_k = 1$
    * $0 \leq \pi_k \leq 1$
- 从贝叶斯的角度分析，边缘概率密度：$p(\mathsf{x}) = \sum_{k=1}^{K} p(k)p(\mathsf{x}|k)$
  - 每个高斯概率密度的混合系数 $\pi_k = (k)$ 看成选择这个成分的先验概率
  - 定义后验概率 $p(k|\mathsf{x})$ 为「责任 (responsibilities)」(Ch 09)

$$\begin{aligned}
\gamma(\mathsf{x}) &\equiv p(k|\mathsf{x}) \\
&= \frac{p(k)p(\mathsf{x}|k)}{\sum_l p(l)p(\mathsf{x}|l)} \\
&= \frac{\pi_k \mathcal{N}(\mathsf{x}|\mu_k, \Sigma_k)}{\sum_l \pi_l \mathcal{N}(\mathsf{x}|\mu_l, \Sigma_l)}
\end{aligned}$$

多元混合高斯

- 前提条件
  - $\pi \equiv \{\pi_1, \cdots, \pi_K\}$

  - $\mu \equiv \{\mu_1, \cdots, \mu_K\}$
  - $\Sigma \equiv \{\Sigma_1, \cdots, \Sigma_K\}$
  - $\mathsf{X} = \{\mathsf{x}_1, \cdots, \mathsf{x}_N\}$

- 对数似然函数

$$\ln p(\mathsf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln\{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathsf{x}_n|\mu_k, \Sigma_k)\}$$

- 参数的最大似然解不再有一个封闭形式的解析解。求解方式有以下两种方法
  - 迭代数值优化方法
  - 期望最大化方法 (Ch 09)

2.4. 指数族分布: 连续分布

指数族 (exponential family) 分布

$$p(\mathsf{x}|\eta) = h(\mathsf{x})g(\eta)\exp\{\eta^T u(\mathsf{x})\}$$

- 变量 x：可能是标量，也可能是向量；可能是离散，也可能是连续。
- 参数为 η 是概率分布的自然参数 (natural parameters)
- $u(\mathsf{x})$ 是 x 的某个函数
- $g(\eta)$ 是系数，用来归一化概率分布

$$g(\eta) \int h(\mathsf{x}) \exp\{\eta^T u(\mathsf{x})\} \mathsf{dx} = 1 \tag{2.195}$$

指数族分布的实例：

- Bernoulli 分布
  - 参数：$\eta = \ln(\frac{\mu}{1-\mu})$
  - Logistic Sigmoid 函数：$\sigma(\eta) = \frac{1}{1+\exp(-\eta)} = \mu$
  - 转化成指数族分布形式
    - $u(x) = x$
    - $h(x) = 1$
    - $g(\eta) = \sigma(-\eta)$

$$\begin{aligned} p(x|\mu) &= \text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \\ &= \exp\{x \ln \mu + (1-x)\ln(1-\mu)\} \\ &= (1-\mu)\exp\left\{\ln(\frac{\mu}{1-\mu})x\right\} \\ &= \sigma(-\eta)\exp(\eta x) \end{aligned}$$

- 一元多项式分布
  - 参数：M 个参数 μ_k
    - $\eta = (\eta_1, \cdots, \eta_M)^T, \eta_k = \ln \mu_k$
    - $\sum_{k=1}^M \mu_k = 1, 0 \le \mu_k \le 1$
  - 转化成指数族分布形式
    - $u(\mathbf{x}) = \mathbf{x}$
    - $h(\mathbf{x}) = 1$
    - $g(\eta) = 1$

$$p(\mathbf{x}|\mu) = \prod_{k=1}^M \mu_k^{x_k} = \exp(\sum_{k=1}^M x_k \ln \mu_k)$$

- 一元多项式分布
  - 参数：$M-1$ 个参数 μ_k，第 M 个参数使用前 $M-1$ 个参数表示
    - $\eta_k = \ln(\frac{\mu_k}{1-\sum_j \mu_j})$
    - $\sum_{k=1}^M \mu_k = 1, 0 \le \mu_k \le 1, \sum_{k=1}^{M-1} \mu_k \le 1$
    - $\mu_k = \frac{\exp(\eta_k)}{1+\sum_j \exp(\eta_j)}$，Softmax 函数，也称为归一化指数
  - 转化成指数族分布形式
    - $u(\mathbf{x}) = \mathbf{x}$
    - $h(\mathbf{x}) = 1$
    - $g(\eta) = \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1}$

$$
\begin{aligned}
p(\mathsf{x}|\mu) &= \prod_{k=1}^{M} \mu_k^{x_k} = \exp(\sum_{k=1}^{M} x_k \ln \mu_k) \\
&= \exp\left\{ \sum_{k=1}^{M-1} x_k \ln \mu_k + (1 - \sum_{k=1}^{M-1} x_k) \ln(1 - \sum_{k=1}^{M-1} \mu_k) \right\} \\
&= \exp\left\{ \sum_{k=1}^{M-1} x_k \ln(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}) + \ln(1 - \sum_{k=1}^{M-1} \mu_k) \right\}
\end{aligned}
$$

- 一元高斯分布
  - 参数 : $\eta = (\mu/\sigma^2, -1/(2\sigma^2))^T$
  - 转化成指数族分布形式
    * $\mathsf{u}(x) = (x, 2x^2)^T$
    * $h(x) = (2\pi)^{-1/2}$
    * $g(\eta) = (-2\eta_2)^{-1/2} \exp(\eta_1^2/(4\eta_2))$

$$
\begin{aligned}
p(x|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(x - \mu)^2\} \\
&= \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{1}{2\sigma^2} x^2 + \frac{\mu}{2\sigma^2} - \frac{1}{2\sigma^2} \mu^2 \right\}
\end{aligned}
$$

2.4.1 最大似然与 充分统计量

(最大似然与充分统计量两个概念都很重要，需要真正理解，因为后面会大量出现)

使用最大似然估计计算一般形式的指数族分布的参数向量 η

- 对 (Eq 2.195) 两边关于 η 求梯度

$$
\nabla g(\eta) \int h(\mathsf{x}) \exp\{\eta^T \mathsf{u}(\mathsf{x})\} d\mathsf{x} + g(\eta) \int h(\mathsf{x}) \exp\{\eta^T \mathsf{u}(\mathsf{x})\} \mathsf{u}(\mathsf{x}) d\mathsf{x} = 0
$$

$$
-\frac{1}{g(\eta)} \nabla g(\eta) = g(\eta) \int h(\mathsf{x}) \exp\{\eta^T \mathsf{u}(\mathsf{x})\} \mathsf{u}(\mathsf{x}) d\mathsf{x} = \mathbb{E}[\mathsf{u}(\mathsf{x})]
$$

- 结论：如果可以对一个指数族分布的概率分布进行归一化，则 u (x) 的矩可以通过求微分的方式获得

$$\mathbb{E}[\mathsf{u}(\mathsf{x})] = -\frac{1}{g(\eta)}\nabla g(\eta) \tag{2.226}$$

实例：一组独立同分布的数据 $\mathsf{X} = \{\mathsf{x}_1, \cdots, \mathsf{x}_N\}$

- 似然函数

$$p(\mathsf{X}|\eta) = \left(\prod_{n=1}^{N} h(\mathsf{x}_n)\right) g(\eta) \exp\left\{\eta^T \sum_{n=1}^{N} \mathsf{u}(\mathsf{x}_n)\right\}$$

- 令对数似然函数 $\ln p(\mathsf{X}|\eta)$ 关于 η 的导数等于零

$$-\nabla \ln g(\eta_{ML}) = \frac{1}{N} \sum_{n=1}^{N} \mathsf{u}(\mathsf{x}_n) \tag{2.228}$$

- 因为最大似然估计只需要通过 $\sum_n \mathsf{u}(\mathsf{x}_n)$ 对数据产生依赖，因此 $\sum_n \mathsf{u}(\mathsf{x}_n)$ 称为概率分布的充分统计量。
  - Bernoulli 分布的 $u(x) = x$，因此 $\sum_n x$ 是充分统计量
  - Gauss 分布的 $\mathsf{u}(x) = (x, x^2)^T$，因此 $\sum_n x$ 和 $\sum_n x^2$ 是充分统计量
- 当 $N \to \infty$ 时，(Eq 2.228) 右侧变为 $\mathbb{E}[\mathsf{u}(x)]$，对比 (Eq 2.226) 可得 $\eta_{ML} \to \eta$

2.4.2 共轭先验

对于给定的概率分布，寻找一个先验使其与似然函数共轭，从而后验分布的函数形式与先验分布相同。

- 指数族分布中的任何分布，都存在一个共轭先验。
  - $f(\chi, \nu)$ 是归一化系数
  - η 是似然函数中的参数

$$p(\eta|\chi, \nu) = f(\chi, \nu) g(\eta)^{\nu} \exp\{\nu \eta^T \chi\}$$

- 后验概率
  - 从贝叶斯的角度，参数 ν 可以看成先验分布中假想观测的有效观测数。
  - 在给定 χ 的情况下，每个假想观测都对充分统计量 $\mathsf{u}(\mathsf{x})$ 的值有贡献

$$p(\eta|\mathsf{X},\chi,\nu) \propto g(\eta)^{\nu+N} \exp\left\{\eta^T\left(\sum_{n=1}^{N}\mathsf{u}(\mathsf{x}_n)+\nu\chi\right)\right\}$$

2.4.3 无信息先验 (non-informative prior)

当没有先验知识时，选择「先信息先验」能够对后验分布产生尽可能小的影响。

当先验分布 $p(\lambda) = \text{const}$ 时存在的问题

- 如果 λ 的取值范围是无界的，那么先验分布无法被正确地归一化。这样的先验分布被称为反常的 (improper)
  - 如果对应的后验分布是正常的，即后验分布可以被正确地归一化，那么反常的先验分布可以作为选择。
- 概率密度分布中变量的非线性变换，使得最初的先验分布经过变换后不再是常数。
  - 假如：函数 $h(\lambda) = \text{const}$，进行变量替换 $\lambda = \eta^2$，得 $\hat{h}(\eta) = h(\lambda) = h(\eta^2) = \text{const}$
  - 假如：概率密度 $p_\lambda(\lambda) = \text{const}$，进行变量替换 $\lambda = \eta^2$，得 η 的概率密度 $p_\eta(\eta) = p_\lambda(\lambda)|=p_\lambda(\eta^2)2\eta \propto \eta$ 不再是常数
    * (Eq 1.27) $p_y(y) = p_x(x)|\frac{dy}{dx}| = p_x(g(y))|g'(y)|$
  - 在最大似然估计中不存在问题，因为 $p(x|\lambda)$ 是 λ 的简单函数，所以可以使用各种对参数操作的方法
  - 在最大后验估计中，如果需要先验分布为常数时，就需要注意参数使用一个合适的表达形式

先验分布的两个例子

- 具有平移不变性 (translation invariance) 的概率分布
  - $p(x|\mu) = f(x-\mu)$
    * μ ：位置参数 (location parameter)
  - $p(\hat{x}|\hat{\mu}) = f(\hat{x}-\hat{\mu})$
    * $\hat{x} = x + c$
    * $\hat{\mu} = \mu + c$
  - 两个概率分布的形式相同
  - 选择先验分布时也需要满足平移不变性
    * $\int_A^B p(\mu)\mathrm{d}\mu = \int_{A-c}^{B-c} p(\mu)\mathrm{d}\mu = \int_A^B p(\mu-c)\mathrm{d}\mu$
    * $p(\mu-c) = p(\mu) = \text{const}$

  - 实例：高斯分布的均值的共轭先验
    * $p(\mu|\mu_0, \sigma_0^2) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$
    * 当 $\sigma_0^2 \to \infty$ 时，得到一个无信息先验

- 具有缩放不变性 (scale invariance) 的概率分布
  - $p(x|\sigma) = \frac{1}{\sigma} f(\frac{x}{\sigma})$
    * $\sigma > 0$
    * σ：缩放参数 (scale parameter)
  - $p(\hat{x}|\hat{\sigma}) = \frac{1}{\hat{\sigma}} f(\frac{\hat{x}}{\hat{\sigma}})$
    * $\hat{x} = cx$
    * $\hat{\sigma} = c\sigma$
  - 两个概率分的形式相同
  - 选择先验分布时也需要满足缩放不变性
    * $p(\sigma) = p(\frac{1}{c}\sigma)\frac{1}{c} \propto \frac{1}{\sigma}$
    * 这是一个反常先验分布：$\int_0^\infty p(\sigma) > \infty$
    * 一个原始区间 $A \le \sigma \le B$，一个缩放区间 $\frac{A}{c} \le \sigma \le \frac{B}{c}$
      · 满足条件 $\int_A^B p(\sigma)\mathrm{d}\sigma = \int_{\frac{A}{c}}^{\frac{B}{c}} p(\sigma)\mathrm{d}\sigma = \int_A^B p(\frac{1}{c}\sigma)\frac{1}{c}\mathrm{d}\sigma$
    * 常用的缩放先验分布：$p(\ln\sigma) = \mathrm{consnt}$
    * 实例：高斯分布的方差的共轭先验
      .

2.5. 非参数化密度估计

(不是本书重点，作者描述较少，可参考 [9] Ch 03 和 [10] Ch 04)

- 概率密度建模的参数化方法
  - 优点
    * 利用数据估计有参数的概率分布的形式
    * 一旦确定数据的概率分布形式，概率密度只需要少量的参数控制
  - 缺点
    * 如果假设的概率模型错误，那么估计的结果也会错误
    * 如果假设的概率模型是多峰的，那么也不容易正确估计

- 概率密度建模的非参数化方法

[9] Andrew R. Webb. 统计模式识别。电子工业出版社。2004.
[10] Duda R O,Peter E Hart,etc. 李宏东等译。模式分类。机械工业出版社。2003.

  - 优点：不需要确定数据的概率分布形式

- 密度估计的基本思路：直方图方法
  - 优点
    * 直方图计算完成后就可以丢弃数据，适合大数据量处理
    * 可以应用到数据顺序到达的情况
    * 适合将一维或者二维数据的可视化应用
  - 缺点
    * 估计的概率密度具有不连续性
    * 可能造成维数放大。例如：D 维蝗每一维变量都划分到 M 个箱子中，那么箱的总数为 M^D
  - 两个常用的概率密度建模的非参数化方法都能更好地处理维度放大问题
    * 核密度估计
    * K 近邻密度估计

- 非参数概率密度估计的特点
  - 需要确定距离度量，方便计算某个领域内的数据点个数；
  - 为了更好地平衡概率密度的细节，需要注意领域的大小，类似于模型复杂度的确定。

- 密度估计的通用形式
  - V 是区域 R 的体积
  - K 是区域 R 内数据点的个数

- 密度估计的核密度方法
  - 固定 V，确定 K
  - 选择一个核函数（kernel function），即 Parzen 窗
    * 为了得到平滑的模型，需要选择平滑的核函数
  - 不需要进行训练阶段的计算
  - 估计概率密度的计算代价随着数据集的规模线性增长。

- 密度估计的近邻方法
  - 固定 K，确定 V
  - 解决核方法中核固定的问题
  - K 近邻法得到的模型不是真实的概率密度模型，在整个空间的积分是发散的。

- 常用密度估计方法的问题：
  - 需要存储整个训练数据

  - 基于树的搜索结构，解决这个问题
    * 不必遍历整个数据集，就可以找到合适的近邻
    * 需要增加计算量

2.5.1 核密度估计

密度估计的通用形式

- 前提条件
  - D 空间是欧氏空间
  - N 次观测服从 D 维空间的某个未知的概率密度分布 $p(\mathrm{x})$
  - 包含 x 的某个小区域 $\mathcal{R}$ 的概率质量 $P = \int_{\mathcal{R}} p(\mathrm{x})\mathrm{dx}$
  - 位于区域 $\mathcal{R}$ 内部的数据点的总数 K 服从二项分布

$$\mathrm{Bin}(K|N,P) = \frac{N!}{K!\,(N-K)!}P^K(1-P)^{N-K}$$

- 计算过程
  - 由 (Eq 2.11) $\mathbb{E}[m] = N\mu$ 得 $\mathbb{E}[K/N] = P$
  - 由 (Eq 2.12) $\mathrm{var}[m] = N\mu(1-\mu)$ 得 $\mathrm{var}[K/N] = P(1-P)/N$
  - 当 $N \to \infty$ 时，$K \simeq NP$，分布在均值处产生尖峰
  - 假定 $\mathcal{R}$ 足够小，使得区域内的 $p(\mathrm{x}) = \mathrm{const}$，得
    * $P \simeq p(\mathrm{x})V$
    * V 是区域 $\mathcal{R}$ 的体积
  - 因此，概率密度的估计公式，得

$$p(\mathrm{x}) = \frac{K}{NV} \tag{2.246}$$

- 注意：(Eq 2.246) 的成立依赖下面两个矛盾
  - 区域 $\mathcal{R}$ 要足够小，即 V 足够小，使得区域内的概率密度近似为常数
  - 区域 $\mathcal{R}$ 要足够大，即 K 足够大，保证区域内的数据点能使二项分布产生尖峰
- 利用 (Eq 2.245) 的两种方法
  - K 近邻方法：固定 K，计算 V
  - 核密度估计：固定 V，计算 K
  - 在极限 $N \to \infty$ 时，V 随着 N 收缩，K 随着 N 增大，则两种方法都能够收敛到真实的概率密度

密度估计的核密度方法

- 前提条件
  - 区域 $\mathcal{R}$ 定义为以 x 为中心的小超立方体
  - D 维边长为 h 的立方体的体积公式 $V = h^D$
  - 区域内的数据点数量 $K = \sum_{n=1}^{N} k(\frac{\mathsf{x}-\mathsf{x}_n}{h})$
    - $k(\cdot)$ 是核函数，在估计问题中也被称为 Parzen 窗
- 选择核函数 $k(\cdot)$，表示一个以原点为中心的单位立方体，存在非连续性问题

$$k(\mathsf{u}) = \begin{cases} 1 & |u_i| \leq 1/2, i = 1, \cdots, D, \\ 0 & \text{others} \end{cases}$$

- 核函数需要满足的条件
  - $h(\mathsf{u}) \geq 0$
  - $\int k(\mathsf{u})\mathrm{du} = 1$
- 密度估计：点 x 处的概率密度估计

$$p(\mathsf{x}) = \frac{K}{NV} = \sum_{n=1}^{N} k(\frac{\mathsf{x}-\mathsf{x}_n}{h})\frac{1}{N}\frac{1}{h^D}$$

- 选择高斯核函数 (平滑的核函数)，得到概率密度模型 (平滑的模型)
  - h 表示高斯分布的标准差
    - h 过小，会造成模型对噪声过于敏感
    - h 过大，会造成模型过度平滑
    - 对 h 的一个模型复杂度问题

$$p(\mathsf{x}) = \frac{K}{NV} = \sum_{n=1}^{N} \exp\{-\frac{\|\mathsf{x}-\mathsf{x}_n\|^2}{2h^2}\}\frac{1}{N}\frac{1}{(2\pi h^2)^{D/2}}$$

2.5.2 密度估计的 K 近邻方法

- 前提条件
  - 区域 $\mathcal{R}$ 定义为以 x 为中心的小球体
    - 球体的半径可以变化，直到包含 K 个数据点为止

        * K 的值控制了概率密度的光滑程度
- 注意：K 近邻方法得到的不是真实的概率密度模型，因为模型在整个空间的积分是发散的。

K 近邻方法在分类问题中的应用

- 前提条件
  - N 个数据点的数据集中 N_k 个数据点属于类别 $\mathcal{C}_k$
    * $N = \sum_k N_k$
- 分类任务
  - 需要对新的数据点 x 进行分类
  - 以 x 为中心建立小球体
    * 球体体积 V
    * 球体包含 K 个数据点
    * 每个类别 $\mathcal{C}_k$ 的数据点为 K_k
    * $K = \sum_k K_k$
- 公式推导
  - 每个类别关联的概率密度估计：$p(\mathbf{x}|\mathcal{C}_k) = K_k/(N_k V)$
  - 数据集本身的概率密度估计：$p(\mathbf{x}) = K/(NV)$
  - 类别的先验概率：$p(\mathcal{C}_k) = N_k/N$
  - 类别的后验概率：$p(\mathcal{C}_k|\mathbf{x}) = p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)/p(\mathbf{x}) = K_k/K$
    * 从后验概率可以看出将新的数据点分配给区域中数据点最多的类别是合理的
    * $K = 1$ 的特例称为最近邻规则
    * $K = 1$ 分类器在 $N \to \infty$ 时，错误率不会超过最优分类器的最小错误率的二倍

02. 小结

这章看起来是基础知识的介绍，实际上是对后面知识的梳理。如果这章看完有太多不理解的内容，再次建议先补充概率与统计的基础，否则就无法通过贝叶斯角度来理解机器学习。

Ch 03. 回归的线性模型

提纲

重点

- 线性基函数模型（Sec 3.1）

- 贝叶斯线性回归 (Sec 3.3)
- 二次正则化项 (Eq 3.27)

难点

- 贝叶斯模型比较 (Sec 3.4)
- 证据近似 (Sec 3.5)

学习要点

- 回归问题：在给定输入变量 x $\in \mathcal{R}^D$ 的情况下，预测一个或者多个连续目标变量 t 的值
- 线性回归模型：具有可调节的参数，具有线性函数的性质
  - 输入变量的线性函数：最简单的形式
  - 参数的线性函数：将一组输入变量的非线性函数 (基函数) 进行线性组合
    * 依然具有简单的分析性质；
    * 同时关于输入变量是非线性的。
    * 非线性变换不会消除数据中的重叠，甚至还可能增加重叠的程度或者在原始观测空间中不存在重叠的地方产生出新的重叠。
    * 恰当地选择非线性变换可以让后验概率的建模过程更加简单。

3.1. 线性基函数模型

线性回归 (Linear Regression) ：输入变量的线性组合。

- 回归问题的最简单模型
  - $y(\mathbf{x},\mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D$
  - $\mathbf{x} = (x_1, \cdots, x_D)^T$
  - $y(\mathbf{x},\mathbf{w})$ 是参数 $w_0, \cdots, w_D$ 的线性函数

线性基函数：输入变量的固定的非线性函数的线性组合。

- 原始模型：$y(\mathbf{x},\mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$
  - 偏置参数 (bias parameter) w_0：使得数据中可以存在任意固定的偏置。
- 简化模型：$y(\mathbf{x},\mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$
  - 定义一个虚「基函数」$\phi_0(\mathbf{x}) = 1$，简化数学模型。

基函数 (Basis Function) $\phi_j(\mathbf{x})$ 的选择

- 线性无关函数集：$k_1 = k_2 = k_3 = 0, k_1 f_1(\cdot) + k_2 f_2(\cdot) + k_3 f_3(\cdot) = 0$
  - 多项式基函数：$\phi_j(x) = x^j$，x 的幂指数形式
    * 局限性：输入变量的全局函数，因此对于输入空间中一个区域的改变将会影响所有其他的区域。
    * 解决办法：样条函数 (spline function)。把输入空间切分成若干个区域，然后对于每个区域用不同的多项式函数拟合。
  - 高斯基函数：$\phi_j(x) = \exp[-\frac{(x-\mu_j)^2}{2s^2}]$
  - Sigmoid 基函数：$\phi_j(x) = \sigma(\frac{x-\mu_j}{s})$
    * Logistic Sigmoid 函数：$\sigma(a) = \frac{1}{1+\exp(-a)}$
- 规范正交函数集：$f_i(\cdot)f_j(\cdot) = \begin{cases} 1 & i = j \\ 0, & i \neq j \end{cases}$
  - Fourier 基函数：使用正弦函数展开。每个基函数代表一个频率，在空间中有无限的延伸。是规范正交函数集
  - 小波 (wavelet) 基函数

3.1.1. 最大似然与最小平方

平方和误差函数等价于高斯噪声模型的下的最大似然解。

线性：有噪声：函数建模

- $t = y(\mathbf{x},\mathbf{w}) + \epsilon$
  - $y \in \mathcal{R}, \mathbf{x} \in \mathcal{R}^D, \mathbf{w} \in \mathcal{R}^{D+1}, \epsilon \sim \mathcal{N}(0, \beta^{-1}), \sigma \in \mathcal{R}$
- (Eq_3.8) : $p(t|\mathbf{x},\mathbf{w},\beta) = \mathcal{N}(t|y(\mathbf{x},\mathbf{w}),\beta^{-1})$
- 条件均值：$\mathbb{E}[t|\mathbf{x}] = \int tp(t|\mathbf{x})\mathrm{d}t = y(\mathbf{x},\mathbf{w})$
  - 前提条件
    * 噪声服从高斯分布
    * 目标变量的条件分布 $p(t|\mathbf{x})$ 也服从高斯分布
  - 补充：这里处理的是单峰的高斯分布，(Sec 14.5.1.) 扩展到多峰的高斯分布

线性：有噪声：概率建模 似然函数：

$$p(\mathsf{t}|\mathsf{X},\mathsf{w},\beta)=\prod_{n=1}^{N}\mathcal{N}(t_n|\mathsf{w}^T\phi(\mathsf{x}_n),\beta^{-1}) \tag{3.10}$$

因为输入变量 x 不是求解目标，并且一直存在于条件变量中，因此简化表达式为

$$p(\mathsf{t}|\mathsf{X},\mathsf{w},\beta)=p(\mathsf{t}|\mathsf{w},\beta)$$

对数似然函数

$$\begin{aligned}\ln p(\mathsf{t}|\mathsf{w},\beta)&=\sum_{n=1}^{N}\ln\mathcal{N}(t_n|\mathsf{w}^T\phi(\mathsf{x}_n),\beta^{-1})\\&=\frac{N}{2}\ln\beta-\frac{N}{2}\ln(2\pi)-\beta E_D(\mathsf{w})\end{aligned} \tag{3.11}$$

平方和误差函数

$$E_D(\mathsf{w})=\frac{1}{2}\sum_{n=1}^{N}\{t_n-\mathsf{w}^T\phi(\mathsf{x}_n)\}^2 \tag{3.12}$$

对数似然函数关于参数 w 求导等于零

$$\nabla_{\mathsf{w}}\ln p(\mathsf{t}|\mathsf{w},\beta)=\beta\sum_{n=1}^{N}\{t_n-\mathsf{w}^T\phi(\mathsf{x}_n)\}\phi(\mathsf{x}_n)^T=0$$

得到极值 w_{ML} 称之为「最小二乘问题」的「规范方程」(normal equation)

- $\mathsf{w}_{ML}=(\Phi^T\Phi)^{-1}\Phi^T\mathsf{t}=\Phi^\dagger\mathsf{t}$
  - 设计矩阵 (Eq 3.16) ：$\Phi=\begin{bmatrix}\phi_0(\mathsf{x}_1)&\phi_1(\mathsf{x}_1)&\cdots&\phi_M(\mathsf{x}_1)\\\phi_0(\mathsf{x}_2)&\phi_1(\mathsf{x}_2)&\cdots&\phi_M(\mathsf{x}_2)\\\vdots&\vdots&\ddots&\vdots\\\phi_0(\mathsf{x}_N)&\phi_1(\mathsf{x}_N)&\cdots&\phi_M(\mathsf{x}_N)\end{bmatrix}$
    - $\Phi\in\mathcal{R}^{(N\times M)}$
    - $\Phi_{nj}=\phi_j(\mathsf{x}_n)$
  - Moore-Penrose 伪逆矩阵 (pseudo-inverse matrix) ：$\Phi^\dagger\equiv(\Phi^T\Phi)^{-1}\Phi^T$
  - 如果 Φ 是方阵并且可逆，则基于 $(AB)^{-1}=B^{-1}A^{-1}$ 性质，得 $\Phi^\dagger\equiv\Phi^{-1}$

单独求解偏置参数 w_0

- 平方和误差函数：$E_D(\mathsf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w_0 - \sum_{j=1}^{M-1} w_j\phi_j(\mathsf{x})\}^2$
- 关于 w_0 求导，得 $w_0 = \bar{t} - \sum_{j=1}^{M-1} w_j\bar{\phi}_j(\mathsf{x})$
  - $\bar{t} = \frac{1}{N}\sum_{n=1}^{N} t_n$
  - $\bar{\phi}_j = \frac{1}{N}\sum_{n=1}^{N}\phi_j(\mathsf{x}_n)$
- w_0 补偿了训练集上的目标值的平均值与基函数的值的平均值的加权求和之间的差

求解噪声精度参数 β

- 噪声精度的倒数是目标值在回归函数周围的残留方差（residual variance）的均值

$$\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{t_n - \mathsf{w}_{ML}^T\phi(\mathsf{x}_n)\}^2 \tag{3.21}$$

3.1.2. 最小平方的几何描述

平方误差函数是 y 和 t 之间的欧氏距离的平方。

- N 个目标数据 t_n 张成一个 N 维空间
  - $\mathsf{t} = (t_1, \cdots, t_N)$ 是一个 N 维向量
  - $\mathsf{y} = (y(x_1, \mathsf{w}), \cdots, y(x_N, \mathsf{w})$ 是一个 N 维向量
  - $\varphi_j = (\phi_j(x_1), \cdots, \phi_j(x_N))$ 是一个 N 维向量
- 如果 $N > M$，则 M 个向量 φ_j 张成 M 维的子空间 S
  - y 是 t 在子空间 S 上的正交投影。
  - w 的最小平方解就是 y 与 t 之间的距离
- 当 $\Phi^T\Phi$ 接近奇异矩阵时，直接求解规范方程存在困难，可以使用 SVD（奇异值分解) 来解决。

3.1.3. 顺序学习、在线学习

随机梯度下降（stochastic gradient descent），也叫顺序梯度下降（sequential gradient descent）

- 数据点和误差函数：$E = \sum_n E_n$

  - 更新公式：$\mathsf{w}^{(\tau+1)} = \mathsf{w}^{(\tau)} - \eta\nabla E_n$
- 平方误差函数：$E_D(\mathsf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathsf{w}^T\phi(\mathsf{x}_n)\}^2$
  - 更新公式：$\mathsf{w}^{(\tau+1)} = \mathsf{w}^{(\tau)} + \eta(t_n - (\mathsf{w}^{(\tau)})^T\phi_n)\phi_n$
  - 这个误差函数的随机梯度下降算法，也叫最小均方误差算法（Least Mean Squares, LMS）。

3.1.4. 正则化最小平方

误差函数：$E_D(\mathsf{w}) + \lambda E_W(\mathsf{w})$

- 基于数据的误差：$E_D(\mathsf{w})$
- 正则化项：$E_W(\mathsf{w})$
  - λ 是正则化系数

具体案例

- 平方和误差函数
  - （Eq 3.26）：$E_D(\mathsf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathsf{w}^T\phi(\mathsf{x}_n)\}^2$
- 二次正则化项
  - $E_W(\mathsf{w}) = \frac{1}{2}\mathsf{w}^T\mathsf{w}$

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathsf{w}^T\phi(\mathsf{x}_n)\}^2 + \frac{\lambda}{2}\mathsf{w}^T\mathsf{w} \tag{3.27}$$

- 正则化项的选择方法，在机器学习的文献中被称为权值衰减（weight decay）
  - 在顺序学习算法中，「权值」向零的方向衰减
- 正则化项的选择方法，在统计学习中称为参数收缩（parameter shrinkage）
  - 在顺序学习算法中，「参数的值」向零的方向收缩
- 其他类型的正则化项
  - q=1 的情形被称为套索（lasso）
    * 性质为：如果 λ 充分大，那么某些系数会变为零，从而产生一个稀疏（sparse）模型。
- 正则化方法通过限制模型的复杂度，使得复杂的模型在有限大小的数据集上进行训练，而不会产生严重的过拟合问题。

- 使得确定最优的模型复杂度的问题从确定合适的基函数数量的问题转移到了确定正则化系数 λ 的合适值的问题。

3.1.5. 多个输出

多个输出：预测多于 1 个目标变量的数据，可以采用下面两种方法：

第一种方法

- 对于 t 的每个分量，引入一个不同的基函数集合，从而转化为多个独立的回归问题

第二种方法 (更常用)

- 对目标向量的所有分量使用一组相同的基函数来建模
  - $y(\mathbf{x},\mathbf{w}) = W^T\phi(\mathbf{x})$
    - $\phi(\mathbf{x}) = \phi_j(\mathbf{x}), \phi_0(\mathbf{x}) = 1$
    - $y \in \mathcal{R}^K, W \in \mathcal{R}^{M\times K}, \phi(\mathbf{x}) \in \mathcal{R}^M$
- 假设目标向量的条件概率分布是一个各向同性的高斯分布

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

- 对数似然函数

$$\begin{aligned}\ln p(\mathbf{T}|\mathbf{X},\mathbf{W},\beta) &= \sum_{n=1}^{N}\ln\mathcal{N}(\mathbf{t}_n|\mathbf{W}^T\phi(\mathbf{x}_n),\beta^{-1}\mathbf{I}) \\ &= \frac{NK}{2}\ln(\frac{\beta}{2\pi}) - \frac{\beta}{2}\sum_{n=1}^{N}\|\mathbf{t}_n - \mathbf{W}^T\phi(\mathbf{x}_n)\|^2\end{aligned}$$

- 求解对数似然函数的最大似然估计
  - $\mathbf{W}_{ML} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{T}$
  - $\mathbf{w}_k = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{t}_k = \Phi^\dagger\mathbf{t}_k$
    - $\mathbf{t}_k = (t_{1k},\cdots,t_{Nk})^T$
    - 保证不同目标变量的回归问题被分解成多个单目标变量的回归问题。

3.2.「偏置——方差」分解

频率学家看待「模型复杂度问题」的角度是 “偏置——方差” 折中 (bias-variance trade-off)

回归问题 (Ref : Sec 1.5.5)

- 最优预测：$h(\mathsf{x}) = \mathbb{E}_t[t|\mathsf{x}] = \int tp(t|\mathsf{x})\mathrm{d}t$
- 平方损失函数的期望：$\mathbb{E}[L] = \int\{y(\mathsf{x}) - h(\mathsf{x})\}^2 p(\mathsf{x})\mathrm{d}\mathsf{x} + \int\int\{h(\mathsf{x}) - t\}^2 p(\mathsf{x}, t)\mathrm{d}\mathsf{x}\,\mathrm{d}t$
  - $\int\{y(\mathsf{x}) - h(\mathsf{x})\}^2 p(\mathsf{x})\mathrm{d}\mathsf{x}$ 与 $y(\mathsf{x})$ 的选择有关，不同的选择会得到不同的回归函数的解 $h(\mathsf{x})$，从而结果会无限逼近于零
  - $\int\int\{h(\mathsf{x}) - t\}^2 p(\mathsf{x}, t)\mathrm{d}\mathsf{x}\,\mathrm{d}t$ 与 $y(\mathsf{x})$ 的选择无关，是受数据的噪声影响，表示期望损失能够达到的最小值。

估计的不确定性 (频率学派)

前提条件

- 多个数据集 $(\mathcal{D}_1, \cdots, \mathcal{D}_C)$
- 每个数据集 $\mathcal{D}_c$ 的大小都是 N
- 每个数据集 $\mathcal{D}_c$ 都独立地从分布 $p(t, \mathsf{x})$ 中抽取的
- 考虑特定数据集 $\mathcal{D}_c$，得到预测函数 $y_c(\mathsf{x}; \mathcal{D}_c)$

公式推导

- 单个数据集 $\mathcal{D}_c$，单个数据点 x

$$\{y(\mathsf{x}; \mathcal{D}_c) - h(\mathsf{x})\}^2$$

$$\begin{aligned} y(\mathsf{x}) - h(\mathsf{x}) &= \{y(\mathsf{x}; \mathcal{D}_c) - \mathbb{E}_{\mathcal{D}}[y(\mathsf{x}; \mathcal{D}_c)] + \mathbb{E}_{\mathcal{D}}[y(\mathsf{x}; \mathcal{D}_c)] - h(\mathsf{x})\}^2 \\ &= \{y(\mathsf{x}; \mathcal{D}_c) - \mathbb{E}_{\mathcal{D}}[y(\mathsf{x}; \mathcal{D}_c)]\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathsf{x}; \mathcal{D}_c)] - h(\mathsf{x})\}^2 \\ &\quad + 2\{y(\mathsf{x}; \mathcal{D}_c) - \mathbb{E}_{\mathcal{D}}[y(\mathsf{x}; \mathcal{D}_c)]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathsf{x}; \mathcal{D}_c)] - h(\mathsf{x})\} \end{aligned}$$

- 多个数据集，单个数据点 x，求期望

$$\begin{aligned} \mathbb{E}_{\mathcal{D}}[\{y(\mathsf{x}; \mathcal{D}) - h(\mathsf{x})\}^2] &= \int\{y(\mathsf{x}; \mathcal{D}) - h(\mathsf{x})\}^2 p(\mathsf{x})\mathrm{d}\mathsf{x} \\ &= \{\mathbb{E}_{\mathcal{D}}[y(\mathsf{x}; \mathcal{D}) - h(\mathsf{x})]\}^2 + \mathbb{E}_{\mathcal{D}}[\{y(\mathsf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathsf{x}; \mathcal{D})]\}^2] \\ &\quad + \mathbb{E}_{\mathcal{D}}[2\{y(\mathsf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathsf{x}; \mathcal{D})]\}\{\mathbb{E}_{\mathcal{D}}[y(\mathsf{x}; \mathcal{D})] - h(\mathsf{x})\}] \\ &= (\text{偏置})^2 + \text{方差} + 0 \end{aligned}$$

- 多个数据集，多个数据点 x，求积分
  - $(\text{偏置})^2 = \int\{\mathbb{E}_{\mathcal{D}}[y(\mathsf{x}; \mathcal{D}) - h(\mathsf{x})]\}^2 p(\mathsf{x})\mathrm{d}\mathsf{x}$

    * $(\text{偏置})^2$：表示所有数据集的平均预测与预期的回归函数之间的差异；
  - 方差 $= \int \mathbb{E}_{\mathcal{D}}[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2]p(\mathbf{x})\mathrm{d}\mathbf{x}$
    * 方差：度量了对于单独的数据集，模型所给出的解在平均值附近变化的情况
  - 噪声 $= \int\int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)\mathrm{d}\mathbf{x}\,\mathrm{d}t$
    * 噪声：数据上叠加的常数噪声项。
  - 平方损失函数展开后得：期望损失 $= (\text{偏置})^2 +$ 方差 $+$ 噪声

- 目标：最小化期望损失，在偏置与方差之间取得最优的平衡的模型
  - 灵活的模型，偏置较小，方差较大。
  - 固定的模型，偏置较大，方差较小。

- 例子：(Fig 3.5) ,100 个数据集，25 阶模型
  - $\bar{y}(x) = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$
  - $(\text{偏置})^2 = \frac{1}{N}\sum_{n=1}^{N}\{\bar{y}(x_n) - h(x_n)\}^2$
  - 方差 $= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\{y^{(l)}(x_n) - \bar{y}(x_n)\}^2$

总结

- 「偏置——方差」分解依赖于对所有的数据集求平均
  - 因为无法有效应用所有的数据，因此产生了贝叶斯线性回归

3.3. 贝叶斯线性回归

在最大似然估计中需要确定模型的复杂度以避免过拟合问题，解决模型复杂度需要使用分割数据集求平均的方式又存在无法有效利用数据问题，并且增加了计算量。

在贝叶斯估计中，既能够避免过拟合问题，还可以基于训练数据确定模型复杂度。

3.3.1. 参数分布

参数分布：引入模型参数 w 的先验概率分布

前提条件

- 似然函数：(噪声精度参数 β 为已知常数)
  - (Eq 3.10) :$p(\mathbf{t}|\mathbf{X},\mathbf{w},\beta) = \prod_{n=1}^{N}\mathcal{N}(\mathbf{t}_n|\mathbf{x}^T\phi(\mathbf{x}_n),\beta^{-1})$
- 对应的共轭分布是高斯分布

  - $p(\mathsf{w}) = \mathcal{N}(\mathsf{w}|\mathsf{m}_0, \mathsf{S}_0)$
    * m_0 为均值
    * S_0 为协方差

公式推导 (Eq 2.116 的推导)

- 后验概率 : $p(\mathsf{w}|\mathsf{t},\mathsf{X}, \beta)p(\mathsf{t}) = p(\mathsf{t}|\mathsf{X},\mathsf{w}, \beta)p(\mathsf{w})$
  - (Eq 3.49) : $p(\mathsf{w}|\mathsf{t}) = \mathcal{N}(\mathsf{w}|\mathsf{m}_N, \mathsf{S}_N)$
    * (Eq 3.50) : $\mathsf{m}_N = \mathsf{S}_N(\mathsf{S}_0^{-1}\mathsf{m}_0 + \beta\Phi^T\mathsf{t})$
    * (Eq 3.51) : $\mathsf{S}_N^{-1} = \mathsf{S}_0^{-1} + \beta\Phi^T\Phi$
    * Φ : 设计矩阵 (Eq 3.16)

具体案例

- 特定形式的先验
  - (Eq 3.52) : $p(\mathsf{w}|\alpha) = \mathcal{N}(\mathsf{w}|0, \alpha^{-1}\mathsf{I})$
    * 参数 w 受精度参数 α 控制
- 后验概率
  - $p(\mathsf{w}|\mathsf{t}) = \mathcal{N}(\mathsf{w}|\mathsf{m}_N, \mathsf{S}_N)$
    * $\mathsf{m}_N = \beta\mathsf{S}_N\Phi^T\mathsf{t}$
    * $\mathsf{S}_N^{-1} = \alpha\mathsf{I} + \beta\Phi^T\Phi$
    * Φ : 设计矩阵 (Eq 3.16)
- 后验概率的对数
  - w 的极值 : 最大化 (后验分布) 等价于最小化 (平方和误差函数 + 二次正则化项)
  - (Fig 3.7) 给出了贝叶斯学习的效果，以及基于贝叶斯的顺序学习的本质

$$\begin{aligned}\ln p(\mathsf{w}|\mathsf{t}) &= -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathsf{w}^T\phi(\mathsf{x}_n)\}^2 - \frac{\alpha}{2}\mathsf{w}^T\mathsf{w} + \text{const} \\ &= -\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathsf{w}^T\phi(\mathsf{x}_n)\}^2 - \frac{1}{2}\frac{\alpha}{\beta}\mathsf{w}^T\mathsf{w} + \text{const}\end{aligned}$$

- 与最大似然估计对比
  - $\lambda = \frac{\alpha}{\beta}$

$$E_D(\mathsf{w}) + \lambda E_W(\mathsf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathsf{w}^T\phi(\mathsf{x}_n)\}^2 + \frac{\lambda}{2}\mathsf{w}^T\mathsf{w} \tag{3.27}$$

- 高斯先验分布的推广
  - $q = 2$ 表示高斯分布

$$p(\mathsf{w}|\alpha) = [\frac{q}{2}(\frac{\alpha}{2})^{1/q}\frac{1}{\Gamma(1/q)}]^M \exp(-\frac{\alpha}{2}\sum_{j=0}^{M-1}|w_j|^q) \tag{3.56}$$

3.3.2. 预测值的分布

预测分布 (predictive distribution) ：用于帮助新的 x 值预测出 t 的值。

- 预测分布的定义：$p(t|\mathsf{t},\alpha,\beta) = \int p(t|\mathsf{w},\beta)p(\mathsf{w}|\mathsf{t},\alpha,\beta)\mathrm{dw}$
  - $\mathsf{t} = (t_1,\cdots,t_N)^T$ ：训练数据的目标变量的值组成的向量
- 目标变量的条件概率分布

$$p(t|\mathsf{x},\mathsf{w},\beta) = \mathcal{N}(t|y(\mathsf{x},\mathsf{w},\beta^{-1})) \tag{3.8}$$

- 预测分布的计算：$p(t|\mathsf{x},\mathsf{t},\alpha,\beta) = \mathcal{N}(t|\mathsf{m}_N^T\phi(\mathsf{x}),\sigma_N^2(\mathsf{x}))$
  - (Eq 3.59) : $\sigma_N^2(\mathsf{x}) = \beta^{-1} + \phi(\mathsf{x})^T\mathsf{S}_N\phi(\mathsf{x})$
    - β^{-1} ：表示数据中的噪声
    - $\phi(\mathsf{x})^T\mathsf{S}_N\phi(\mathsf{x})$ ：表示了与参数 w 关联的不确定性，当 $N \to \infty$ 时，超于 0
- (Fig 3.8) ：每个点的预测方差与 x 的函数关系
  - 预测的不确定性依赖于数据 x，并且在数据点的领域内最小。
  - 不确定性的程度随着观测数据的增多而逐渐减小
- (Fig 3.9) ：不同数据点 x 的预测之间的协方差

3.3.4. 等价核

预测均值

$$y(\mathsf{x},\mathsf{m}_N) = \mathsf{m}_N^T\phi(\mathsf{x}) = \beta\phi(\mathsf{x})^T\mathsf{S}_N\Phi^T\mathsf{t} = \sum_{n=1}^N \beta\phi(\mathsf{x})^T\mathsf{S}_N\phi(\mathsf{x}_n)\mathsf{t} = \sum_{n=1}^N k(\mathsf{x},\mathsf{x}_n)t_n$$

- (Eq 3.3) : $y(\mathrm{x},\mathrm{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathrm{x}) = \mathrm{w}^T \phi(\mathrm{x})$
- (Eq 3.53) : $\mathrm{m}_N = \beta \mathrm{S}_N \Phi^T \mathrm{t}$
- (Eq 3.51) : $\mathrm{S}_N^{-1} = \mathrm{S}_0^{-1} + \beta \Phi^T \Phi$
- 等价核 (Eq 3.62) ：$k(\mathrm{x}, \mathrm{x}') = \beta \phi(\mathrm{x})^T \mathrm{S}_N \phi(\mathrm{x}')$
- (Eq 3.49) : $p(\mathrm{w}|\mathrm{t}) = \mathcal{N}(\mathrm{w}|\mathrm{m}_N, \mathrm{S}_N)$
- $\mathrm{cov}[y(\mathrm{x}), y(\mathrm{x}')] = \mathrm{cov}[\phi(\mathrm{x})^T \mathrm{w}, \mathrm{w}^T \phi(\mathrm{x}')] = \phi(\mathrm{x})^T \mathrm{S}_N \phi(\mathrm{x}') = \beta^{-1} k(\mathrm{x}, \mathrm{x}')$
- 等价核 (equivalent kernel) 也称为平滑矩阵 (smoother matrix)
  - 线性平滑 (linear smoother) ：通过对训练集里目标值进行线性组合做预测。
  - 等价核定义了模型的权值：$\sum_{n=1}^{N} k(\mathrm{x}, \mathrm{x}_n) = 1$
    * 距离 x 较近的数据点可以赋予一个较高的权值，距离 x 较远的数据点可以赋予一个较低的权值。
    * 附近的点预测均值相关性较高，远处的点预测均值相关性较低。
  - 等价核可以表示为非线性函数的向量的内积的形式：$k(\mathrm{x}, \mathrm{z}) = \psi(\mathrm{x})^T \psi(\mathrm{z})$
- 用核函数表示线性回归模型的方法
  - 线性基函数模型的后验均值可以解释为核方法，即隐式地定义了一个等价核
  - 直接定义一个局部的核函数，然后在给定观测数据集的条件下，使用这个核函数对新的输入变量 x 做预测，这个框架称为高斯过程 (Gaussian process)

3.4. 基于贝叶斯方法的模型比较

(从贝叶斯的角度考虑模型选择问题，如果理解有困难，还可以参考 [11] Ch 09. P 392)

- 前提条件
  - 模型：即观测数据集 $\mathcal{D}$ 上的概率分布
  - 模型 $\{\mathcal{M}_i\}, i = 1, \cdots, L$ 的比较
  - 后验分布：$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i) p(\mathcal{D}|\mathcal{M}_i)$

模型证据 (model evidence) ：表达了数据展现出的不同模型的优先级。也叫边缘似然 (marginal likelihood)。因为可以被看作模型空间中的似然函数。还是估计参数的后验分布时出现在贝叶斯定理的分母中的归一化项。

- $p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathrm{w}, \mathcal{M}_i) p(\mathrm{w}|\mathcal{M}_i) \mathrm{dw}$

贝叶斯因子 (Bayes factor) $p(\mathcal{D}|\mathcal{M}_i) / p(\mathcal{D}|\mathcal{M}_j)$ ：是两个模型的模型证据的比值。

模型有一个参数 w 的情形

[11] Duda R O,Peter E Hart,etc. 李宏东等译。模式分类。机械工业出版社。2003.

- $p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)\mathrm{d}w \simeq p(\mathcal{D}|w_{MAP})\frac{\Delta w_{后验}}{\Delta w_{先验}}$
- $\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{MAP}) + \ln(\frac{\Delta w_{后验}}{\Delta w_{先验}})$
  - $\ln p(\mathcal{D}|w_{MAP})$：基于最可能的参数（$w_{MAP}$）拟合数据
  - $\ln(\frac{\Delta w_{后验}}{\Delta w_{先验}})$：根据模型的复杂度来惩罚模型，即 $\lambda = (\frac{\Delta \mathrm{w}_{后验}}{\Delta \mathrm{w}_{先验}})$

模型有 M 个参数的情形

- $\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\mathrm{w}_{MAP}) + M\ln(\frac{\Delta \mathrm{w}_{后验}}{\Delta \mathrm{w}_{先验}})$
  - 模型复杂度增加时，$\ln p(\mathcal{D}|\mathrm{w}_{MAP})$ 会变大，因为模型越复杂，拟合度越好。
  - 复杂度惩罚项的大小随着模型中可调节参数 M 的数量线性增加
  - 模型复杂度增加时，$M\ln(\frac{\Delta \mathrm{w}_{后验}}{\Delta \mathrm{w}_{先验}})$ 会变小，因为受限于 M。

预测分布：对各个模型的预测分布 $p(t|\mathrm{x}, \mathcal{M}_i, \mathcal{D})$ 求加权平均，权值为这些模型的后验概率 $p(\mathcal{M}_i|\mathcal{D})$。

$$p(t|\mathrm{x}, \mathcal{D}) = \sum_{i=1}^{L} p(t|\mathrm{x}, \mathcal{M}_i, \mathcal{D})p(\mathcal{M}_i|\mathcal{D})$$

模型选择（model selection）：对于「模型求平均」的简单近似是使用最有可能的模型做预测。

- 简单的模型拟合数据的效果较差
- 复杂的模型的预测概率会分散到过多的可能的数据集当中，从而每个数据集赋予的概率都相对较小。

贝叶斯模型比较

- 生成数据的真实概率分布包含在考虑的模型集合当中
- 贝叶斯模型会倾向于选择更加正确的模型
- 贝叶斯方法需要对模型的形式作出假设，如果假设不合理，那么结果就会出错。

最优评估方式：保留一个独立的测试数据集，使用这个数据集来评估最终系统的表现。

3.5. 证据的近似计算

近似计算方法用于解决模型证据中无法对所有的变量进行完整积分的问题。

- 计算过程
  - 首先对参数 w 求积分，得到边缘似然函数（marginal likelihood function）
  - 然后通过最大边缘似然函数，确定超参数的值。

  - 这个框架在统计学中称为经验贝叶斯 (empirical Bayes) 或者第二类最大似然 (type 2 maximum likelihood) 或者推广的最大似然 (generalized maximum likelihood)，在机器学习中称为证据近似 (evidence approximation)。

公式推导

- (Eq 3.74) : $p(t|\mathrm{t}) = \int\int\int p(t|\mathrm{w},\beta)p(\mathrm{w}|\mathrm{t},\alpha,\beta)p(\alpha,\beta|\mathrm{t})\mathrm{dw}\ \mathrm{d}\alpha\ \mathrm{d}\beta$
  - (Eq 3.8) : $p(t|\mathrm{w},\beta) = p(t|\mathrm{x},\mathrm{w},\beta) = \mathcal{N}(t|y(\mathrm{x},\mathrm{w}),\beta^{-1})$
  - (Eq 3.49) : $p(\mathrm{w}|\mathrm{t},\alpha,\beta) = \mathcal{N}(\mathrm{w}|\mathrm{m}_N,\mathrm{S}_N)$
    * (Eq 3.53) : $\mathrm{m}_N = \beta\mathrm{S}_N\Phi^T\mathrm{t}$
    * (Eq 3.54) : $\mathrm{S}_N^{-1} = \alpha\mathrm{I} + \beta\Phi^T\Phi$
  - 如果 $p(\alpha,\beta|\mathrm{t})$ 在 $\widehat{\alpha}$ 和 $\widehat{\beta}$ 附近有尖峰
  - 则 $\alpha = \widehat{\alpha}, \beta = \widehat{\beta}$
  - 得 $p(t|\mathrm{t}) \simeq p(t|\mathrm{t},\widehat{\alpha},\widehat{\beta}) = \int p(t|\mathrm{w},\widehat{\beta})p(\mathrm{w}|\mathrm{t},\widehat{\alpha},\widehat{\beta})\mathrm{dw}$
  - 否则 : $p(\alpha,\beta|\mathrm{t}) \propto p(\mathrm{t}|\alpha,\beta)p(\alpha,\beta)$

最大化对数证据

- 解析地计算证据函数。然后令其导数为零，得到对于超参数的重新估计方程 (Sec 3.5.2)
- 期望最大化 (EM) 算法 (Sec 9.3.4)

证据函数的计算与最大化 (有点难，建议手工推导，看懂图的含义，对于理解**贝叶斯模型比较**有帮助)

3.5.1 计算证据函数

边缘似然函数

$$p(\mathrm{t}|\alpha,\beta) = \int p(\mathrm{t}|\mathrm{w},\beta)p(\mathrm{w}|\alpha)\mathrm{dw}$$

计算积分的方法

- 使用线性——高斯模型的条件概率分布

$$p(y) = \mathcal{N}(y|A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$$

- 对指数项配平方，使用高斯分布的归一化系数的基本形式
  - (Eq 3.78) : $p(\mathrm{t}|\alpha,\beta) = (\frac{\beta}{2\pi})^{N/2} + (\frac{\alpha}{2\pi})^{M/2}\int \exp\{-E(\mathrm{w})\}\mathrm{dw}$

    * (Eq 3.10) : $p(\mathrm{t}|\mathrm{X},\mathrm{w},\beta)=\prod_{n=1}^{N}\mathcal{N}(t_n|\mathrm{w}^T\phi(\mathrm{x}_n),\beta^{-1})$
    * (Eq 3.12) :$E_D(\mathrm{w})=\frac{1}{2}\sum_{n=1}^{N}\{t_n-\mathrm{w}^T\phi(\mathrm{x}_n)\}^2$
    * (Eq 3.52) : $p(\mathrm{w}|\alpha)=\mathcal{N}(\mathrm{w}|0,\alpha^{-1}\mathrm{I})$
    * N : 数据集中数据点的个数
    * M : 参数 w 的维数
  - (Eq 3.79) :$E(\mathrm{w})=\beta E_D(\mathrm{w})+\alpha E_W(\mathrm{w})=\frac{\beta}{2}\|\mathrm{t}-\Phi\mathrm{w}\|^2+\frac{\alpha}{2}\mathrm{w}^T\mathrm{w}$
    * (Eq 3.27) :$\frac{1}{2}\sum_{n=1}^{N}\{t_n-\mathrm{w}^T\phi(\mathrm{x}_n)\}^2+\frac{\lambda}{2}\mathrm{w}^T\mathrm{w}$
  - $E(\mathrm{w})=E(\mathrm{m}_N)+\frac{1}{2}(\mathrm{w}-\mathrm{m}_N)^TA(\mathrm{w}-\mathrm{m}_N)$
    * $A=\alpha I+\beta\Phi^T\Phi=\nabla\nabla E(\mathrm{w})$, 也称为 Hessian 矩阵
    * $E(\mathrm{m}_N)=\frac{\beta}{2}\|\mathrm{t}-\Phi\mathrm{m}_N\|^2+\frac{\alpha}{2}\mathrm{m}_N^T\mathrm{m}_N$
    * $\mathrm{m}_N=\beta A^{-1}\Phi^T\mathrm{t}$
      · (Eq 3.53) : $\mathrm{m}_N=\beta\mathrm{S}_N\Phi^T\mathrm{t}$
      · (Eq 3.54) : $\mathrm{S}_N^{-1}=\alpha I+\beta\Phi^T\Phi$
      · $A=\mathrm{S}_N^{-1}$

比较多元高斯分布的归一化系数，关于 w 的积分计算

$$\begin{aligned}\int\exp\{-E(\mathrm{w})\}\mathrm{dw}&=\exp\{-E(\mathrm{m}_N)\}\int\exp\{-\frac{1}{2}(\mathrm{w}\text{-}\mathrm{m}_N)A\mathrm{w}\text{-}\mathrm{m}_N)\}\mathrm{dw}\\&=\exp\{-E(\mathrm{m}_N)\}(2\pi)^{M/2}|A|^{-1/2}\end{aligned}$$

模型证据函数：基于边缘似然函数 (Eq 3.78) 的对数得

$$\ln p(\mathrm{t}|\alpha,\beta)=\frac{M}{2}\ln\alpha+\frac{N}{2}\ln\beta-E(\mathrm{m}_N)-\frac{1}{2}\ln|A|-\frac{N}{2}\ln(2\pi)\tag{3.86}$$

(理解 Fig 3.14，明白模型对数证据与阶数 M 之间的关系)

3.5.2 最大化证据函数

边缘似然函数 (Eq 3.78) 关于 α 的最大化

- 定义特征向量方程：$(\beta\Phi^T\Phi)\mathrm{u}_i=\lambda_i\mathrm{u}_i$
  - $A=\alpha I+\beta\Phi^T\Phi$ 的特征值为 $\alpha+\lambda_i$

$$\frac{d}{d\alpha}\ln|A|=\frac{d}{d\alpha}\ln\prod_i(\lambda_i+\alpha)=\frac{d}{d\alpha}\sum_i\ln(\lambda_i+\alpha)=\sum_i\frac{1}{\lambda_i+\alpha}$$

- 证据函数关于 α 的驻点满足

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\mathsf{m}_N^T\mathsf{m}_N - \frac{1}{2}\sum_i \frac{1}{\lambda_i + \alpha}$$

- 两侧乘以 2α 得

$$\alpha\mathsf{m}_N^T\mathsf{m}_N = M - \alpha\sum_i \frac{1}{\lambda_i + \alpha} = \gamma \tag{3.90}$$

- 由于 i 的求和式中一共有 M 项，因此 γ 重写为

$$\gamma = \sum_i \frac{\lambda_i}{\lambda_i + \alpha} \tag{3.91}$$

- 基于 (Eq 3.90)，最大化边缘似然函数的 α 为

$$\alpha = \frac{\gamma}{\mathsf{m}_N^T\mathsf{m}_N}$$

- 注 1：α 是一个隐式解，因为 α 既与 γ 相关，还与众数 m_N 相关，因此只能通过迭代方法求解。
- 注 2：α 的求解基于对训练集的观察确定的，不像最大似然方法需要独立的数据集优化模型的复杂度

边缘似然函数 (Eq 3.78) 关于 β 的最大化

$$\frac{d}{d\beta}\ln|A| = \frac{d}{d\beta}\sum_i \ln(\lambda_i + \alpha) = \frac{1}{\beta}\sum_i \lambda_i\lambda_i + \alpha = \frac{\gamma}{\beta}$$

- 证据函数关于 β 的驻点满足

$$0 = \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^N \{t_n - \mathsf{m}_N^T\phi(\mathsf{x}_n)\}^2 - \frac{\gamma}{2\beta} \tag{3.94}$$

- 基于 (Eq 3.94)，最大化边缘似然函数的 β 为

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^N \{t_n - \mathsf{m}_N^T\phi(\mathsf{x}_n)\}^2 \tag{3.95}$$

- 注：β 是一个隐式解，因此只能通过迭代方法求解

3.5.3 参数的有效数量

(Eq 3.91) 的 γ 度量了已经良好确定的参数的数目。

对比 β 的估计公式

- (Eq 3.95) : $\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}\{t_n - \mathsf{m}_N^T\phi(\mathsf{x}_n)\}^2$
  - 分母中的因子 $N-\gamma$ 抵消了最大似然解的偏差
  - 参数的总数量 M，由数据确定的有效参数的数量 γ，剩余的 $M-\gamma$ 个参数被先验概率分布设置为较小的值
- (Eq 3.21) : $\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1}^{N}\{t_n - \mathsf{w}_{ML}^T\phi(\mathsf{x}_n)\}^2$
- (Eq 1.56) : $\mathbb{E}[\sigma^2_{ML}] = \frac{N-1}{N}\sigma^2$
- (Eq 1.59) : $\tilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$
- (Eq 3.97) : $\sigma^2_{MAP} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$
  - 分母中的因子 $N-1$ 反映了模型中的一个自由度被用于拟合均值，抵消了最大似然解的偏差

极限情况 $N >> M$，基于 (Eq 3.87) 可知特征值 λ_i 随着数据集规模的增加而增大，可得

$$\gamma = M$$
$$\alpha = \frac{M}{2E_W(\mathsf{m}_N)}$$
$$\beta = \frac{M}{2E_D(\mathsf{m}_N)}$$

3.6. 固定基函数的局限性

- 优点：通过选择合适的基函数，可以建立输入向量到目标向量之间的任意非线性函数映射。
- 缺点：因为基函数在观测到任何数据之前就被固定下来，因此基函数的数量会随着输入空间的维度 D 的增长而呈指数方式增长。(参考 Sec 1.4.)

真实数据集的两个性质

- 数据向量通常位于一个非线性流形内部
  - 基于输入变量之间的相关性，流形本身的维度小于输入空间的维度
  - 使用局部基函数，基函数只分布在输入空间中包含数据的区域。
  - 具体应用

        * 径向基函数网络
        * 支持向量机
        * 相关向量机
        * 神经网络：使用可调节的基函数，通过调节参数使基函数会按照数据流形发生变化。

- 目标变量可能只依赖于数据流形中的少量可能的方向
    - 具体应用
        * 神经网络：选择输入空间中基函数产生响应的方向

03. 小结

这章的标题在许多模式识别与机器学习的书中都见过，但是作者采用贝叶斯和核函数的视角来分析和解释这个模型，使模型的理解难度加大，但是对于知识的扩展和补充很有裨益。如果对贝叶斯方法了解不足，可以参考 [12] 和 [13]，虽然这些书中的内容并不能减轻阅读这一章的难度，但至少可以为理解贝叶斯方法打下基础

[12] Duda R O,Peter E Hart,etc. 李宏东等译。模式分类。机械工业出版社。2003.
[13] Andrew R. Webb. 统计模式识别。电子工业出版社。2004.