31

学 数 应 用

随机模型与应用概率

Luc Devroye
László Györfi 著
Gábor Lugosi

王业江 译

模式识别的概率理论

Springer

模式识别的概率理论

Luc Devroye
László Györfi 著
Gábor Lugosi
王业江 译

Spriger 出版社

爱人，爱自己

序

生活不过是一段长的随机漫步。事物因为情况正好是对的而被创造。更多时候，创造，如本书，是偶然间的。非参数估计在 50 年代和 60 年代产生并在 60 年代末开始以疯狂的速度发展，其增长淹没模式识别。在 60 年代中叶，两个年轻人，Tom Cover 和 Peter Haart 展示简单最近邻规则的世界，其保证误差最多两倍于最好的判别法。Tom Cover 的结果对 Terry Wagner 有深刻的影响，Wagner 成为德克萨斯大学奥斯汀分校的教授，为年轻的非参数估计领域带来概率严格性。在 1971 年，Vapnik 和 Chervonenkis 开始发表一系列革命性论文，对模式识别具有深刻影响。但他们的工作在那时没有广为人知。但是，Tom Cover 与 Terry Wagner 已注意到这工作的潜力，Terry 叫 Luc Devroye（本书作者）阅读这些论文为他在德克萨斯大学的博士论文做准备。这一年是 1974 年。Luc Devroye 最后在德克萨斯很偶然，多亏朋友和同伴 BelgianWilly Wouters 的一个建议，他为 Luc 和 Terry 牵线搭桥。在 Luc 的学位论文于 1976 年发表时，模式识别已经开始认真起步。在理论层面，重要的属性仍在被发现。在 1977 年，Stone 通过表明存在非参数规则（rule）在数据的所有分布下收敛震惊了非参数社区。这称为分布无关性（distribution-free）或普遍一致性（universal consistency）。是它令非参数方法如此迷人。然而，少有研究人员关心普遍一致性，除了 Laci Györfi 这个明显的例外，他当时在布达佩斯一个有活力的非参数专家团队工作，团队人员包括 Sándor Csibi，Józssef Fritz 和 Pál Révész。

所以由该共同观点连接，Luc 和 Laci 决定在 80 年代初通力合作。在 1982 年，他们写了六章关于非参数回归函数估计的一本书，但没有发表。事实上，这些笔记至今仍在办公室的抽屉中。他们觉得这个主题还未成熟。关于非参数密度估计的书在 1985 年看见了曙光。不幸的是，作为真正的婴儿潮一代，Luc 和 Laci 都没有时间，在 1985 年后，来写一本关于非参数模式识别的教材。Gabor Lugosi 加了进来。他在 Laci 的指导下于 1991 年获得博士学位。Gabor 在 1992 年准备一组关于该主题粗糙的课堂笔记，并在 1993 年准备加入该项目（本书）。有了新的活力，我们开始写作本书，这本该至少在过去十年中写完的书。讨论与工作会议在布达佩斯，蒙特利尔，鲁汶和卢旺-拉-纽维召开。在鲁汶，我们热情的东道主是 Ed van der Meulen 和 Jan Beirlant。在卢旺-拉-纽维，我们美食与精神上受到 Leopold Simar 和 Irene Gijbels 的支持。我们感谢他们。新的结果累积，我们抵抗尝试在期刊上发表他们的冲动。最终，在 1995 年 5 月，书稿已膨胀到必须送到出版商不可的程度，否则它将变成百科全书。一些重要的悬而未决问题很快变成自虐的习题或胡乱猜测。我们将在导论中解释课题选择，课堂使用，章节关系和个人观点。我们很抱歉，确实，为书中存在的所有错误。

我们受许多人触动、影响、指引和教导。Terry Wagner 对美丽的非参数问题的严格和品味感染了我们一生。我们感谢过去和现在在非参数论文中的合作者，Alain Berlinet, Michel Broniatowski, Ricardo Cao, Paul Deheuvels, Andras Farago, Adam

Krzyzak，Tamas Linder，Andrew Nobel，Mirek Pawlak，Igor Vajda，HarroWalk 和 Ken Zeger。Tamas Linder 读了本书的绝大部分并提供了无价的反馈。它的帮助尤为感激。几章被布达佩斯的学生批判阅读过。我们感谢他们，尤其是 Andras Antos，Miklos Csuros，Balazs Kegl，Istvan Pali 和 Marti Pinter。最后，下面是一字母顺序朋友表，他们直接或间接的有助于我们的非参数的掌握与热爱：Andrew、Roger Barron、Denis Bosq、Prabhir Burman、Tom Cover、Antonio Cuevas、Pierre Devijver、Ricardo Fraiman、Ned Glick、Wenceslao Gonzalez-Manteiga、Peter Hall、Eiichi Isogai、Ed Mack、Arthur Nadas、Georg Pflug、George Roussas、Winfried Stute、Tamas Szabados、Godfried Toussaint、Sid Yakowitz 和 Yannis Yatracos。

Gábor 勤奋的手打了整部书稿和协调所有贡献。他在这个过程中快变成一个 Tex 专家。几张图由 Gabor 和 Luc 使用 *idraw* 和 *xfig* 制作。大多数绘画由 Luc 和一个 McGill 大学本科生 Hisham Petry 直接在 PostScript 中编程，对他我们表示感激。对 Gábor，本书在他工作之初到来。不幸的是，其它两个作者没有这么幸运。因为 Luc 和 Laci 两人都觉得他们不会再写另外一本关于非参数模式识别的书，随机漫游必须继续，他们决定记述他们对主题领域的总览到纸上，同时尝试区分开不相关点与关键点。当然，这已造成本文的长度。

至今，我们的随机涉足已是很开心。无独有偶，Luc 与 Bea 结婚了，她是这个世界上最善解人意的人，并有了两个女儿，Natasha 和 Birgit，她两没有迷失在她们的随机课程中。相似地，Laci 有一个一样棒的妻子 Kati 和两个性格稳重的孩子 Kati 和 Janos。在准备本书期间，Gábor 遇到了一个令人妙龄少女 Arrate。他们最近决定将他们的生活绑在一起。

在不那么多情和迷人的方面,我们真诚的感激 NSERC CANADA、FCAR QUEBEC、OTKA HUNGARY 和匈牙利科学院与比利时皇家科学院的交换项目的研究支持。本书早期版本在布达佩斯理工大学、卡托利克鲁汶大学、斯图加特大学和蒙彼利埃第二大学的一些课程上做尝试。我们为这些学生给予的帮助使本书更好表示感谢。

Luc Devroye
László Györfi
Gábor Lugosi

目录

1 导论

模式识别（pattern recognition）或模式判别（discrimination）是关于猜测或预测观察（observation）未知属性，是一个离散量，例如黑或白，一或零，生病或健康，真实或虚假。一个观察是一组数值度量，例如图像（可以是一列比特值，每个像素一个比特），天气数据向量，心电图或一个在支票上经数字化的签名。更正式的说，一个观察是一个 d 维向量 x。观察的未知属性称为类（class）。类被表示为 y，其在一个有限集 $\{1,2,...,M\}$ 取值。在模式识别中，需构建一个函数 $g(x):\mathcal{R}^d \to \{1,2,...,M\}$，代表给定 x 时对 y 的猜测。映射 g 被称为分类器。我们的分类器出现误差若 $g(x) \neq y$。

观察

类

误差

该怎么构建一个规则 g 依赖于具体的问题。专家能被召集进行医疗诊断或者地震预测，他们尝试建模 g 靠他们的知识和经验，常常是反复试错。理论上而言，每个专家操作一个内置的分类器 g，但描述这个 g 显式以数学形式并不是简单的事。x 空间纯粹的量级和丰富性将击败即使是最好的专家。不可能辨别出 g 为一个人能在未来观测到的所有可能的 x 值。我们必须准备与不完美的分类器共存。实际上，我们该如何度量分类器的质量？我们不能放弃一个分类器仅因为它错误分类了某个特定 x 值。首先，如果观察没有完全描述潜在的过程（即若 y 不是 x 的确定性函数），则可能同一个 x 将导致两个不同的 y 值在不同的情况下。例如，如果我们仅测量人身体的水容量，我们发现人是脱水的，则原因（类）可能范围从热的天气下低水摄入到严重腹泻。因此我们引入一个概率设置，令 (X,Y) 是 $\mathcal{R} \times \{1,2,...,M\}$ 值随机对。(X,Y) 的分布描述在现实中遇见特定对的频率。错误发生如果 $g(X) \neq Y$，分类器 g 的误差概率为

误差概率

$$L(g) = \mathbf{P}\{g(X) \neq Y\}.$$

存在一个最优分类器 g^*，其定义为

$$g^* = \underset{g:\mathcal{R}\to\{1,...,M\}}{\arg\min} \mathbf{P}\{g(X) \neq Y\}.$$

注意 g^* 依赖于 (X,Y) 的分布。如果这个分布是已知的，则 g^* 能被计算出来。寻找 g^* 的问题被称为贝叶斯问题，分类器 g^* 被称为贝叶斯分类器（或贝叶斯规则）。最小误差概率被称为贝叶斯误差，其被表示为 $L^* = L(g^*)$。大多数时候，(X,Y) 的分布是未知的，所以 g^* 也是未知的。

贝叶斯分类器

我们不咨询专家来尝试重建 g^*，但可以访问过去观测的好的对 $(X_i,Y_i), 1 \leq i \leq n$ 数据库。这个数据库可能是经验观测的结果（如气象数据、指纹数据、心电图数据或手写字母）。它也可能通过专家或教授在看见 X_i 值后填写 Y_i 的值获得。寻找具有小错误概率的分类器 g 是无望的，除非能保证 (X_i,Y_i) 联合 [概率] 一定程度上代表未知的分布。在本书中我们假设数据 $(X_1,Y_1),...,(X_n,Y_n)$ 是一列独立同分布（$i.i.d$）的随机对，每对具有与 (X,Y) 相同的分布。这确实是一个非常强的假设。

数据，独立同分布

但是，一些新出现的理论结果表明基于轻微相关数据对和独立同分布数据对的分类器表现大致相同。而且，简单模型更容易理解和解释。

学习

分类器构建在 $X_1, Y_1, ..., X_N, Y_N$ 的基础上，表示为 $g_n; Y$ 由 $g_n(X; X_1, Y_1, ..., X_n, Y_n)$ 预测。构建 g_n 的过程称为学习（有监督学习或有教师学习）。g_n 的表现由条件误差概率测量

$$L_n = L(g_n) = \mathbf{P}\{g_n(X; X_1, Y_1, \ldots, X_n, Y_n) \neq Y \mid X_1, Y_1, \ldots, X_n, Y_n)\}.$$

这是一个随机变量，因为它依赖于数据 $\{X_i, Y_i\}$。故 L_n 在 (X, Y) 分布上取期望，但数据 $\{X_i, Y_i\}$ 是固定的。[继续] 在数据上取平均是不自然的，因为在一个给定的应用中，人们与手头的数据接触。[因此] 知道 EL_n 是略微有用的，因为这个数值表示的是平均数据序列而不是你自己的数据序列的质量。本文因此关于 L_n，其表示误差的条件概率。

分类器，规则

一个单独映射 $g_n : \mathcal{R}^d \times \left\{\mathcal{R}^d \times \{1, \ldots, M\}\right\}^n \longrightarrow \{1, \ldots, M\}$ 仍称为为分类器。一个序列 $\{g_n, n \geq 1\}$ 称为（判别）规则。因此分类器是函数，规则是一列函数。

初学者可能会问几个简单的问题像：该怎么构建好的分类器？分类器能做的多好？分类器 A 是否比分类器 B 更好？我们能否估计分类器的性能？最好的分类器是哪一个？本书部分回答了这些简单问题。大量的精力花在初学者问题的数学形式化之上。对我们而言，一个规则（不是分类器）是优的，如果它是一致的，即若

一致的

$$\lim_{n \to \infty} \mathbf{E} L_n = L^*$$

或等价的，若在概率上 $L_n \to L^*$ 当 $n \to \infty$。我们假设读者良好掌握概率论的基础知识，包括例如按概率收敛、平均强大数定律和条件概率等概念。在附录给出结果和定义的选择对本章可能有用。一致规则保证取更多样本本质上足够粗略重建未知的，因为 L_n 能根据需要任意接近 L^*。换言之，无穷信息能从有限样本中收集。没有该保证，我们不会有动机来取更多的样本。我们应小心不能为了规则的一致性附加条件在 (X, Y) 上，因为这些条件可能是不可验证的。如果一个规则对所有 (X, Y) 的分布是一致的，则称为是普遍一致的。

普遍一致性

有趣的是，直至 1977 年，仍未清楚是否存在一个普遍一致规则。所有 1977 年前的一致性结果均来自对 (X, Y) 的限制。在 1977 年，Stone 表明可以取任意 k 最近邻规则令 $k = k(n) \to \infty$ 和 $k/n \to 0$。k 最近邻分类器 $g_n(x)$ 取 $\{(X_1, Y_1), ..., (X_n, Y_n)\}$ 的子集 k 对 (X_i, Y_i) 中 Y_i 的多数票，使 $\|X_i - x\|$ 取得最小值（即，X_i 与 x 最接近）。自 Stone 证明 k 最近邻规则具有普遍一致性后，几个其他规则也被证明一样具有普遍一致性。本书强调普遍性（universality）并期望给出合理解释该方向的发展。

Stone 定理

k 最近邻分类器

概率学家可能想知道为什么我们没有使用概率收敛性在我们的一致性定义中。确实，强一致性——L_n 概率收敛至 L^*——蕴含当 n 增长时几乎每个样本收敛。幸运的是，对大多表现良好的规则，一致性和强一致性是等价的。例如 k 最近邻规则 $k \to \infty$ 和 $k/n \to 0$ 蕴含着按概率 1 收敛 $L_n \to L^*$。等价性将会被处理，但它不是主要的关心点。大多数情况，不是全部，等价结果基于一些强大的中心不等式，如 McDiarmid 不等式。例如，我们可以证明对 k 最近邻规则存在 $c > 0$，任 $\forall \epsilon > 0$，$\exists N(\epsilon) > 0$ 依赖于 (X, Y) 的分布，有

$$\mathbf{P}\{L_n - L^* > \epsilon\} \leq e^{-cn\epsilon^2}, n \geq N(\epsilon).$$

这也说明了本书的另一个关注点——不等式。只要有可能，我们通过显式不等式说明理由或做证明。各种参数能在这些不定式中被替换以允许使用者关于样本大小做出结论或允许识别出最重要的参数。

书中的材料常是技术性和枯燥的。因此为了聚焦于主要问题，我们保持问题尽量简单：

1. 我们仅处理二进制分类（M=2）。类 Y 在 $\{0,1\}$ 中取值，分类器 g_n 是一个映射 $\mathcal{R}^d \times \left\{\mathcal{R}^d \times \{0,1\}\right\}^n \longrightarrow \{0,1\}$.
2. 我们仅考虑独立同分布的数据序列。我们不允许主动学习（active learning），用户可以确切地选择 X_i。
3. 我们不考虑无限空间。例如 X 不能是随机函数如心电图。X 必须是 $\mathcal{R}^d$-值随机向量。读者需要清楚许多在这给出的结果能不费吹灰之力扩展至某些无限维度量空间。

让我们回到初学者问题。我们知道存在好的规则，但分类器能好的什么程度？明显，在所有情况下 $L^n \geq L^*$。因此重要的是知道 L^* 或预测它，因为若 L^* 很大，则任意分类器将表现不好。但即使 L^* 为零，L_n 仍可能很大。因此，如果有显式概率不等式如

$$\mathbf{P}\{L_n \geq L^* + \epsilon\}$$

将很好。

但是，这些不等式必须依赖于 (X,Y) 的分布。即对任规则，

$$\liminf_{n\to\infty} \sup_{\forall \mathcal{D}(X,Y) \text{ with } L^*+\epsilon<1/2} \mathbf{P}\{L_n \geq L^* + \epsilon\} > 0$$

保证收敛的通用速率并不存在。收敛率学习必须包含某个 (X,Y) 分布的子族。因为这个原因，极少有例外，我们将避开收敛速率的流沙。

即使不存在普遍的性能保证，但我们仍可以满足初学者疑惑，如果我们能足够对手头的规则估计 L_n 通过数据的函数 $\widehat{L_n}$。该函数称为*误差估计*。例如，对于 k 最近邻分类器，我们可以使用*删除估计*

误差估计

$$\widehat{L}_n = \frac{1}{n}\sum_{i=1}^{n} I_{\{g_{ni}(X_i)\neq Y_i\}}$$

其中 $g_{ni}(X_i)$ 基于删除 (X_i,Y_i) 的数据 $(X_1,Y_1),\ldots,(X_n,Y_n)$ 通过 k 最近邻方法分类 X_i。完成该步骤后，我们得到与分布无关的不等式

Rogers-Wagner 不等式

$$\mathbf{P}\left\{\left|\widehat{L}_n - L_n\right| > \epsilon\right\} \leq \frac{6k+1}{n\epsilon^2}$$

前提是距离是紧的以适当的方式。换言之，若不知道 (X,Y) 的分布，我们能有一定信心认为 L_n 位于 $[\widehat{L_n} - \epsilon, \widehat{L_n} - \epsilon]$ 中。因此，对许多分类器用手头数据估计 L_j 确实是可能的。但是，不可能普遍好地估计 L^*：对任 n，任基于数据序列的 L^* 的估计，总存在 (X,Y) 的分布使估计任意差。

我们能否比较规则 $\{g_n\}$ 与 $\{g'_n\}$？再次，答案是否定的：不存在“最优的”分类器（或超分类器），对任意规则 $\{g_n\}$，存在一个 (X,Y) 的分布和另外一个规则 $\{g'_n\}$ 使得对任 n，$\mathbf{E}\left\{L\left(g'_n\right)\right\} < \mathbf{E}\{L\left(g_n\right)\}$。假若存在一个普遍最优分类器，本书可能就没必要了：我们将一直使用它。

这一不存在性意味着实用模式识别器的争辩将永无尽头，模拟特定示例绝不应被用来比较分类器。例如，考虑 1-最近邻分类器，一个简单但不是普遍一致的规则。然而，在所有 k 最近邻分类器中，1-最近邻分类器是可接受的——存在分布使它的期望误差概率比任意 k 最近邻分类器（$k \geq 1$）更优。故它绝不能被忽略。因此我们必须学习所有简单规则，并将为最近邻规则及其衍生规则保留几页空间。我们将举例证明 Cover-Hart 不等式（Cover 和 Hart，1967），表明对所有 (X,Y) 分布上

Cover-Hart 不等式

$$\limsup_{n\to\infty} \mathbf{E}L_n \leq 2L^*$$

这里 L_n 是 1-最近邻规则的误差概率，由于 L^* 常是小的（否则，你将不想做判别），即 L^* 也是小的。因此 1-最近邻规则将表现良好。

最优分类器的不存在性可能会使初学者失望。但我们可以一定程度上改变设置和限制分类器在一个族 $\mathscr{C}$ 中，例如所有的 k 最近邻分类器，k 取遍所有可能的值。是否可以从这个族中取到最优的分类器？用这种方式表达，我们不可能做的比

$$L \stackrel{\text{def}}{=} \inf_{g_n \in \mathscr{C}} \mathbf{P}\{g_n(X) \neq Y\}$$

更好。一般而言 $L > L^*$。有趣的是，存在一个从 $\mathscr{C}$ 中选择分类器并保证获得普遍性能保证的通用范式。它使用经验风险最小化，能在 Vapnik 和 Chervonenkis（1971）的著作中详细的学习该方法。例如，如果我们从 $\mathscr{C}$ 中选择 g_n 通过最小化

经验风险最小化

$$\frac{1}{n}\sum_{i=1}^{n} I_{\{g_n(X_i)\neq Y_i\}}$$

则相应的误差概率 L_n 对任 $\epsilon > 0$ 满足如下不等式：

$$\mathbf{P}\{L_n > L + \epsilon\} \leq 8\left(n^V + 1\right) e^{-n\epsilon^2/128}$$

其中 $V > 0$ 是一个仅依赖于 $\mathscr{C}$ 容量的整数。V 被称为 $\mathscr{C}$ 的*VC 维*，对大的族 $\mathscr{C}$ 其值可以无限大。对充分限制的族 $\mathscr{C}$，V 是有限的，上面给出的显式普遍界能被用来为选定的 g_n（相对于 L，不是 L^*）获得性能保证。上述的界是有效的仅当 $\mathscr{C}$ 独立于数据对 $(X_1,Y_1),...,(X_n,Y_n)$。固定族，例如所有分类器在某个半空间中取 1，在其补空间中取 0 都可以。我们另采样 m 对（除了当前已存在的 n 对）数据，并基于 m 对数据，使用上述 n 对数据来选择最好的 k 值用于 k 最近邻分类器。如你所见，选定的规则是普遍一致的，如果 m 和 n 差异很大，即 $n/logm \to \infty$。我们已自动的解决了选择 k 的问题。回忆一下 Stone 的普遍一致性定理仅告诉我们选择 $k = \mathcal{O}(m)$ 和令 $k \to \infty$，但它并未告诉我们是否 $k \approx m^{0.01}$ 比 $k \approx m^{0.99}$ 更好。经验风险最小化产生一个依赖于数据的随机 k 值，甚至无法保证其趋于无穷大或者其值为 $o(m)$，但选定的规则是普遍一致的。

VC 维

我们实际上没有像标准教材一样提供算法的帮助，除了两个显著例外。易于计算，存储和解释刺激了一些规则的发展。例如，树分类器构建一棵树来存储数据，通过某种典型的垂直于坐标轴的切割方法切分 $\mathscr{R}^d$，是可解释的——向量 X 在构建树的早期阶段切割的分量对获取决策而言是最重要的。专家系统、自动医疗诊断和一系列其他识别规则使用树分类。例如，在自动医疗诊断中，医生会首先检查患者的脉搏（第一分量）。如果其值为 0，则表明患者已死亡。如果其值低于

树分类器

40，则表明患者很虚弱。第一分量至此一共被切割了两次。在每次切割中，我们可能考虑另一分量并继续分解成更多特定情况。几个新的有趣的普遍一致树分类器将在第 20 章中介绍。

第二组分类器是*神经网络分类器*，其发展部分依赖于易实现性。神经网络是 Rosenblatt *感知器* 的后代（Rosenblatt，1956）。这些分类器具有必须训练或靠数据选择的未知参数，在这里我们让数据为 k 最近邻分类器挑选 k。大多数神经网络研究论文处理训练层面，但我们并非如此。当我们说“经由经验风险最小化挑选参数”时，我们将不回答重要的算法复杂性问题。感知器通过一个超平面切分空间并附决策 1 与 0 给两个半空间. 这种简单分类器不是一致的，除了少部分分布外。情况确实如此，例如当 X 在 $\{0,1\}^d$ 中取值（超立方体）并令 X 的分量是相互独立的。单隐藏层神经网络是普遍一致的，如果参数是精心挑选过的。我们将看到考虑双隐藏层会存在一点增益，但是超过两层则是不必要的。

神经网络

感知器

训练算法——在该阶段分类器 g_n 从 $\mathscr{C}$ 中选择——的复杂度无疑是重要的。有些时候，我们像获得在某些变换下仍保持不变的分类器。例如 k 最近邻分类器在坐标轴的非线性变换下并未保持不变。这是一个缺陷因为分量常常是在任意标度（比例）的测量值。切换到对数标度或使用华氏而不是摄氏表示一个标度不应影响优的判别规则。存在 k 最近邻规则的变体具有这里给出的不变性。在字符识别中，有时代表一个字符的向量 X 的所有分量是真实测量值，仅涉及所选字符的向量差，例如最左边与最右边的点，几何中心点，所有黑色像素的权重中心点，最上边和最下边的点。在这种情况下，标度具有至关重要的信息，标度改动下的不变性是不利的。但是，一些正交旋转不变性是有利的。

我们遵循概率论教材的标准概念。因此随机变量标为大写字母如 X，Y 和 Z。概率测度以希腊字母表示如 μ 和 υ。数值和向量用小写字母表示，如 a，b，c，x 和 y。集合也用罗马大写字母表示，但是存在明显的记忆方法：S 表示球，B 表示 Borel 集，诸如此类。如果需要多种集合，我们一般将使用字母表的前几个 (A,B,C)。大多数函数由 f,g,ϕ 和 ψ 表示。花体字母如 $\mathscr{A}$，$\mathscr{C}$ 和 $\mathscr{F}$ 被用来表示函数族或集族。频繁使用到的符号短列表位于本书的末尾。

在本章末尾，你将看见一个有向非循环图，其描述了各章间的依赖关系。很明显未来教师必须选择这些章的一部分子集。所有的章无一例外都是毫无顾忌的理论。我们没有给本书预留繁重的仿真或快捷但肮脏的工程解决方案。本书中的方法必须附加健康剂量的工程知识。理想情况下，学生应该拥有具美丽应用的配套文本，如自动病毒识别，电话窃听语言识别，安全系统语音识别，指纹识别或手写字符识别等。为了从头开始运行一个真实模式识别项目，许多经典的统计模式识别教材都能也应该去参考，因为本书仅限制在模式识别的一般概率理论层面。本书具有超过 430 道习题来帮助学者学习。这些习题包括技能磨练联系，智力题，可爱的拼图，开放性问题和严肃的数学挑战。没有解题手册。本书仅是一个*开始*。把它当成是一个玩具——阅读一些证明，享受一些不等式，学习新的技巧，并研究伪装一个问题看起来像另一个问题的艺术。为了学习而学习。

开始

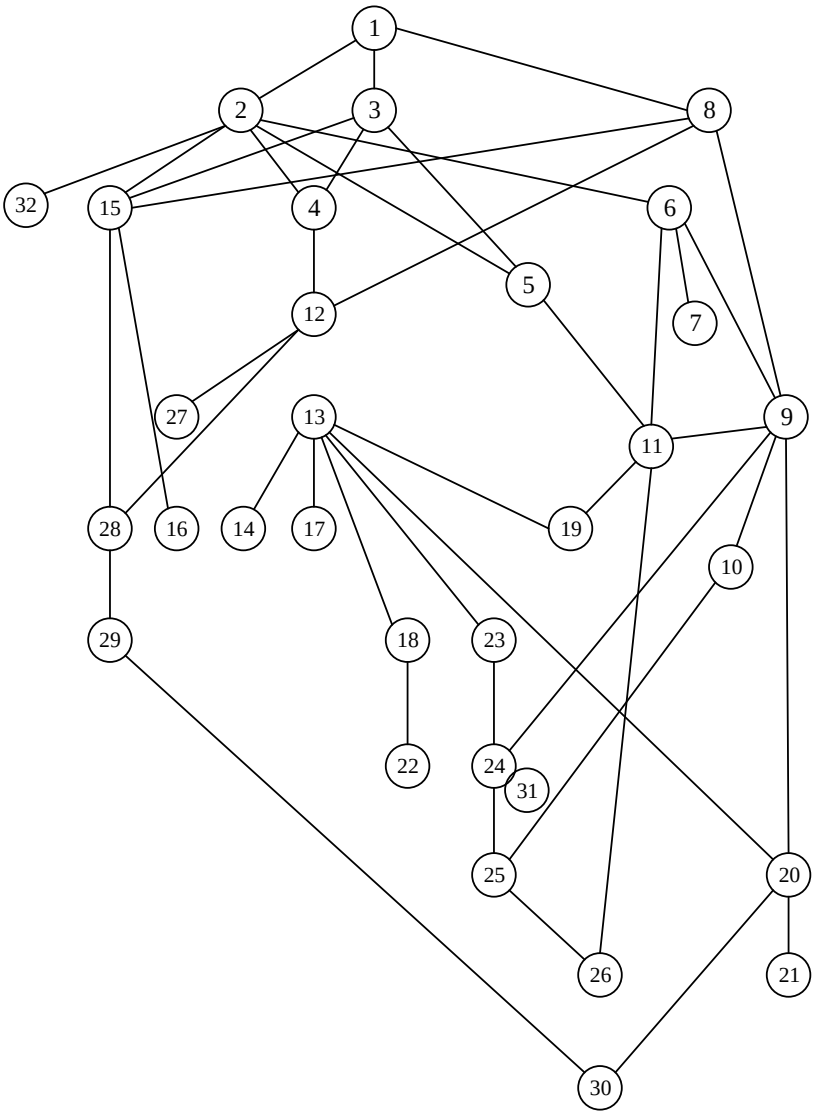

Figure 1.1: 章节依赖图

1 导论 2 贝叶斯误差 3 不等式与替代距离度量 4 线性判别 5 最近邻规则 6 一致性 7 慢收敛速率 8 误差估计 9 常规直方图规则 10 核规则 11 k-最近邻规则的一致性 12 Vapnik-Chervonenkis 理论 13 Vapnik-Chervonenkis 理论的组合视角 14 经验分类器选择的下界 15 最大似然原则 16 参数分类 17 广义线性判别 18 复杂性正则化 19 简化与编辑最近邻规则 20 树分类器 21 数据相关切分 22 切割数据 23 重置估计 24 误差概率的删除估计 25 自动核规则 26 自动最近邻规则 27 超立方体核离散空间 28 Epsilon 熵与完全有界集 29 一致大数定律 30 神经网络 31 其他误差估计 32 特征提取

贝叶斯误差 2

2.1 贝叶斯问题

在本节中，我们定义数学模型和引入整本书都会使用的概念。令 (X,Y) 为一对随机变量，其值从 $\mathcal{R}^d$ 和 $\{0,1\}$ 中取得。随机对 (X,Y) 用许多方式描述：例如可以被一对 (μ,η) 定义，其中 μ 是 X 的*概率测度*，η 是在 X 上 Y 的回归。更精确地说，对于一个 Borel 可测集 $A \subseteq \mathcal{R}^d$，

$$\mu(A) = \mathbf{P}\{X \in A\}$$

和对任 $x \in \mathcal{R}^d$，

$$\eta(x) = \mathbf{P}\{Y = 1 \mid X = x\} = \mathbf{E}\{Y \mid X = x\}$$

因此，$\eta(x)$ 是给定 $X = x$ 时 $Y = 1$ 的条件概率。这样便足够描述 (X,Y) 的分布。对任 $C \subseteq \mathcal{R}^d \times \{0,1\}$，我们有

$$C = \left(C \cap \left(\mathcal{R}^d \times \{0\}\right)\right) \cup \left(C \cap \left(\mathcal{R}^d \times \{1\}\right)\right) \stackrel{\text{def}}{=} C_0 \times \{0\} \cup C_1 \times \{1\}$$

和

$$\begin{aligned}\mathbf{P}\{(X,Y) \in C\} &= \mathbf{P}\{X \in C_0, Y = 0\} + \mathbf{P}\{X \in C_1, Y = 1\} \\ &= \int_{C_0} (1 - \eta(x))\mu(dx) + \int_{C_1} \eta(x)\mu(dx)\end{aligned}$$

这对任意 Borel 可测集 C 都是成立的，(X,Y) 的分布取决于 (μ,η)。函数 η 有时称为*后验概率*。

任意函数 $g : \mathcal{R}^d \to \{0,1\}$ 定义一个分类器或决策函数。g 的误差概率为 $L(g) = \mathbf{P}\{g(x) \neq Y\}$。我们对贝叶斯决策函数更感兴趣

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise} \end{cases}$$

这个决策函数使误差概率达到最小。

定理 2.1 *对任决策函数 $g : \mathcal{R}^d \to \{0,1\}$，*

$$\mathbf{P}\{g^*(X) \neq Y\} \leq \mathbf{P}\{g(X) \neq Y\}$$

即，g^* 是最优决策。

证. 给定 $X = x$，任一决策 g 的条件误差概率可被表示为

$$\begin{aligned}
&\mathbf{P}\{g(X) \neq Y \mid X = x\} \\
&\quad = 1 - \mathbf{P}\{Y = g(X) \mid X = x\} \\
&\quad = 1 - (\mathbf{P}\{Y = 1, g(X) = 1 \mid X = x\} + \mathbf{P}\{Y = 0, g(X) = 0 \mid X = x\}) \\
&\quad = 1 - \left(I_{\{g(x)=1\}}\mathbf{P}\{Y = 1 \mid X = x\} + I_{\{g(x)=0\}}\mathbf{P}\{Y = 0 \mid X = x\}\right) \\
&\quad = 1 - \left(I_{\{g(x)=1\}}\eta(x) + I_{\{g(x)=0\}}(1 - \eta(x))\right),
\end{aligned}$$

其中 I_A 表示集 A 的指示函数。因此对任 $x \in \mathscr{R}^d$，

$$\begin{aligned}
&\mathbf{P}\{g(X) \neq Y \mid X = x\} - \mathbf{P}\{g^*(X) \neq Y \mid X = x\} \\
&\quad = \eta(x)\left(I_{\{g^*(x)=1\}} - I_{\{g(x)=1\}}\right) + (1 - \eta(x))\left(I_{\{g^*(x)=0\}} - I_{\{g(x)=0\}}\right) \\
&\quad = (2\eta(x) - 1)\left(I_{\{g^*(x)=1\}} - I_{\{g(x)=1\}}\right) \\
&\quad \geq 0
\end{aligned}$$

通过 g^* 的定义。在两边对 $\mu(dx)$ 积分，可得到上表述。 [证毕]

备注 2.1 g^* 被称为贝叶斯决策，$L^* = \mathbf{P}\{g^*(X) \neq Y\}$ 被称为贝叶斯误差概率、贝叶斯误差或贝叶斯风险。上述给出的证明表明

$$L(g) = 1 - \mathbf{E}\left\{I_{\{g(X)=1\}}\eta(X) + I_{\{g(X)=0\}}(1 - \eta(X))\right\}$$

特别是

$$L^* = 1 - \mathbf{E}\left\{I_{\{\eta(X)>1/2\}}\eta(X) + I_{\{\eta(X)\leq 1/2\}}(1 - \eta(X))\right\}$$

我们观察到后验概率

$$\eta(x) = \mathbf{P}\{Y = 1 \mid X = x\} = \mathbf{E}\{Y \mid X = x\}$$

最小化均方误差，当 Y 被 $f(X)$ 预测时，其中函数 $f : \mathscr{R}^d \to \mathscr{R}$：

$$\mathbf{E}\left\{(\eta(X) - Y)^2\right\} \leq \mathbf{E}\left\{(f(X) - Y)^2\right\}$$

为了明白上述不等式为真，观察到对任 $x \in \mathscr{R}^d$，

$$\begin{aligned}
&\mathbf{E}\left\{(f(X) - Y)^2 \mid X = x\right\} \\
&\quad = \mathbf{E}\left\{(f(x) - \eta(x) + \eta(x) - Y)^2 \mid X = x\right\} \\
&\quad = (f(x) - \eta(x))^2 + 2(f(x) - \eta(x))\mathbf{E}\{\eta(x) - Y \mid X = x\} \\
&\qquad + \mathbf{E}\left\{(\eta(X) - Y)^2 \mid X = x\right\} \\
&\quad = (f(x) - \eta(x))^2 + \mathbf{E}\left\{(\eta(X) - Y)^2 \mid X = x\right\}
\end{aligned}$$

条件中值，即最小化绝对误差 $\mathbf{E}\{|\ f(X) - Y|\}$ 的函数 f 与贝叶斯规则关系更为密切（见问题 2.12）。

2.2 一个简单例子

让我们考虑预测学生课程表现（及格/不及格）在给定一些重要因子时。首先，令 $Y = 1$ 表示及格，而 $Y = 0$ 代表不及格。取单个观察 X 为每周学习时间。这自身并不是一个万无一失的学生表现预测函数，

因为我们需要更多关于学生思维是否敏捷，健康与否，社交习惯等信息。回归函数 $\eta(x) = \mathbf{P}\{Y = 1 \mid X = x\}$ 可能是在 x 上单调升的。如果已知 $\eta(x) = x/(c + x)$，我们的问题将被解决，因为贝叶斯决策为

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2(\text{ i.e. }, x > c) \\ 0 & \text{otherwise.} \end{cases}$$

相应的贝叶斯误差为

$$L^* = L(g^*) = \mathbf{E}\{\min(\eta(X), 1 - \eta(X))\} = \mathbf{E}\left\{\frac{\min(c, X)}{c + X}\right\}$$

同时我们可以仅从 η 推导出贝叶斯决策，但不能推出贝叶斯风险 L^*——它需要 X 的分布信息。若 $X = c$ 按概率 1 成立（像在军事学校里所有学生每周都专心学习 c 个小时），则 $L^* = 1/2$。若我们的总体很分散，即 X 在 $[0, 4c]$ 中均匀分布，则情况可以有些许改善：

$$L^* = \frac{1}{4c}\int_0^{4c} \frac{\min(c, x)}{c + x} dx = \frac{1}{4}\log\frac{5e}{4} \approx 0.305785$$

与 $x = c$ 不同，判别相当简单。一般来说，判别比估计更加简单是由于这种现象。

2.3 另一简单例子

让我们解决第二个简单例子，其中根据是否一个学生通过或没有通过一门课程决定 $Y = 1$ 或 $Y = 0$。X 代表一个或多个对学生的观察。X 的分量在本例中分别标记为 T,B 和 E，其中 T 是学生看电视的平均时长（小时），B 为每天喝啤酒的平均瓶数，E 是个无形的量，测量额外负面因子的数值，如懒惰程度和学习难度。在这个生造的示例中，我们有

$$Y = \begin{cases} 1 & \text{if } T + B + E < 7 \\ 0 & \text{otherwise.} \end{cases}$$

因此，如果 T，B 和 E 是已知的，则 Y 也是可知的。贝叶斯分类器取 1，若 $T + B + E < 7$，否则取 0。相应的贝叶斯误差概率为 0。不幸的是，E 是无形的，对于观察者是不可用的。我们仅知道 T 和 B 的值。给定 T 和 B，我们何时可以猜测 $Y = 1$? 为了回答这个问题，我们需要知道 (T, B, E) 的联合概率分布，或等价地，(T, B, Y) 的联合概率分布。因此我们假设 T，B 和 E 是独立同分布的指数分布随机变量（它们在 $[0, \infty]$ 上具有密度 e^{-u}）。贝叶斯规则比较 $\mathbf{P}\{Y = 1 \mid T, B\}$ 和 $\mathbf{P}\{Y = 0 \mid T, B\}$，得到的决策最大化这两个值。简单的计算如下所示

$$\begin{aligned} \mathbf{P}\{Y = 1 \mid T, B\} &= \mathbf{P}\{T + B + E < 7 \mid T, B\} \\ &= \mathbf{P}\{E < 7 - T - B \mid T, B\} \\ &= \max\left(0, 1 - e^{-(7-T-B)}\right). \end{aligned}$$

两种决策交叉出现，当这个值为 1/2 时。因此贝叶斯分类器如下所示：

$$g^*(T, B) = \begin{cases} 1 & \text{if } T + B < 7 - \log 2 = 6.306852819\ldots \\ 0 & \text{otherwise.} \end{cases}$$

当然，这个分类器是不完美的。误差概率为

$$\begin{aligned}
&\mathbf{P}\{g^*(T,B) \neq Y\} \\
&\quad = \mathbf{P}\{T+B<7-\log 2, T+B+E \geq 7\} \\
&\qquad + \mathbf{P}\{T+B \geq 7-\log 2, T+B+E<7\} \\
&\quad = \mathbf{E}\left\{e^{-(7-T-B)} I_{\{T+B<7-\log 2\}}\right\} \\
&\qquad + \mathbf{P}\left\{\left(1-e^{-(7-T-B)}\right) I_{\{7>T+B\geq 7-\log 2\}}\right\} \\
&\quad = \int_0^{7-\log 2} x e^{-x} e^{-(7-x)} dx + \int_{7-\log 2}^{7} x e^{-x}\left(1-e^{-(7-x)}\right) dx \\
&\qquad \text{(since the density of } T+B \text{ is } ue^{-u} \text{ on } [0,\infty)\text{)} \\
&\quad = e^{-7}\left(\frac{(7-\log 2)^2}{2} + 2(8-\log 2) - 8 - \frac{7^2}{2} + \frac{(7-\log 2)^2}{2}\right) \\
&\qquad \left(\text{as} \int_x^{\infty} u e^{-u} du = (1+x)e^{-x}\right) \\
&\quad = 0.0199611\ldots
\end{aligned}$$

如果我们仅知道 T 的值，则贝叶斯分类器仅允许使用 T 值。首先，我们有

$$\begin{aligned}
\mathbf{P}\{Y=1 \mid T\} &= \mathbf{P}\{E+B<7-T \mid T\} \\
&= \max\left(0, 1-(1+7-T)e^{-(7-T)}\right).
\end{aligned}$$

交叉发生在值为 1/2 时，则 $T = c \stackrel{\text{def}}{=} 5.321653009\ldots$，因此贝叶斯分类器如下给出

$$g^*(T) = \begin{cases} 1 & \text{if } T<c \\ 0 & \text{otherwise} \end{cases}$$

误差概率为

$$\begin{aligned}
&\mathbf{P}\{g^*(T) \neq Y\} \\
&\quad = \mathbf{P}\{T<c, T+B+E \geq 7\} + \mathbf{P}\{T \geq c, T+B+E<7\} \\
&\quad = \mathbf{E}\left\{(1+7-T)e^{-(7-T)} I_{\{T<c\}}\right\} \\
&\qquad + \mathbf{P}\left\{\left(1-(1+7-T)e^{-(7-T)}\right) I_{\{7>T\geq c\}}\right\} \\
&\quad = \int_0^c e^{-x}(1+7-x)e^{-(7-x)} dx + \int_c^7 e^{-x}\left(1-(1+7-x)e^{-(7-x)}\right) dx \\
&\quad = e^{-7}\left(\frac{8^2}{2} - \frac{(8-c)^2}{2} + e^{-(c-7)} - 1 - \frac{(8-c)^2}{2} + \frac{1}{2}\right) \\
&\quad = 0.02235309002\ldots.
\end{aligned}$$

贝叶斯误差有些许增加，但不会太多。最后，若我们不能知道所有三个变量 T，B 和 E 的值，我们能做的最好的是看哪一类最有可能出现。因此，我们计算

$$\mathbf{P}\{Y=0\} = \mathbf{P}\{T+B+E \geq 7\} = \left(1+7+7^2/2\right)e^{-7} = .02963616388\ldots$$

若令 $g \equiv 1$ 总成立，我们出错的概率为 0.02963616388…。

实际上，贝叶斯分类器是未知的，仅是因为 (X,Y) 的分布是未知的。考虑基于 (T,B) 的分类器。Rosenblatt 感知器（见第四章）基于数

据寻找最优的线性分类器。即，决策具有如下形式

$$g(T,B) = \begin{cases} 1 & \text{if } aT + bB < c \\ 0 & \text{otherwise} \end{cases}$$

基于数据选择 a，b 和 c。如果我们有大量数据可以利用，则有可能取到一个接近最优的线性分类器。如上面所见，贝叶斯分类器恰好是线性的。当然这纯粹是一个巧合。如果贝叶斯分类器不是线性的——例如如果我们有 $Y = I_{\{T+B^2+E<7\}}$——则即使是最好的感知器也是次优的，不管我们有多少对数据。如果我们使用 3-最近邻规则（第五章），渐进误差概率不会超过贝叶斯误差的 1.3155 倍，在我们的例子中大约是 0.02625882705 倍。上述例子也表明需要注意单个分量，估计多少个，哪几个分量对判别最有用。这一主题在特征提取一章中介绍（第 32 章）。

2.4 贝叶斯风险的其他公式

下述形式的贝叶斯误差更方便：

$$\begin{aligned} L^* &= \inf_{g:\mathcal{R}^d \to \{0,1\}} \mathbf{P}\{g(X) \neq Y\} \\ &= \mathbf{E}\{\min\{\eta(X), 1-\eta(X)\}\} \\ &= \frac{1}{2} - \frac{1}{2}\mathbf{E}\{|2\eta(X)-1|\} \end{aligned}$$

在特殊情况下，我们能得到其他有用的形式。例如，如果 X 具密度 f，则

$$\begin{aligned} L^* &= \int \min(\eta(x), 1-\eta(x)) f(x)\,dx \\ &= \int \min\left((1-p)f_0(x), p f_1(x)\right) dx \end{aligned}$$

其中 $p = P\{Y=1\}$，$f_i(x)$ 是给定 $Y=i$ 时 X 的密度。p 和 $1-p$ 称为类概率，f_0，f_1 为类条件密度。如果 f_0 和 f_1 不重叠，即 $\int f_0 f_1 = 0$，则明显 $L^* = 0$。另外假设 $p = 1/2$。则

类概率

$$\begin{aligned} L^* &= \frac{1}{2}\int \min\left(f_0(x), f_1(x)\right) dx \\ &= \frac{1}{2}\int f_1(x) - (f_1(x) - f_0(x))_+\,dx \\ &= \frac{1}{2} - \frac{1}{4}\int |f_1(x) - f_0(x)|\,dx \end{aligned}$$

这里 g_+ 代表函数 g 的正部。因此贝叶斯误差与类密度间的 L_n 距离直接相关。

2.5 插件决策

从 X 关于 Y 的最好猜测为贝叶斯决策

$$g^*(x) = \begin{cases} 0 & \text{if } \eta(x) \leq 1/2 \\ 1 & \text{otherwise} \end{cases} = \begin{cases} 0 & \text{if } \eta(x) \leq 1-\eta(x) \\ 1 & \text{otherwise.} \end{cases}$$

函数 η 一般是未知的。假设我们可以知道非负函数 $\tilde{\eta}(x), 1-\tilde{\eta}(x)$ 的值，它们分别是对 $\eta(x)$ 和 $1-\eta(x)$ 的近似。这种情况下，使用插件决策函数

$$g(x) = \begin{cases} 0 & \text{if } \tilde{\eta}(x) \le 1/2 \\ 1 & \text{otherwise} \end{cases}$$

近似贝叶斯决策是是相当自然的。下面的著名定理（见，Van Ryzin (1966), Wolverton and Wagner (1969a), Glick (1973), Csibi (1971), Gyorfi (1975), (1978), Devroye and Wagner (1976b), Devroye (1982b), and Devroye and Gyorfi (1985)）表明若 $\tilde{\eta}(x)$ 在 L_1 意义上接近于实际的后验概率，则决策 g 的误差概率接近于最优决策 g^*。

定理 2.2 *对于上述定义的插件决策 g 的误差概率，我们有*

$$\mathbf{P}\{g(X) \neq Y\} - L^* = 2\int_{\mathcal{R}^d} |\eta(x) - 1/2| I_{\{g(x)\neq g^*(x)\}} \mu(dx)$$

与

$$\mathbf{P}\{g(X) \neq Y\} - L^* \le 2\int_{\mathcal{R}^d} |\eta(x) - \tilde{\eta}(x)| \mu(dx) = 2\mathbf{E}|\eta(X) - \tilde{\eta}(X)|.$$

证. 若对某 $x \in \mathcal{R}^d$，$g(x) = g^*(x)$，则明显 g 和 g^* 的条件误差概率之差值为零：

$$\mathbf{P}\{g(X) \neq Y \mid X = x\} - \mathbf{P}\{g^*(X) \neq Y \mid X = x\} = 0$$

否则，若 $g(x) \neq g^*(x)$，则如在定理2.1中所见，差值可写成

$$\begin{aligned} &\mathbf{P}\{g(X) \neq Y \mid X = x\} - \mathbf{P}\{g^*(X) \neq Y \mid X = x\} \\ &\qquad = (2\eta(x) - 1)\left(I_{\{g^*(x)=1\}} - I_{\{g(x)=1\}}\right) \\ &\qquad = |2\eta(x) - 1| I_{\{g(x)\neq g^*(x)]} \end{aligned}$$

因此，

$$\begin{aligned} \mathbf{P}\{g(X) \neq Y\} - L^* &= \int_{\mathcal{R}^d} 2|\eta(x) - 1/2| I_{\{g(x)\neq g^*(x)]} \mu(dx) \\ &\le \int_{\mathcal{R}^d} 2|\eta(x) - \tilde{\eta}(x)| \mu(dx) \end{aligned}$$

因为 $g(x) \neq g^*(x)$ 蕴含着 $|\eta(x) - \tilde{\eta}(x)| \ge |\eta(x) - 1/2|$。 [证毕]

当分类器 $g(x)$ 能表成形式

$$g(x) = \begin{cases} 0 & \text{if } \tilde{\eta}_1(x) \le \tilde{\eta}_0(x) \\ 1 & \text{otherwise} \end{cases}$$

其中 $\tilde{\eta}_1(x)$，$\tilde{\eta}_0(x)$ 分别是 $\eta_1(x)$ 和 $\eta_0(x)$ 的某个近似。如果 $\tilde{\eta}_1(x) + \tilde{\eta}_0(x)$ 不必等于 1，则情况与在定理2.2中讨论的不同。但是一于定理2.2相似的不等式保持成立：

定理 2.3 *如上定义的决策误差概率有上界*

$$\mathbf{P}\{g(X) \neq Y\} - L^* \leq \int_{\mathcal{R}^d} |(1-\eta(x)) - \tilde{\eta}_0(x)|\, \mu(dx) + \int_{\mathcal{R}^d} |\eta(x) - \tilde{\eta}_1(x)|\, \mu(dx)$$

定理的证明留给读者（问题 2.9）。

备注 2.2 假设类条件密度 f_0，f_1 存在并由密度 $\tilde{f}_0(x)$ 和 $\tilde{f}_1(x)$ 近似。另外假设类概率 $p = \mathbf{P}\{Y=1\}$ 和 $1-p = \mathbf{P}\{Y=0\}$ 的近似分别为 $\tilde{p}_1$ 和 $\tilde{p}_0$。则插件决策函数的误差概率为

$$g(x) = \begin{cases} 0 & \text{if } \tilde{p}_1 \tilde{f}_1(x) \leq \tilde{p}_0 \tilde{f}_0(x) \\ 1 & \text{otherwise} \end{cases}$$

$$\mathbf{P}\{g(X) \neq Y\} - L^* \leq \int_{\mathcal{R}^d} \left|(1-p) f_0(x) - \tilde{p}_0 \tilde{f}_0(x)\right| dx + \int_{\mathcal{R}^d} \left|p f_1(x) - \tilde{p}_1 \tilde{f}_1(x)\right| dx$$

见问题 2.10.

2.6 贝叶斯误差与维数

在贝叶斯分类器里重要的 X 的分量是那些显式出现在 $\eta(X)$ 中的。实际上，因此，所有判别问题都是一维的，因为我们可以等价地替换 X 成 $\eta(X)$ 或通过任一 $\eta(X)$ 的严格单调增函数，例如 $\eta^7(X) + 5\eta^3(X) + \eta(X)$。不幸的是，$\eta$ 一般是未知的。在 2.3 节的示例中，我们有如下情况

$$\eta(T, B) = \max\left(0, 1 - e^{-(7-T-B)}\right)$$

另一情况下

$$\eta(T) = \max\left(0, 1 - (1+7-T)e^{-(7-T)}\right)$$

前者形式表明我们可以使所有决策基于 $T+B$。这意味着若我们无法个别地访问 T 和 B 的值，但可以知道 $T+B$ 的值，则我们也可以得到同样的结果！因为 η 是未知的，但是无关紧要的。

一般来说，贝叶斯风险增加，如果我们替换 X 成 $T(X)$ 对任意变换 T（见问题 2.1），因为这会破坏信息。另一方面，存在一种变换（如 $\eta(X)$）保持贝叶斯误差不变。对更多贝叶斯误差与维度的关系，参考第 32 章。

2.7 问题与习题

问题 2.1

令 $T : \mathcal{X} \to \mathcal{X}'$ 是任一可测函数。若 L_X^* 和 $L_{T(X)}^*$ 分别表示 (X, Y) 和 $(T(X), Y)$ 的贝叶斯误差概率，则证明

$$L_{T(X)}^* \geq L_X^*$$

（这表示 X 的变换破坏信息，因为贝叶斯风险增加了。）

问题 2.2

令 X' 与 (X,Y) 相独立，证明

$$L^*_{(X,X')} = L^*_X$$

问题 2.3

证明 $L^* \leq \min(p, 1-p)$，其中 $p, 1-p$ 是类概率。证明等式成立，仅当 X 和 Y 是独立的。给出一个分布，其中 X 与 Y 不独立，但 $L^* = \min(p, 1-p)$。

问题 2.4

纽曼-皮尔逊引理

再次考虑决策问题，有一个决策 g，我们设置两个误差概率，

$$L^{(0)}(g) = \mathbf{P}\{g(X) = 1 \mid Y = 0\} \text{ and } L^{(1)}(g) = \mathbf{P}\{g(X) = 0 \mid Y = 1\}$$

假设类条件密度 f_0, f_1 存在。对 $c > 0$，定义决策

$$g_c(x) = \begin{cases} 1 & \text{if } c f_1(x) > f_0(x) \\ 0 & \text{otherwise} \end{cases}$$

证明对任一决策 g，若 $L^{(0)}(g) \leq L^{(0)}(g_c)$，则 $L^{(1)}(g) \geq L^{(1)}(g_c)$。换言之，若 $L^{(0)}$ 被要求保持再某个水平之下，则最小化 $L^{(1)}$ 的决策类似 $g_c, \exists c$ 的形式。注意 g^* 正是如此。

问题 2.5

拒绝决策

有时在决策问题中，可以允许一个人说“我不知道”，如果这种情况不那么频繁发生的话。这些决策被称为有拒绝项的决策（见 Forney (1968), Chow (1970)）。正式的说，决策 $g(x)$ 可以取三个值：0, 1 和“拒绝（reject）”。有两个性能测量量：拒绝概率 $\mathbf{P}\{g(X) = \text{"reject"}\}$ 和错误概率 $\mathbf{P}\{g(X) \neq Y \mid g(X) \neq \text{"reject"}\}$。对 $0 < c < 1/2$，定义决策

$$g_c(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 + c \\ 0 & \text{if } \eta(x) \leq 1/2 - c \\ \text{"reject"} & \text{otherwise} \end{cases}$$

证明对任一决策 g，若

$$\mathbf{P}\{g(X) = \text{"reject"}\} \leq \mathbf{P}\{g_c(X) = \text{"reject"}\}$$

则

$$\mathbf{P}\{g(X) \neq Y \mid g(X) \neq \text{"reject"}\} \geq \mathbf{P}\{g_c(X) \neq Y \mid g_c(X) \neq \text{"reject"}\}$$

因此，为了保持拒绝概率在某个水平之下，决策具有 g_c 的形式是最优的（Gyorfi，Gyorfi 和 Vajda（1978））。

问题 2.6

考虑预测学生是否不及格基于变量 T, B，其中 $Y = I_{\{T+B+E<7\}}$ 且 E 的值不可知（见第 2.3 节）。

- 令 T, B 和 E 相互独立。仅通过改变 E 的分布，证明基于 (T, B) 分类的贝叶斯误差能任意接近 1/2。
- 令 T, B 相互独立且都服从指数分布。找到一个联合分布 $\mathcal{D}(T, B, E)$ 使得贝叶斯分类器不是一个线性分类器。
- 令 T, B 相互独立且都服从指数分布。找到一个联合分布 $\mathcal{D}(T, B, E)$ 使得贝叶斯分类器为

$$g^*(T, B) = \begin{cases} 1 & \text{if } T^2 + B^2 < 10 \\ 0 & \text{otherwise} \end{cases}$$

- 找到基于 (T, B, E) 分类的贝叶斯分类器和贝叶斯误差，如果 (T, B, E)（Y 如上给出）是 $[0, 4]^3$ 上的均匀分布。

问题 2.7

假设 T, B, E 是在 $[0, 4]$ 上独立的均匀随机变量，如 2.3 节中所介绍的一样。令 $Y = 1$（或 0）表示是否一个学生通过（或没通过）一门课程。假设 $Y = 1$ 当且仅当 $TBE < 8$。

- 找到贝叶斯决策，如果所有变量都是未知的，或如果仅 T 是可知的，或如果 T, B 都可知。
- 确定所有三种情况下的贝叶斯误差。
- 确定仅基于 T, B 最优的线性分类器。

问题 2.8

令 $\eta', \eta'' : \mathcal{R}^d \to [0, 1]$ 为任意可测函数，定义相应的决策为 $g'(x) = I_{\{\eta'(x)>1/2\}}$ 和 $g''(x) = I_{\{\eta''(x)>1/2\}}$。证明

$$|L(g') - L(g'')| \le \mathbf{P}\{g'(X) \neq g''(X)\}$$

和

$$|L(g') - L(g'')| \le \mathbf{E}\left\{|2\eta(X) - 1| I_{\{g'(X)\neq g''(X)\}}\right\}$$

问题 2.9

证明定理2.3。

问题 2.10

假设类条件密度 f_0, f_1 存在并由密度 $\tilde{f}_0, \tilde{f}_1$ 分别近似。另设类概率 $p = \mathbf{P}\{Y = 1\}$ 和 $1 - p = \mathbf{P}\{Y = 0\}$ 分别由 $\tilde{p}_1, \tilde{p}_0$ 近似。证明对插件决策函数的误差概率

$$g(x) = \begin{cases} 0 & \text{if } \tilde{p}_1 \tilde{f}_1(x) \le \tilde{p}_0 \tilde{f}_0(x) \\ 1 & \text{otherwise} \end{cases}$$

我们有

$$\mathbf{P}\{g(X) \neq Y\} - L^* \leq \int_{\mathcal{R}^d} \left|pf_1(x) - \tilde{p}_1\tilde{f}_1(x)\right| dx + \int_{\mathcal{R}^d} \left|(1-p)f_0(x) - \tilde{p}_0\tilde{f}_0(x)\right| dx$$

问题 2.11

使用问题 2.10 的概念，证明如果对序列 $\tilde{f}_{m,n}(x)$ 和 $\tilde{p}_{m,n}(x)$ $(m = 0, 1)$ 有

$$\lim_{n\to\infty} \int_{\mathcal{R}^d} \left|pf_1(x) - \tilde{p}_{1,n}\tilde{f}_{1,n}(x)\right|^2 dx + \int_{\mathcal{R}^d} \left|(1-p)f_0(x) - \tilde{p}_{0,n}\tilde{f}_{0,n}(x)\right|^2 dx = 0$$

则对相应的插件决策 $\lim_{n\to\infty} \mathbf{P}\{g_n(X) \neq Y\} = L^*$（Wolverton 和 Wagner（l969a））。

[提示]：根据问题 2.10，可以证明如果我们给定一个密度函数的确定序列 $f, f_1, f_2, f_3, \ldots$，则

$$\lim_{n\to\infty} \int (f_n(x) - f(x))^2 dx = 0$$

蕴含着

$$\lim_{n\to\infty} \int |f_n(x) - f(x)| dx = 0$$

（函数 f 称为密度函数，如果它是非负的而且 $\int f(x)dx = 1$）。要看到这一点，请注意

$$\int |f_n(x) - f(x)| dx = 2\int (f_n(x) - f(x))_+ dx = 2\sum_i \int_{A_i} (f_n(x) - f(x))_+ dx$$

其中 $A_1, A_2, \ldots$ 是 $\mathcal{R}^d$ 的划分单位立方体，f_+ 表示函数 f 的正部。关键的地方是无穷和的每一项收敛到 0 意味着整个积分收敛（利用控制收敛定理），因为 $\int (f_n(x) - f(x))_+ dx \leq \int f_n(x)dx = 1$。使用 Cauchy-Schwarz 不等式处理右边。

问题 2.12

通过 $J(f) = \mathbf{E}\{|f(X) - Y|\}$ 定义函数 $f : \mathcal{R}^d \to \mathcal{R}$ 的 L_1 误差。证明最小化 $J(f)$ 的函数正好是贝叶斯规则 g^*，即，$J^* = \inf_f J(f) = J(g^*)$。因此，$J^* = L^*$。定义决策

$$g(x) = \begin{cases} 0 & \text{if } f(x) \leq 1/2 \\ 1 & \text{otherwise} \end{cases}$$

证明误差概率 $L(g) = \mathbf{P}\{g(X) \neq Y\}$ 满足不等式

$$L(g) - L^* \leq J(f) - J^*.$$

不等式与替代距离度量 3

3.1 度量判别信息

在二类判别问题中，最优的规则具有（贝叶斯）误差概率

$$L^* = \mathbf{E}\{\min(\eta(X), 1-\eta(X))\}.$$

该量能度量判别问题的难度。它也作为模式识别中 (X,Y) 分布的质量度量。换言之，若 ψ_1 和 ψ_2 是某多到一映射，L^* 能被用来比较基于 $(\psi_1(X),Y)$ 和 $(\psi_2(X),Y)$ 的判别。当 ψ_1 通过取前 d_1 个坐标把 $\mathcal{R}^d$ 投影到 $\mathcal{R}^{d_1}$，而 ψ_2 取后 d_2 个坐标。相应的贝叶斯误差将帮助我们决定哪一个投影更优。在这种意义下，L^* 在特征提取中是基础量。

多年来，其他量也被提出用以度量隐藏在 (X,Y) 分布中判别能力。在一些设置下这些量可能是有用的。例如，在理论研究或某些证明中，L^* 和 (X,Y) 分布之间的关系能通过某些连接 L^* 和其他分布泛函的不等式从而变得清晰。我们都明白矩和方差的含义，但这些简单的泛函与 L^* 的关系是什么？可能我们将学到是什么使 L^* 变小。在特征选择中，一些涉及 L^* 的显式不等式仅提供数值信息，允许我们对哪种特征实践中更优作出某种判断。简言之，我们获得对模式识别具多种用处的 L^* 的更多信息。

在下面几节中，我们避免对 (X,Y) 的分布施加任何条件。

3.2 Kolmogorov 变分距离

受分布间全变分距离 的启发，Kolmogorov 变分距离

问：已解决。

$$\begin{aligned}\delta_{\mathrm{KO}} &= \frac{1}{2}\mathbf{E}\{|\mathbf{P}\{Y=1 \mid X\} - \mathbf{P}\{Y=0 \mid X\}|\} \\ &= \frac{1}{2}\mathbf{E}\{|2\eta(X)-1|\}\end{aligned}$$

捕获两个类间的距离。我们不需要任何特别的东西来处理 δ_{KO}

$$\begin{aligned}L^* &= \mathbf{E}\left\{\frac{1}{2} - \frac{1}{2}|2\eta(X)-1|\right\} \\ &= \frac{1}{2} - \frac{1}{2}\mathbf{E}\{|2\eta(X)-1|\} \\ &= \frac{1}{2} - \delta_{\mathrm{K0}}.\end{aligned}$$

3.3 最近邻误差

最近邻规则的渐进误差为

$$L_{\mathrm{NN}} = \mathbf{E}\{2\eta(X)(1-\eta(X))\}$$

（见第五章）。明显，当 $2\max(\eta, 1-\eta) \geq 1$ 时有 $2\eta(1-\eta) \geq \min(\eta, 1-\eta)$。另外，使用概念 $A = \min(\eta(X), 1-\eta(X))$，我们有

$$\begin{aligned} L^* \leq L_{\mathrm{NN}} &= 2\mathbf{E}\{A(1-A)\} \\ &\leq 2\mathbf{E}\{A\} \cdot \mathbf{E}\{1-A\} \\ &\quad \text{(通过定理 A.19 第二个分配不等式)} \\ &= 2L^*(1-L^*) \leq 2L^*, \end{aligned}$$

Cover-Hart 不等式

这是著名的 Cover-Hart 不等式（1967）。L_{NN} 给我们提供了相当多的关于 L^* 的更多信息。

度量 L_{NN} 能以其他方式表示：Devijver 和 Kittler（1982，263 页）和 Vajda（1968）称之为量子熵，Mathai 和 Rathie（1975）称为调和平均系数。

3.4 Bhattacharyya 亲和力

Bhattacharyya 亲和力度量（Bhattacharyya，1946）为——$\log(\rho)$，其中

$$\rho = \mathbf{E}\{\sqrt{\eta(X)(1-\eta(X))}\}$$

被称为 Matushita 误差。它没有作为任意标准判别规则的极限自然出现（但是，见问题 6.11）。ρ 由 Matushita（1956）建议作为模式识别的距离度量。它在数理统计中也以其他形式出现——可见于 Hellinger 距离的文献中（Le Cam（1970），Beran（1977））。

显然，$\rho = 0$ 当且仅当 $\eta(X) \in \{0,1\}$ 概率 1 成立。即 $L^* = 0$。另外，ρ 取得最大值当且仅当 $\eta(X) = 1/2$ 概率 1 成立。ρ 和 L^* 的关系不是线性的。我们将证明对所有分布，如果 L_{NN} 能用来近似 L^*，则它比 ρ 更有用。

定理 3.1 *对所有分布，有*

$$\begin{aligned} \frac{1}{2} - \frac{1}{2}\sqrt{1-4\rho^2} &\leq \frac{1}{2} - \frac{1}{2}\sqrt{1-2L_{\mathrm{NN}}} \\ &\leq L^* \\ &\leq L_{\mathrm{NN}} \\ &\leq \rho. \end{aligned}$$

证. 首先,

$$\begin{aligned}\rho^2 &= \mathbf{E}^2\{\sqrt{\eta(X)(1-\eta(X))}\}\\ &\le \mathbf{E}\{\eta(X)(1-\eta(X))\} \quad \text{(使用 Jensen 不等式)}\\ &= \frac{L_{\mathrm{NN}}}{2}\\ &\le L^*(1-L^*) \quad \text{(使用 Cover-Hart 不等式 (3.1)).}\end{aligned}$$

第二，由于 $\sqrt{\eta(1-\eta)} \ge 2\eta(1-\eta)$，对任$\eta \in [0,1]$，我们能得到 $\rho \ge L_{\mathrm{NN}} \ge L^*$。最后，通过使用 Cover-Hart 不等式,

$$\sqrt{1-2L_{\mathrm{NN}}} \ge \sqrt{1-4L^*(1-L^*)} = 1-2L^*.$$

将所有这些式子放一起可建立不等式链。 [证毕]

不等式 $L_{\mathrm{NN}} \le \rho$ 来自 Ito（1972)。$L_{\mathrm{NN}} \ge 2\rho^2$ 来自于 Horibe（1970)。$1/2 - \sqrt{1-4\rho^2}/2 \le L^* \le \rho$ 能在 Kailath（1967）中找到。最后一个不等式的右边项首先出现在 Hudimoto（1957)。所有这些不等式是紧的（见问题 3.2)。像 L_{NN} 和 ρ 这些量的吸引力在于其涉及到 η 的多项式，然而 $L^* = \mathbf{E}\{\min(\eta(X), 1-\eta(X))\}$ 是非多项式的。对某些判别问题，X 的分布的参数是已知的，我们可以显式计算出 L_{NN} 和 ρ，它们作为分布参数的函数。通过这些不等式，可以用以获得插入类型（plug-in type）参数判别规则的精度保证。（见第 16 章）

为了完整性，我们提及 Bhattacharyya 的亲和力度量，率先由 Chernoff（1952）提出:

$$\delta_C = -\log\left(\mathbf{E}\left\{\eta^\alpha(X)(1-\eta(X))^{1-\alpha}\right\}\right),$$

其中 $\alpha \in (0,1)$ 是固定的。对 $\alpha = 1/2$, 则 $\delta_C = -\log\rho$。但经由取 $\alpha \ne 1/2$ 引出的渐进性并没有实际的解释。

3.5 熵

离散概率分布 $(p_1, p_2, \ldots)$ 的熵定义为 熵

$$\mathscr{H} = \mathscr{H}(p_1, p_2, \ldots) = -\sum_{i=1}^{\infty} p_i \log p_i,$$

其中，令 $0\log 0 = 0$（Shannon（1948))。熵是信息论中的中心量（见 Cover 和 Thomas（1991)），在计算机科学，数理统计和物理学中具有相当多的应用。熵的主要属性总结如下:

1. $\mathscr{H} \ge 0$， 等号成立当且仅当 $p_i = 1, \exists i$。[证：由 $\log p_i \le 0, \forall i$，等号成立当且仅当 $p_i = 1, \exists i$。因此熵是最小对于退化分布，即“散布”程度最小的分布。] 译注：因为 $p_i = 1, \exists i$ 而 $p_j = 0, \forall j \ne i$，因此分布均“凝聚”在 p_i 上。
2. $\mathscr{H}(p_1, \ldots, p_k) \le \log k$。其等号成立当且仅当 $p_1 = p_2 = \cdots = p_k = 1/k$。换言之，当分布被最大程度“抹去”时熵取最大值。[证：由不等式 $\log x \le x - 1, \forall x > 0$ 得

$$\mathscr{H}(p_1, \ldots, p_k) - \log k = \sum_{i=1}^{k} p_i \log\left(\frac{1}{kp_i}\right) \le 0.]$$

3. 对伯努利分布 $(p, 1-p)$，二进制熵 $\mathscr{H}(p,1-p) = p\log p - (1-p)\log(1-p)$ 在 p 上是凹的。

假设 X 是离散随机变量，对某个集合 A，可通过询问“是否 $X \in A$?”来猜测其值。令 N 为猜出 X 确切值所需问题数目的最小期望。众所周知

$$\frac{\mathscr{H}}{\log 2} \le N < \frac{\mathscr{H}}{\log 2} + 1$$

（见 Cover 和 Thomes（1991））。因此 $\mathscr{H}$ 不仅测量 X 的质量如何散开，而且也给某些算法提供一个具体上下界。在上面的简单例子中，$\mathscr{H}$ 实际上与最优算法的计算时间是成比例的。

本书对信息论本身（per se) 不感兴趣，但它在模式识别中相当有用。回到上述讨论，若固定 $X = x$，则 Y 是伯努利的（$\eta(x)$）。因此，给定 $X = x$ 时 Y 的条件熵 为

条件熵

$$\mathscr{H}(\eta(x), 1-\eta(x)) = -\eta(x)\log\eta(x) - (1-\eta(x))\log(1-\eta(x)).$$

它测量给定 $X = x$ 时 Y 的不确定程度或混乱程度。其值为 0（当 $\eta(x) \in \{0,1\}$）或为 $\log 2$（当 $\eta(x) = 1/2$），故对 $\eta(x)$ 是凹的。我们定义期望条件熵为

期望条件熵

$$\begin{aligned}\mathscr{E} &= \mathbf{E}\{\mathscr{H}(\eta(X), 1-\eta(X))\} \\ &= -\mathbf{E}\{\eta(X)\log\eta(X) + (1-\eta(X))\log(1-\eta(X))\}.\end{aligned}$$

为方便起见，我们称 $\mathscr{E}$ 为熵。$\mathscr{E} = 0$ 当且仅当 $\eta(X) \in \{0,1\}$ 以概率 1 成立。因此 $\mathscr{E}$ 和 L^* 是彼此相关的。

Fano 不等式

定理 3.2 *1.* $\mathscr{E} \le \mathscr{H}(L^*, 1-L^*) = -L^*\log L^* - (1-L^*)\log(1-L^*)$. *（Fano（1952），见 Cover 和 Thomas（1991，39 页））。*

2. $\mathscr{E} \ge -\log(1-L_{\mathrm{NN}}) \ge -\log(1-L^*)$.

3. $\mathscr{E} \le \log 2 - \frac{1}{2}(1-2L_{\mathrm{NN}}) \le \log 2 - \frac{1}{2}(1-2L^*)^2$.

证. 第一部分：令 $A = \min(\eta(X), 1-\eta(X))$，则

$$\begin{aligned}\mathscr{E} &= \mathbf{E}\{\mathscr{H}(A, 1-A)\} \\ &\le \mathscr{H}(\mathbf{E}A, 1-\mathbf{E}A) \text{ (由}\mathscr{H}\text{ 是凸的，使用 Jensen 不等式)} \\ &= \mathscr{H}(L^*, 1-L^*).\end{aligned}$$

第二部分：

$$\begin{aligned}\mathscr{E} &= -\mathbf{E}\{A\log A + (1-A)\log(1-A)\} \\ &\ge -\mathbf{E}\left\{\log\left(A^2 + (1-A)^2\right)\right\} \text{ (使用 Jensen 不等式)} \\ &= -\mathbf{E}\{\log(1-2A(1-A))\} \\ &\ge -\log(1-\mathbf{E}\{2A(1-A)\}) \text{ (使用 Jensen 不等式)} \\ &= -\log(1-L_{NN}) \\ &\ge -\log(1-L^*).\end{aligned}$$

第三部分：通过自变量 A 函数 $(A, 1-A)$ 的凹性及泰勒级数展开，

$$\mathscr{H}(A, 1-A) \le \log 2 - \frac{1}{2}(2A-1)^2.$$

因此，由第一部分，

$$\begin{aligned}
\mathscr{C} &\le \log 2 - \mathbf{E}\left\{\frac{1}{2}(2A-1)^2\right\} \\
&= \log 2 - \frac{1}{2}(1 - 2L_{\mathrm{NN}}) \\
&\le \log 2 - \frac{1}{2} + 2L^*(1-L^*) \text{ (使用 Cover-Hart 不等式 (3.1))} \\
&= \log 2 - \frac{1}{2}(1-2L^*)^2 .
\end{aligned}$$

[证毕]

备注 3.1 $\mathscr{C}$ 和 L^* 接近单调的关系有大有用处。我们提醒读者，在接近原点的地方，L^* 关于 $\mathscr{C}$ 可能会线性递减，但也可以比 $\mathscr{C}^{209}$ 递减得还快。L^* 和 L_{NN}——线性——（或 L^* 和 ρ——介于线性与二次之间——）之间的关系变化不会如此大。

3.6 Jeffrey 散度

Jeffrey 散度是 Kullback-Leibler 散度（1951）

$$\delta_{\mathrm{KL}} = \mathbf{E}\left\{\eta(X)\log\frac{\eta(X)}{1-\eta(X)}\right\}.$$

的对称形式

$$\mathcal{J} = \mathbf{E}\left\{(2\eta(X)-1)\log\frac{\eta(X)}{1-\eta(X)}\right\}.$$

为理解 $\mathcal{J}$，注意函数 $(2\eta-1)\log\frac{\eta}{1-\eta}$ 关于 $\eta = 1/2$ 对称，且是凸的，在 $\eta = 1/2$ 处具有最小值（0）。当 $\eta \downarrow 0, \eta \uparrow 1$（$\eta \to 0, \eta \to 1$），函数趋于无界。因此 $\mathcal{J} = \infty$，若$\mathbf{P}\{\eta(X) \in \{0,1\}\} > 0$。因此在判别中使用 Jeffrey 散度时有必要加以限制。对 $\mathcal{J}$ 的进一步推广可参考 Renyi（1961），Burbea and Rao（1982），Taneja（1983；1987），and Burbea（1984）。因此无法使用 L_{NN} 和/或 L^* 的函数给出 $\mathcal{J}$ 的上界。但是下界可以获得。因为 $x\log((1+x)/(1-x))$ 对 x 是凸的，且

译注：注意 $\{0,1\}$ 是集合而不是区间。

$$(2\eta-1)\log\frac{\eta}{1-\eta} = |2\eta-1|\log\left(\frac{1+|2\eta-1|}{1-|2\eta-1|}\right),$$

通过 Jensen 不等式，

$$\begin{aligned}
\mathcal{J} &\ge \mathbf{E}\{|2\eta(X)-1|\}\log\left(\frac{1+\mathbf{E}\{|2\eta(X)-1|\}}{1-\mathbf{E}\{|2\eta(X)-1|\}}\right) \\
&= (1-2L^*)\log\left(\frac{1+(1-2L^*)}{1-(1-2L^*)}\right) \\
&= (1-2L^*)\log\left(\frac{1-L^*}{L^*}\right) \\
&\ge 2(1-2L^*)^2 .
\end{aligned}$$

该下界是紧的（当 $\eta(x)$ 在整个空间中取常数时等号成立）。另外，对固定的 L^*，任意在下界之上 $\mathcal{J}$ 值都可能对应某个 (X,Y) 分布。根据 $\mathcal{J}$ 的

定义，$\mathcal{J}=0$ 当且仅当 $\eta\equiv 1/2$ 概率 1 成立（或 $L^*=1/2$）。

相关的另一个下界由 Toussaint（1974b）得到：

$$\mathcal{J}\geq\sqrt{1-2L_{\mathrm{NN}}}\log\left(\frac{1+\sqrt{1-2L_{\mathrm{NN}}}}{1-\sqrt{1-2L_{\mathrm{NN}}}}\right)\geq 2\left(1-2L_{\mathrm{NN}}\right).$$

该下界严格优于前面给出的 L^* 界。见问题 3.7。

3.7 F-误差

F-误差

目前为止的误差度量都与 $\eta(X)=\mathbf{P}\{Y=1\mid X\}$ 的凹函数的期望值相关。一般来说，如果 F 是 $[0,1]$ 上的凹函数，我们定义相对于 (X,Y) 的 F-误差如下

$$d_F(X,Y)=\mathbf{E}\{F(\eta(X))\}.$$

举几个 F-误差的例子：

(a) 贝叶斯误差 L^*：$F(x)=\min(x,1-x)$,
(b) 渐进最近邻误差 L_{NN}：$F(x)=2x(1-x)$,
(c) Matushita 误差 ρ：$F(x)=\sqrt{x(1-x)}$,
(d) 期望条件熵 $\mathcal{C}$：$F(x)=-x\log x-(1-x)\log(1-x)$,
(e) 负 Jeffrey 散度 $-\mathcal{J}$：$F(x)=-(2x-1)\log\frac{x}{1-x}$。

Hashlamoun，Varshney 和 Samarasooriya（1994）指出若 $F(x)\geq\min(x,1-x),\forall x\in[0,1]$, 则相应的 F-误差是贝叶斯误差的上界。$F(x)$ 越接近 $\min(x,1-x)$，该上界越紧。例如，$F(x)=(1/2)\sin(\pi x)\leq 2x(1-x)$ 推出比 L_{NN} 更紧的上界。所有这些误差具有一个共同属性：若 X 变换成其他任意函数，误差增加。

定理 3.3 *令 $t:\mathcal{R}^d\to\mathcal{R}^k$ 是任一可测函数。则对任 (X,Y) 分布，*

$$d_F(X,Y)\leq d_F(t(X),Y).$$

证. 令 $\eta_t:\mathcal{R}^k\to[0,1]$ 由$\eta_t(x)=\mathbf{P}\{Y=1\mid t(X)=x\}$，则有

$$\eta_t(t(X))=\mathbf{E}\{\eta(X)\mid t(X)\}.$$

因此，

$$\begin{aligned}d_F(t(X),Y)&=\mathbf{E}\left\{F\left(\eta_t(t(X))\right)\right\}\\&=\mathbf{E}\{F(\mathbf{E}\{\eta(X)\mid t(X)\})\}\\&\geq\mathbf{E}\{\mathbf{E}\{F(\eta(X))\mid t(X)\}\}\quad\text{(Jensen 不等式)}\\&=\mathbf{E}\{F(\eta(X))\}=d_F(X,Y).\end{aligned}$$

[证毕]

备注 3.2 从证明中我们还能看到 F-误差保持不变，如果变换 t 是可逆的。定理3.3证明了 F-误差有点类似贝叶斯误差——当信息丢失时（替换 X 成 $t(X)$），F-误差增加。

3.8 Mahalanobis 距离

假设两个均值不同的条件分布具有相同的协方差矩阵，则这两个分布可能具有良好的可分离性而使 L^* 值很小。[1]两个随机向量 X_0和X_1[2]的视距（visual distance）度量被称为 Mahalanobis 距离（Mahalanobis，（1936））：

$$\Delta = \sqrt{(m_1 - m_0)^T \Sigma^{-1} (m_1 - m_0)}.$$

其中 $m_1 = \mathbf{E}X_1, m_0 = \mathbf{E}X_0$ 是均值，$\Sigma_1 = \mathbf{E}\left\{(X_1 - m_1)(X_1 - m_1)^T\right\}$ 和 $\Sigma_0 = \mathbf{E}\left\{(X_0 - m_0)(X_0 - m_0)^T\right\}$ 是协方差矩阵，$\Sigma = p\Sigma_1 + (1-p)\Sigma_0$，$(\cdot)^T$ 是向量的转置，$p = 1 - p$ 为混合参数。若 $\Sigma_1 = \Sigma_0 = \sigma^2 I$，其中 I 为单位矩阵，则

$$\Delta = \frac{\|m_1 - m_0\|}{\sigma}$$

为均值之间距离的缩放。若 $\Sigma_1 = \sigma_1^2 I, \Sigma_0 = \sigma_0^2 I$，则当 p 从 1 到 0 改变时，

$$\Delta = \frac{\|m_1 - m_0\|}{\sqrt{p\sigma_1^2 + (1-p)\sigma_0^2}}$$

在 $\|m_1 - m_0\|/\sigma_1$ and $\|m_1 - m_0\|/\sigma_0$ 之间变化。假设有一个判别问题，给定 $Y = 1$，X 是作为 X_1 的分布；给定 $Y = 0$，X 是作为 X_0 的分布。令 $p = \mathbf{P}\{Y = 1\}$，$1 - p$ 是类概率。则有趣的是，Δ 在一般意义下与贝叶斯误差相关。若两个类条件概率之间的 Mahalanobis 距离很大，则 L^* 很小。

定理 3.4 *对任 (X, Y) 分布，且 $\mathbf{E}\left\{\|X\|^2\right\} < \infty$，我们有*

$$L^* \leq L_{\mathrm{NN}} \leq \frac{2p(1-p)}{1 + p(1-p)\Delta^2}.$$

Devijver 和 Kittler（1982，166 页）

备注 3.3 对一个具均值 m 和协方差矩阵 Σ 的分布，点 $x \in \mathcal{R}^d$ 到 m 的 Mahalanobis 距离为

$$\sqrt{(x-m)^T \Sigma^{-1} (x-m)}.$$

在一维情况，可简单解释成以标准差为单位到均值的距离。[3] 在判别中的 Mahalanobis 距离基于这样一种直观概念：我们应基于最小标准差单位内的类进行分类。至少对看起来好看的球状云分布，这一建议可能是有意义的。

证. 首先假设 $d = 1$，即 X 是一个实数值。令 u 和 c 为实数，考虑 $\mathbf{E}\left\{(u(X-c) - (2\eta(X) - 1))^2\right\}$。我们将证明若选择 u, c 最小化该量，则它满足

$$0 \leq \mathbf{E}\left\{(u(X-c) - (2\eta(X) - 1))^2\right\} = 2\left(\frac{2p(1-p)}{1 + p(1-p)\Delta^2} - L_{\mathrm{NN}}\right), \quad (3.1)$$

这证明了 $d = 1$ 时的情况。为说明这一点，$\mathbf{E}\left\{(u(X-c) - (2\eta(X) - 1))^2\right\}$

1: 许有问题：Two conditional distributions with about the same covariance matrices and means that are far away from each other are probably so well separated that L^* is small.

2: 译注：X_i 是列向量。

译注：令 $P\{X|Y=1\} = P\{X_1\}$

译注：$\Delta = (EX_1 - EX_0)^T \Sigma^{-1} (EX_1 - EX_0)^{-1}$

3: 许有问题：In one dimension, this is simply interpreted as distance from the mean as measured in units of standard deviation.

在 $c = \mathbf{E}X - \mathbf{E}\{2\eta(X)-1\}/u$ 时取得最小值。则

$$\begin{aligned}&\mathbf{E}\left\{(u(X-c)-(2\eta(X)-1))^2\right\}\\&\quad = \operatorname{Var}\{2\eta(X)-1\} + u^2\operatorname{Var}\{X\} - 2u\operatorname{Cov}\{X, 2\eta(X)-1\},\end{aligned}$$

其中 $\operatorname{Cov}\{X,Z\} = \mathbf{E}\{(X-\mathbf{E}X)(Z-\mathbf{E}Z)\}$ 当 $u = \operatorname{Cov}\{X,2\eta(X)-1\}/\operatorname{Var}\{X\}$ 时取最小值。直接计算可证得式（3.1）成立。为将不等式（3.1）扩展至多维问题，将其应用到一维决策问题 (Z,Y)，其中 $Z = X^T\Sigma^{-1}(m_1 - m_0)$。则通过定理3.3，

$$L_{\mathrm{NN}}(X,Y) \le L_{\mathrm{NN}}(Z,Y),$$

其中 $L_{\mathrm{NN}}(X,Y)$ 是相对于 (X,Y) 的最近邻误差。 [证毕]

当 X_1 和 X_0 是协方差矩阵相同正态分布，我们有

Matushita（1973）；见问题 3.11

定理 3.5 *当 X_1 和 X_0 是多元正态随机变量且 $\Sigma_1 = \Sigma_0 = \Sigma$，则*

$$\rho = \mathbf{E}\{\sqrt{\eta(X)(1-\eta(X))}\} = \sqrt{p(1-p)}e^{-\Delta^2/8}.$$

若类条件密度 f_1, f_0 分别为 $(x-m_1)^T\Sigma_1^{-1}(x-m_1)$ 和 $(x-m_0)^T\Sigma_0^{-1}(x-m_0)$ 函数，则 Δ 与 L^* 的连接相当紧密（Mitchell 和 Krzanowski（1985））, 但这些分布是例外而不是常规。一般来说，当 Δ 值很小时，不可能推出 L^* 也很小或者相反。（见问题 3.12）。

3.9 f-散度

我们已定义误差度量为 $\eta(X)$ 的凹函数的期望值。这使得连接这些度量到贝叶斯误差 L^* 和其他误差概率概率更加容易。本节中我们简单的介绍更多在传统统计理论中概率测度间的距离的联系。这些距离度量的通用概念称为f-散度，由 Csiszár（1967）提出。Vajda（1989）总结了相关理论。如果已知类概率 $p, 1-p$ 和 X 的给定 $\{Y=0\}$ 和$\{Y=1\}$ 时的条件分布 μ_0, μ_1，

f-散度

$$\mu_i(A) = \mathbf{P}\{X \in A \mid Y = i\} \quad i = 0, 1.$$

则可以计算更早定义的 F-误差。对固定的类概率，F-误差很小若这两个条件分布离得“很远”。在下面我们将定义一个量表示这个距离。令 $f : [0,\infty) \to \mathcal{R} \cup \{-\infty, \infty\}$ 为凸函数且 $f(1) = 0$。两个 $\mathcal{R}^d$ 上概率测度 μ 和 υ 的 f-散度定义为

$$D_f(\mu,\nu) = \sup_{\mathcal{A}=\{A_j\}} \sum_j \nu(A_j) f\left(\frac{\mu(A_j)}{\nu(A_j)}\right),$$

其中上确界取 $\mathcal{R}^d$ 的所有有限可测部分 $\mathcal{A}$。若 λ 是控制 μ 和 υ 的度量——即 μ 和 υ 是关于 λ 绝对连续的——且 $p = d\mu/d\lambda$ 和$q = d\nu/d\lambda$ 是相应的密度，则 f-散度能写成

$$D_f(\mu,\nu) = \int q(x) f\left(\frac{p(x)}{q(x)}\right)\lambda(dx).$$

显然，这个量与 λ 值的选取无关。例如，我们取 $\lambda = \mu + \nu$。若 μ 和 υ 关于 Lebesgue 测度是绝对连续的，则 λ 可被选为 Lebesgue 测度 。通过使用 Jensen 不等式，可得 $D_f(\mu,\nu) \ge 0$ 和$D_f(\mu,\mu) = 0$。

译注：不理解。

f-散度一个重要的例子是全变分，或通过令 $f(x) = |x-1|$ 得到的变分距离，推出

$$V(\mu, \nu) = \sup_{\mathcal{A}=\{A_j\}} \sum_j \left|\mu\left(A_j\right) - \nu\left(A_j\right)\right|.$$

对这个散度，这两个定义的等价性由 Scheffé 定理给出（见问题 12.13）。

定理 3.6 Scheffé（1947）

$$V(\mu, \nu) = 2 \sup_A |\mu(A) - \nu(A)| = \int |p(x) - q(x)| \lambda(dx),$$

其中上确界取遍 $\mathcal{R}^d$ 的所有 Borel 子集。

另外一个重要例子为 Hellinger 距离，通过 $f(x) = (1-x)^2$ 给出:

$$\begin{aligned} H_2(\mu, \nu) &= \sup_{\mathcal{A}=\{A_j\}} 2\left(1 - \sum_j \sqrt{\mu\left(A_j\right) \nu\left(A_j\right)}\right) \\ &= 2\left(1 - \int \sqrt{p(x)q(x)}\lambda(dx)\right). \end{aligned}$$

量 $I_2(\mu, \nu) = \int \sqrt{p(x)q(x)}\lambda(dx)$ 常称为 Hellinger 积分。我们在这方面会提到两个有用的不等式。为简便起见，仅叙述其离散形式。（积分形式是类似的，见问题 3.21）。

引理 3.1 *对正数序列 a_i, b_i，两序列的和均为 1，则* LeCam（1973）

$$\sum_i \min\left(a_i, b_i\right) \geq \frac{1}{2}\left(\sum_i \sqrt{a_i b_i}\right)^2.$$

证. 使用 Cauchy-Schwarz 不等式，得

$$\sum_{i:a_i<b_i} \sqrt{a_i b_i} \leq \sqrt{\sum_{i:a_i<b_i} a_i} \sqrt{\sum_{i:a_i<b_i} b_i} \leq \sqrt{\sum_{i:a_i<b_i} a_i}.$$

该式与 $(x+y)^2 \leq 2x^2 + 2y^2$ 一起，加上对称性，可得

$$\begin{aligned} \left(\sum_i \sqrt{a_i b_i}\right)^2 &= \left(\sum_{i:a_i<b_i} \sqrt{a_i b_i} + \sum_{i:a_i \geq b_i} \sqrt{a_i b_i}\right)^2 \\ &\leq 2\left(\sum_{i:a_i<b_i} \sqrt{a_i b_i}\right)^2 + 2\left(\sum_{i:a_i \geq b_i} \sqrt{a_i b_i}\right)^2 \\ &\leq 2\left(\sum_{i:a_i<b_i} a_i + \sum_{i:a_i \geq b_i} b_i\right) \\ &= 2\sum_i \min\left(a_i, b_i\right). \end{aligned}$$

[证毕]

Devroye 和 Györfi（1985，225 页）

引理 3.2 *令 $a_1,\ldots,a_k,b_1,\ldots b_k$ 为非负值，使得 $\sum_{i=1}^k a_i=\sum_{i=1}^k b_i=1$。则*

$$\sum_{i=1}^k \sqrt{a_i b_i} \le 1-\frac{\left(\sum_{i=1}^k |a_i-b_i|\right)^2}{8}.$$

证.

$$\begin{aligned}\left(\sum_{i=1}^k |a_i-b_i|\right)^2 &\le \left(\sum_{i=1}^k \left(\sqrt{a_i}-\sqrt{b_i}\right)^2\right)\left(\sum_{i=1}^k \left(\sqrt{a_i}+\sqrt{b_i}\right)^2\right)\\ &\quad \text{(使用 Cauchy-Schwarz 不等式)}\\ &\le 4\sum_{i=1}^k \left(\sqrt{a_i}-\sqrt{b_i}\right)^2\\ &= 8\left(1-\sum_{i=1}^k \sqrt{a_i b_i}\right),\end{aligned}$$

可证引理。 [证毕]

信息散度可通过取 $f(x)=x\log x$ 得到：

$$I(\mu,\nu)=\sup_{\mathcal{A}=\{A_j\}}\sum_j \mu(A_j)\log\left(\frac{\mu(A_j)}{\nu(A_j)}\right).$$

$I(\mu,\nu)$ 也称为 Kullback-Leibler 数。

最后一个例子是χ-散度，通过令 $f(x)=(x-1)^2$ 得到：

$$\begin{aligned}\chi^2(\mu,\nu)&=\sup_{\mathcal{A}=\{A_j\}}\sum_j \frac{\left(\mu(A_j)-\nu(A_j)\right)^2}{\nu(A_j)}\\ &=\int \frac{p^2(x)}{q(x)}\lambda(dx)-1.\end{aligned}$$

接下来，我们强调 F-误差与 f-散度直接的联系。令给定 $\{Y=0\}$ 和$\{Y=1\}$ 时的条件分布 μ_0,μ_1。假设类概率是相等的 $p=1/2$。若 F 是凹函数，则 F-误差 $d_F(X,Y)$ 可重写为

$$d_F(X,Y)=F\left(\frac{1}{2}\right)-D_f(\mu_0,\mu_1),$$

其中

$$f(x)=-\frac{1}{2}F\left(\frac{x}{1+x}\right)(1+x)+F\left(\frac{1}{2}\right),$$

且 D_f 是对应的 f-散度。易见 f 是凸的，因为 F 是凹函数。这种对应的特殊情况是

$$L^*=\frac{1}{2}\left(1-\frac{1}{2}V(\mu_0,\mu_1)\right),$$

若 $p=1/2$。而且，容易验证

$$\rho=\sqrt{p(1-p)}I_2(\mu_0,\mu_1),$$

其中 ρ 是 Matushita 误差。想了解更多相关联系，请读者参考本章习题。

3.10 问题与练习

问题 3.1

证明对任 (l_1, l^*)，$0 \le l^* \le l_1 \le 2l^*(1-l^*) \le 1/2$，存在某 (X,Y) 分布使 $L_{NN} = l_1$和$L^* = l^*$ 成立。因此，Cover-Hart 不等式无法（对所有分布——普遍性——）进一步改进。

问题 3.2

定理3.1界是紧的

定理3.1不能进一步改进。

(1) 证明对所有 $\alpha \in [0,1/2]$，存在某 (X,Y) 分布使 $L_{\mathrm{NN}} = L^* = \alpha$ 成立。

(2) 证明对所有 $\alpha \in [0,1/2]$，存在某 (X,Y) 分布使 $L_{\mathrm{NN}}\alpha, L^* = \frac{1}{2} - \frac{1}{2}\sqrt{1-2\alpha}$ 成立。

(3) 证明对所有 $\alpha \in [0,1/2]$，存在某 (X,Y) 分布使 $L_{\mathrm{NN}} = \rho = \alpha$ 成立。

(4) 证明对所有 $\alpha \in [0,1/2]$，存在某 (X,Y) 分布使 $L_{\mathrm{NN}} = \alpha, L^* = \frac{1}{2} - \frac{1}{2}\sqrt{1-4\rho^2}$ 成立。

问题 3.3

证明 $\mathscr{C} \ge L^*$。

问题 3.4

对任 $\alpha \le 1$，找到具条件熵 $\mathscr{C}_n$ 与贝叶斯误差 L_n^* 的分布序列 (X_n, Y_n) 使得当 $n \to \infty$ 时 $L_n^* \to 0$ 成立，且 $\mathscr{C}_n$ 以 $\left(L_n^*\right)^\alpha$ 同样得速度趋于 0。

问题 3.5

误差度量的凹性质

令 Y 为混合随机变量，以概率 p 取 Y_1 值，概率 $1-p$ 取 Y_2 值。令 X 为固定的 $\mathcal{R}^d$-值随机变量，定义 $\eta_1(x) = \mathbf{P}\{Y_1 = 1 \mid X = x\}, \eta_2(x) = \mathbf{P}\{Y_2 = 1 \mid X = x\}$，其中 Y_1, Y_2 为伯努利随机变量。显然，$\eta(x) = p\eta_1(x) + (1-p)\eta_2(x)$。哪一个误差度量 $L^*, \rho, L_{\mathrm{NN}}, \mathscr{C}$ 关于 p 是凹的，对固定的 X, Y_1, Y_2 联合分布？是否每一个判别问题 (X,Y) 都能按这种方式分解，对某些 Y_1, p, Y_2，其中 $\eta_1(x), \eta_2(x) \in \{0,1\}, \forall x$？如果不是，换成条件 $\eta_1(x), \eta_2(x) \in \{0, 1/2, 1\}, \forall x$ 呢？

问题 3.6

证明对任 $l^* \in [0,1/2]$，存在一个 (X,Y) 分布使得 $L^* = l^*$ 和$\mathscr{C} = \mathscr{H}(L^*, 1-L^*)$。因此 Fano 不等式是紧的。

Toussaint 不等式（1974B）

问题 3.7

模仿前面的方式证明不等式

$$\mathcal{J} \geq \sqrt{1-2L_{\mathrm{NN}}} \log\left(\frac{1+\sqrt{1-2L_{\mathrm{NN}}}}{1-\sqrt{1-2L_{\mathrm{NN}}}}\right) \geq 2\left(1-2L_{\mathrm{NN}}\right).$$

问题 3.8

证明 $L^* \leq e^{-\delta_C}$，其中 δ_C 为参数 $\alpha \in (0,1)$Chernoff 亲和力度量。

问题 3.9

证明 $L^* = p - \mathbf{E}\{(2\eta(X)-1)_+\}$，其中 $p = \mathbf{P}\{Y=1\}$，和$(x)_+ = \max(x,0)$。

Toussaint（1974b）

问题 3.10

证明 $\mathcal{J} \geq -2\log\rho - 2\mathcal{H}(p,1-p)$，其中 $p = \mathbf{P}\{Y=1\}$。

问题 3.11

令 f_1, f_0 为两个多元正态密度，均值为 $m1, m0$，具相同的协方差矩阵 Σ。若 $p = \mathbf{P}\{Y=1\}$，且 f_1, f_0 分别为给定 $Y=1, Y=0$ 时 X 的条件密度。证明

$$\rho = \sqrt{p(1-p)}e^{-\Delta^2/8},$$

其中 ρ 为 Matushita 误差，Δ 为 Mahalanobis 距离。

问题 3.12

对任 $\delta \in [0,\infty)$ 和$l^* \in [0,1/2]$ 具$l^* \leq 2/\left(4+\delta^2\right)$，寻找给定 $Y=1, Y=0$ 时 X 的分布 μ_0, μ_1 使得 Mahalanobis 距离 $\Delta = \delta$，但（yet）$L^* == l^*$。因此，Mahalanobis 距离不能普遍地与 Bayes 风险相关。

问题 3.13

证明 Mahalanobis 距离 Δ 在 X 的线性可逆变换下是不变的。

问题 3.14

Lissack 和 Fu（1976）提出度量

$$\delta_{\mathrm{LF}} = \mathbf{E}\{|2\eta(X)-1|^\alpha\}, \quad 0 < \alpha < \infty$$

对 $\alpha = 1$，该度量是 Kolmogorov 距离 δ_{KO} 的两倍。证明

(1) 若 $0 < \alpha \leq 1$, 则$\frac{1}{2}\left(1-\delta_{\mathrm{LF}}\right) \leq L^* \leq \left(1-\delta_{\mathrm{LF}}^{1/\alpha}\right)$。

(2) 若 $1 \le \alpha < \infty$, then $\frac{1}{2}\left(1-\delta_{\mathrm{LF}}^{1/\alpha}\right) \le L^* \le (1-\delta_{\mathrm{LF}})$。

问题 3.15

Hashlamoun，Varshney 和 Samarasooriya（1994）提出使用具

$$F(x) = \frac{1}{2}\sin(\pi x)e^{-1.8063\left(x-\frac{1}{2}\right)^2}$$

的 F-误差来获得 L^* 的紧上界。证明 $F(x) \ge \min(x, 1-x)$，因此相应的 F-误差确实是 Bayes 风险的上界。

问题 3.16

证明 $L^* \le \max(p(1-p))\left(1-\frac{1}{2}V(\mu_0,\mu_1)\right)$。

问题 3.17

证明 $L^* \le \sqrt{p(1-p)}I_2(\mu_0\mu_1)$。[提示：$\min(a,b) \le \sqrt{ab}$。]

问题 3.18

假设 $X = \left(X^{(1)},\ldots,X^{(d)}\right)$ 的分量是条件独立（给定 Y），同分布的。即 $\mathbf{P}\left\{X^{(i)} \in A \mid Y = j\right\} = v_j(A)$ 对 $i = 1,\ldots,d$ 和 $j = 0,1$。使用上一个问题证明

$$L^* \le \sqrt{p(1-p)}\left(I_2(v_0,v_1)\right)^d.$$

问题 3.19

证明 $\chi^2(\mu_1,\mu_2) \ge I(\mu_1,\mu_2)$。[提示：$x-1 \ge \log x$。]

问题 3.20

证明下述定理（类似定理3.3）。令 $t : \mathcal{R}^d \to \mathcal{R}^k$ 为可测函数，μ, v 为 $\mathcal{R}^d$ 上的概率测度。定义在 $\mathcal{R}^k$ 上的测度 $\mu_t(A) = \mu\left(t^{-1}(A)\right)$ 和 $v_t(A) = v\left(t^{-1}(A)\right)$。证明对任一凸函数 f，有 $D_f(\mu,v) \ge D_f(\mu_t,v_t)$。

问题 3.21

证明下述 Hellinger 积分与全变分的联系：

$$\left(1-\frac{1}{2}V(\mu,v)\right) \ge \frac{1}{2}\left(I_2(\mu,v)\right)^2,$$

和

$$\left(V(\mu,v)\right)^2 \le 8\left(1-I_2(\mu,v)\right).$$

[提示：过程类似引理3.1和3.2。]

问题 3.22

Pinsker 不等式（Csiszar（1967），Kullback（1967）和 Kemperman（1969））

证明

$$(V(\mu,\nu))^2 \leq 2I(\mu,\nu)$$

[提示：首先证明若 μ,ν 集中在同样的两个原子上时不等式成立。然后定义 $A = \{x : p(x) \geq q(x)\}$，并通过 $\mu^*(0) = 1-\mu^*(1) = \mu(A)$ 和 $\nu^*(0) = 1-\nu^*(1) = \nu(A)$ 定义在 $\{0,1\}$ 上的 μ^*,ν^*，应用上述结果。最后指出 Scheffé 定理表明 $V(\mu^*,\nu^*) = V(\mu,\nu)$，故 $I(\mu^*,\nu^*) \leq I(\mu,\nu)$。]

线性判别 4

本章中，我们使用一个超平面切割空间并分配两个不同的类到每个半空间。这些规则具有巨大的优点——易于解释基于 $\sum_{i=1}^{d} a_i x^{(i)} + a_0$ 的符号给出的每个决策，其中 $x = \left(x^{(1)}, \ldots, x^{(d)}\right)$，$a_i$ 是权重。权重向量决定每个分量的相对重要性。同时，决策是容易实现的——在标准软件解决方案中，决策时间与 d 成比例——制造一个小芯片来作出几乎是瞬间决策的前景是尤为令人激动的。

Rosenblatt（1962）实现了这种线性规则的巨大潜能，称它们为*感知器*。当新数据到达时改变一个或多个权重允许我们快速且容易的使权重适应新的情况。以人类大脑为模式的训练或学习因此成为一种现实。本章仅介绍一些感知器的理论性质。我们从简单的一维情况开始，然后进一步处理在 $\mathcal{R}^d$ 中权重的选择。除非足够幸运，不然线性判别规则无法提供接近贝叶斯风险的误差概率。但不能因此忽视本章的价值。线性判别几乎是每个成功的模式识别方法的核心，包括树分类器（第 20 和 21 章），广义线性分类器（第 17 章）和神经网络（第 30 章）。我们也会首次遇到参数（权重）依赖于数据的规则。

4.1 一元判别与 Stoller 切割

作为引入示例，令 X 为一元变量。最简单的规则为线性判别规则

$$g(x) = \begin{cases} y' & \text{if } x \le x' \\ 1 - y' & \text{otherwise}, \end{cases}$$

其中 x' 为割点，$y' \in \{0, 1\}$ 是类。一般而言，x', y' 是数据 D_n 的可测函数。在这个简单规则族中，如果我们知道具体的分布，则显然存在一个最优的规则。例如假设对 (X, Y)：令 $\mathbf{P}\{Y = 1\} = p$。给定 $Y = 1$，X 的分布函数为 $F_1(x) = \mathbf{P}\{X \le x \mid Y = 1\}$；给定 $Y = 0$，为 $F_0(x) = \mathbf{P}\{X \le x \mid Y = 0\}$，称 F_0, F_1 为*类条件分布函数*。则理论最优规则可通过割点 x^* 和类 y^*

$$(x^*, y^*) = \arg\min_{(x', y')} \mathbf{P}\{g(X) \ne Y\}$$

得到（若允许 $x' = \infty$ 和$x' = -\infty$ 则最小值总可以取得）。我们将相应的的*最小误差概率*表示为

$$L = \inf_{(x', y')} \left\{ I_{\{y'=0\}} \left(pF_1(x') + (1-p)(1 - F_0(x')) \right) + I_{\{y'=1\}} \left(p(1 - F_1(x')) + (1-p)F_0(x') \right) \right\}$$

由 (x^*, y^*) 定义的切割称为理论 Stoller *切割*（Stoller（1954））。

引理 4.1 *$L \le 1/2$，等式成立当且仅当 $L^* = 1/2$。*

证. 取 $(x', y') = (-\infty, 0)$。则误差概率为 $1 - p = \mathbf{P}\{Y = 0\}$。取 $(x', y') = (-\infty, 1)$，则误差概率为 p。显然

$$L^* \le L \le \min(p, 1-p)$$

译注：由于 L 是取最小值 1/2

这证明了引理的第一部分。对第二部分，若 $L = 1/2$，则 $p = 1/2$，且对任 x，有 $pF_1(x) + (1-p)(1 - F_0(x)) \ge 1/2$ 和 $p(1 - F_1(x)) + (1-p)F_0(x) \ge 1/2$。第一个不等式说明 $pF_1(x) - (1-p)F_0(x) \ge p - 1/2$；第二个不等式说明 $pF_1(x) - (1-p)F_0(x) \le p - 1/2$。因此 $L = 1/2$ 意味着对任 x，有 $pF_1(x) - (1-p)F_0(x) = p - 1/2$。因此对任 x，$F_1(x) = F_0(x)$，故 $L^* = 1/2$。 [证毕]

引理 4.2

$$L = \frac{1}{2} - \sup_x \left| pF_1(x) - (1-p)F_0(x) - p + \frac{1}{2} \right|.$$

特别是，若 $p = 1/2$，则

$$L = \frac{1}{2} - \frac{1}{2} \sup_x |F_1(x) - F_0(x)|.$$

证. 令 $\rho(x) = pF_1(x) - (1-p)F_0(x)$。则，由定义，

$$\begin{aligned} L &= \inf_x \min\{\rho(x) + 1 - p, p - \rho(x)\} \\ &= \frac{1}{2} - \sup_x \left| \rho(x) - p + \frac{1}{2} \right| \\ &\quad (\text{由 } \min\{a, b\} = (a + b - |a - b|)/2). \end{aligned}$$

[证毕]

最后这个性质连接了理论 Stoller 切割质量与 Kolmogorov-Smirnov 距离 $\sup_x |F_1(x) - F_0(x)|$，其中后者度量类条件分布函数的距离。作为有趣的练习，考虑两个类具均值 $m_0 = \mathbf{E}\{X \mid Y = 0\}, m_1 = \mathbf{E}\{X \mid Y = 1\}$，方差 $\sigma_0^2 = \mathrm{Var}\{X \mid Y = 0\}$ 和 $\sigma_1^2 = \mathrm{Var}\{X \mid Y = 1\}$。则下述不等式成立。

定理 4.1

$$L^* \le L \le \frac{1}{1 + \frac{(m_0 - m_1)^2}{(\sigma_0 + \sigma_1)^2}}.$$

备注 4.1 当 $p = 1/2$，Chernoff（1971）证明

$$L \le \frac{1}{2 + 2\frac{(m_0 - m_1)^2}{(\sigma_0 + \sigma_1)^2}}.$$

另外，Becker（1968）指出该界是最优的（见问题 4.2）。

证. 不失一般性，假设 $m_0 < m_1$。若令规则 A 使类取 0 当 $x \le m_0 + \Delta_0$，否则取 1，其中 $m_1 - m_0 = \Delta_0 + \Delta_1$，令 $\Delta_0, \Delta_1 > 0$。显然，L 比该规则的

误差概率小。规则 A 的误差概率为

$$
\begin{aligned}
&pF_1(m_0+\Delta_0)+(1-p)(1-F_0(m_0+\Delta_0))\\
&\quad=p\mathbf{P}\{X\le m_1-\Delta_1\mid Y=1\}+(1-p)\mathbf{P}\{X>m_0+\Delta_0\mid Y=0\}\\
&\quad\le p\frac{\sigma_1^2}{\sigma_1^2+\Delta_1^2}+(1-p)\frac{\sigma_0^2}{\sigma_0^2+\Delta_0^2}\\
&\quad\text{(使用 Chebyshev-Cantelli 不等式；见附录，定理 A.17)}\\
&\quad=\frac{p}{1+\frac{\Delta_1^2}{\sigma_1^2}}+\frac{1-p}{1+\frac{\Delta_0^2}{\sigma_0^2}}\\
&\quad\left(\text{取}\Delta_1=(\sigma_1/\sigma_0)\Delta_0,\ \text{和}\Delta_0=|m_1-m_0|\,\sigma_0/(\sigma_0+\sigma_1)\right)\\
&\quad=\frac{1}{1+\frac{(m_0-m_1)^2}{(\sigma_0+\sigma_1)^2}}.
\end{aligned}
$$

[证毕]

我们因此有了又一个例子证明切割良好的类产生小的 L，因此 L^* 也是小的。切割的度量（measured）通过关于 $\sigma_0+\sigma_1$ 的 $|m_1-m_0|$ 值的大小进行。同样思路的另一个不等式可见问题 4.1。

理论 Stoller 切割的局限体现在下面一个简单示例中。考虑均匀分布 [0,1] 随机变量 X，定义

$$
Y=\begin{cases}1 & \text{if } 0\le X\le \frac{1}{3}+\epsilon\\ 0 & \text{if } \frac{1}{3}+\epsilon<X\le\frac{2}{3}-\epsilon\\ 1 & \text{if } \frac{2}{3}-\epsilon\le X\le 1\end{cases}
$$

对某些小 $\epsilon>0$。因为 Y 是 X 的函数，我们有 $L^*=0$。若我们必须做一个平凡 X-独立决策，则最优决策是 $g(x)\equiv 1$。该决策的误差概率为 $\mathbf{P}\{1/3+\epsilon<X<2/3-\epsilon\}=1/3-2\epsilon$。接下来考虑理论 Stoller 切割。可以看出最好的的割点为 $x'=0$ 或 $x'=1$，因此得到 $L=1/3-2\epsilon$。换句话说，即使是最好的理论切割也是不足的。同时从该例子我们也注意到，$m_0=m_1=1/2$，因此定理4.1的不等式说明 $L\le 1$——它退化了。

我们现在考虑当切割必须是基于数据的该怎么做。Stoller（1954）建议取 (x',y') 使得经验误差最小。他找到 (x',y') 使得

$$
(x',y')=\underset{(x,y)\in\mathcal{R}\times\{0,1\}}{\arg\min}\frac{1}{n}\sum_{i=1}^{n}\left(I_{\{X_i\le x,Y_i\ne y\}}+I_{\{X_i>x,Y_i\ne 1-y\}}\right).
$$

（x' 和 y' 现在是随机变量，尽管有前面约定的惯例，但是我们现在使用小写表示随机变量。）我们称之为 Stoller 规则。该切割称为*经验 Stoller 切割*。将集合 $\{(-\infty,x]\times\{y\}\}\cup\{(x,\infty)\times\{1-y\}\}$ 表示为 $C(x,y)$。则

$$
(x',y')=\underset{(x,y)}{\arg\min}v_n(C(x,y))
$$

其中 v_n 为数据 $D_n=(X_1,Y_1),\ldots,(X_n,Y_n)$ 的经验度量，即，对任一可测集 $A\in\mathcal{R}\times\{0,1\}$，$v_n(A)=(1/n)\sum_{i=1}^{n}I_{\{(X_i,Y_i)\in A\}}$。将 (X,Y) 在 $\mathcal{R}\times\{0,1\}$ 的度量称为 v，明显 $\mathbf{E}\{v_n(C)\}=v(C)=\mathbf{P}\{X\le x,Y\ne y\}+\mathbf{P}\{X>x,Y\ne 1-y\}$。令 $L_n=\mathbf{P}\{g_n(X)\ne Y\mid D_n\}$ 为切割规则 g_n 的误差概率，其中 g_n

基于上述给出的数据相关的选择 (x', y')，在数据上的条件。则

$$\begin{aligned} L_n =& v\left(C\left(x', y'\right)\right) \\ =& v\left(C\left(x', y'\right)\right) - v_n\left(C\left(x', y'\right)\right) + v_n\left(C\left(x', y'\right)\right) \\ \leq& \sup_{(x,y)}\left(v(C(x,y)) - v_n(C(x,y))\right) + v_n\left(C\left(x^*, y^*\right)\right) \\ & \left(\text{其中 } (x^*, y^*) \text{ 最小化} v(C(x,y))\forall(x,y)\right) \\ \leq& 2\sup_{(x,y)}\left|v(C(x,y)) - v_n(C(x,y))\right| + v\left(C\left(x^*, y^*\right)\right) \\ =& 2\sup_{(x,y)}\left|v(C(x,y)) - v_n(C(x,y))\right| + L \end{aligned}$$

从下一个定理我们将看到上述的上确界是小的，即使中等大小的 n，因此 Stoller 规则的表现接近最优切割对任意 (X, Y) 分布。

定理 4.2 *对 Stoller 规则，令 $\epsilon > 0$，有*

$$\mathbf{P}\left\{L_n - L \geq \epsilon\right\} \leq 4e^{-n\epsilon^2/2},$$

和

$$\mathbf{E}\left\{L_n - L\right\} \leq \sqrt{\frac{2\log(4e)}{n}}.$$

证. 通过上面定理给出的不等式，

$$\begin{aligned} \mathbf{P}\left\{L_n - L \geq \epsilon\right\} \leq& \mathbf{P}\left\{\sup_{(x,y)}\left|v(C(x,y)) - v_n(C(x,y))\right| \geq \frac{\epsilon}{2}\right\} \\ \leq& \mathbf{P}\left\{\sup_{x}\left|v(C(x,0)) - v_n(C(x,0))\right| \geq \frac{\epsilon}{2}\right\} \\ & + \mathbf{P}\left\{\sup_{x}\left|v(C(x,1)) - v_n(C(x,1))\right| \geq \frac{\epsilon}{2}\right\} \\ \leq& 4e^{-2n(\epsilon/2)^2} \end{aligned}$$

定理 12.9

和通过两次运用 Dvoretzky-Kiefer-Wolfowitz 不等式（1956）的 Massart（1990）紧版本。见问题 4.5。我们在这不证明该不等式，但我们将在第 12 章以更一般地形式详细地讨论这类不等式。问题 12.1 可证其后的第二个不等式。 [证毕]

Stoller 规则的误差概率在任意分布上一致趋近于 L。这只是未来的预演，在后面我们能在限制的规则族中获得好的性能保证。

4.2 线性判别

Rosenblatt 感知器（Rosenblatt（1962）；在 Nilsson（1965）叙述更详）使用超平面将 $\mathcal{R}^d$ 切割成两个部分。定义具权重 $a_0, a_1, ..., a_d$ 的线性判别规则为

$$g(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{d} a_i x^{(i)} + a_0 > 0 \\ 0 & \text{otherwise} \end{cases}$$

其中 $x = (x^{(1)}, \ldots, x^{(d)})$。其误差概率标为 $L(a, a_0)$, 其中 $a = (a_1, \ldots, a_d)$。再次，我们令

$$L = \inf_{a \in \mathcal{R}^d, a_0 \in \mathcal{R}} L(a, a_0)$$

为这一类下的最优误差概率。令 $a_1 X^{(1)} + \cdots + a_d X^{(d)}$ 的类条件分布函数根据 $Y = 0$ 或 $Y = 1$ 分别表示为 $F_{0,a}$ 和 $F_{1,a}$。对 $L(a, a_0)$，我们使用引理4.2的界并应用至 $F_{0,a}$ 和 $F_{1,a}$。因此，

$$L = \frac{1}{2} - \sup_a \sup_x \left| pF_{1,a}(x) - (1-p)F_{0,a}(x) - p + \frac{1}{2} \right|,$$

对 $p = 1/2$，有

$$L = \frac{1}{2} - \frac{1}{2} \sup_a \sup_x \left| F_{1,a}(x) - F_{0,a}(x) \right|.$$

因此，$L = 1/2$ 当且仅当 $p = 1/2$，且对任 a，$F_{1,a} \equiv F_{0,a}$。然后应用下述简单引理。

引理 4.3 *令 X_1, X_2 是在 $\mathcal{R}^d$ 取值的随机变量，则它们是同分布的当且仅当 $a^T X_1$ 和 $a^T X_2$ 对 $\forall a \in \mathcal{R}^d$ 具有相同的分布。*

Cramér 和 Wold（1936）

证. 两个随机变量具有相同的分布当且仅当它们具有相同的特征函数——见 Lukács 和 Laha（1964）。现在，$X_1 = \left(X_1^{(1)}, \ldots, X_1^{(d)}\right)$ 的特征函数为

$$\begin{aligned}
\psi_1(a) &= \psi_1(a_1, \ldots, a_d) \\
&= \mathbf{E}\left\{ e^{i\left(a_1 X_1^{(1)} + \cdots + a_d X_1^{(d)}\right)} \right\} \\
&= \mathbf{E}\left\{ e^{i\left(a_1 X_2^{(1)} + \cdots + a_d X_2^{(d)}\right)} \right\} \quad \text{(通过假设)} \\
&= \psi_2(a_1, \ldots, a_d),
\end{aligned}$$

即 X_2 的特征函数。 [证毕]

因此我们证明了如下定理：

定理 4.4 $L \le 1/2$，*其中等式成立当且仅当* $L^* = 1/2$。

因此，在一维情况，只要 $L^* < 1/2$，一个有意义（$L < 1/2$）的超平面切割是可能存在的。但也存在一种情况，其中任意切割在规则 $g(x) \equiv y, \forall x,$ 某 y 上都是平凡的，即使 $L^* = 0, L > 1/4$ 成立。[5]为推广定理4.1，我们提供如下结果。相关不等式可见问题 4.7。使用 Chebyshev 不等式来获得这些界的思想来自于 Yau 和 Lin（1968）。（也可见于 Devijver 和 Kittler（1982，162 页））。

5: 许有问题：There are also examples in which no cut improves over a rule in which $g(x) \equiv y$ for some y and all x, yet $L^* = 0$ and $L > 1/4$ (say).

定理 4.4 *令 X_0 和 X_1 为给定 $Y = 0, Y = 1$ 时 X 的随机变量。令 $m_0 = \mathbf{E}\{X_0\}, m_1 = \mathbf{E}\{X_1\}$。定义协方差矩阵 $\Sigma_1 = \mathbf{E}\left\{(X_1 - m_1)(X_1 - m_1)^T\right\}$ 和 $\Sigma_0 = \mathbf{E}\left\{(X_0 - m_0)(X_0 - m_0)^T\right\}$。则*

$$L^* \le L \le \inf_{a \in \mathcal{R}^d} \frac{1}{1 + \frac{\left(a^T(m_1 - m_0)\right)^2}{\left(\left(a^T \Sigma_0 a\right)^{1/2} + \left(a^T \Sigma_1 a\right)^{1/2}\right)^2}}.$$

证. 对 $\forall a \in \mathcal{R}^d$ 我们应用定理4.1至 $a^T X_0$ 和 $a^T X_1$。通过

$$\begin{aligned}\mathbf{E}\left\{a^T X_0\right\} &= a^T \mathbf{E}\left\{X_0\right\} = a^T m_0,\\ \mathbf{E}\left\{a^T X_1\right\} &= a^T m_1,\end{aligned}$$

和

$$\begin{aligned}\mathrm{Var}\left\{a^T X_0\right\} &= \mathbf{E}\left\{a^T \left(X_0 - m_0\right)\left(X_0 - m_0\right)^T a\right\} = a^T \Sigma_0 a\\ \mathrm{Var}\left\{a^T X_1\right\} &= a^T \Sigma_1 a\end{aligned}$$

可证定理。 [证毕]

我们能通过 a 的不同选择获得显式不等式。$a = m_1 - m_0$ 推出方便的表达形式。我们在下一节可以看到 $a = \Sigma(m_1 - m_0)$ 也是有意义的选择，其中 $\Sigma = p\Sigma_1 + (1-p)\Sigma_0$。（见问题 4.7）。

4.3 Fisher 线性判别

a 基于数据的值在好几个准则中出现。第一个方法是由 Fisher (1936) 提出的。令 $\widehat{m}_1, \widehat{m}_0$ 为两个类的样本均值 (即, $\widehat{m}_1 = \sum_{i:Y_i=1} X_i / |\{i : Y_i = 1\}|$)。图片（显示了）投影 $X_1, ..., X_n$ 至一条以 a 为方向的直线上。注意 a 垂直于超平面 $a^T x + a_0 = 0$。投影后的值为 $a^T X_1, \ldots, a^T X_n$。那些在超平面 $a^T x = 0$ 上的 X_i 对应的投影值均为 0，若数据点离开超平面，投影值的绝对值增加。令 $\widehat{\sigma}_1^2$ 和 $\widehat{\sigma}_0^2$ 为类 1 和 0 的样本散布（sample scatters），即

$$\widehat{\sigma}_1^2 = \sum_{i:Y_i=1} \left(a^T X_i - a^T \widehat{m}_1\right)^2 = a^T S_1 a$$

对 $\widehat{\sigma}_0^2$ 也是类似的，其中

$$S_1 = \sum_{i:Y_i=1} \left(X_i - \widehat{m}_1\right)\left(X_i - \widehat{m}_1\right)^T$$

是类 1 的散布矩阵。

Fisher 线性判别是使准则（criterion）

$$J(a) = \frac{\left(a^T \widehat{m}_1 - a^T \widehat{m}_0\right)^2}{\widehat{\sigma}_1^2 + \widehat{\sigma}_0^2} = \frac{\left(a^T \left(\widehat{m}_1 - \widehat{m}_0\right)\right)^2}{a^T \left(S_1 + S_0\right) a}$$

最大化的线性函数 $a^T x$。这相当于寻找一个方向 a，相对于样本散布[6]，它使 $a^T \hat{m}_1$ 与 $a^T \hat{m}_0$ 达到最大程度的分离。幸运的是，寻找 a 不需要数值迭代——其解析解为

6: 许有问题

$$a = \left(S_1 + S_0\right)^{-1} \left(\widehat{m}_1 - \widehat{m}_0\right).$$

Fisher 建议替换 $(X_1, Y_1), \ldots, (X_n, Y_n)$ 成 $\left(a^T X_1, Y_1\right), \ldots, \left(a^T X_n, Y_n\right)$ 并进行一维判别。一般来说，规则对某常量 a_0 使用切割

$$g_{a_0}(x) = \begin{cases} 1 & \text{if } a^T x + a_0 > 0\\ 0 & \text{otherwise .}\end{cases} \tag{4.1}$$

不幸的是，Fisher 判别可以是任意差的：存在一些判别，在两个类是线性可分（即 $L = 0$）的情况下，Fisher 线性判别的误差概率接近于 1（见问题 4.9）。

4.4 正态分布

在一些纯粹偶然的情况下，贝叶斯规则是线性判别。然而这并不是主要问题，关键在于识别最一般的情况，即多元正态分布。一般多元正态密度记为

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2}(x-m)^T \Sigma^{-1}(x-m)},$$

其中 m 为均值（x 和 m 为 d-分量列向量），Σ 为 $d \times d$ 协方差矩阵，Σ^{-1} 为 Σ 的逆，$det(\Sigma)$ 表示行列式。我们记 $f \sim N(m, \Sigma)$。显然，若 X 具密度 f，则 $m = \mathbf{E}X$ 和 $\Sigma = \mathbf{E}\left\{(X-m)(X-m)^T\right\}$。

多元正态密度能被 $d + \binom{d}{2}$ 个（形式）参数（m 和 Σ）完全确定。该密度的样本聚集在一朵椭圆云中。等密度点的轨迹是椭圆

$$(x-m)^T \Sigma^{-1}(x-m) = r^2$$

对某常数 $r > 0$。数 r 为从 x 到 m 的 Mahalanobis 距离，即使基本分布（underlying distribution）不是正态分布 r 也是有用的。它考虑 Σ 决定的空间 的方向拉伸（stretch）。给定一个两类问题，X 的密度为 $(1-p)f_0(x) + pf_1(x)$，其中 f_0, f_1 都是参数为 $m_i, \Sigma_i, i = 0, 1$ 的多元正态密度，则对应的贝叶斯规则

问：什么空间?

$$g^*(x) = \begin{cases} 1 & \text{if } pf_1(x) > (1-p)f_0(x) \\ 0 & \text{otherwise.} \end{cases}$$

取对数并注意到 $g^*(x) = 1$ 当且仅当

问：对什么取对数?
译者注：将多元正态密度公式代入 $pf_1(x) > (1-p)f_0(x)$。

$$\begin{aligned} &(x-m_1)^T \Sigma_1^{-1}(x-m_1) - 2\log p + \log(\det(\Sigma_1)) \\ &\quad < (x-m_0)^T \Sigma_0^{-1}(x-m_0) - 2\log(1-p) + \log(\det(\Sigma_0)). \end{aligned}$$

实际上，我们可能想要从数据估计 $m_1, m_0, \Sigma_1, \Sigma_0$ 和 p 并使用这些估计计算 g^*。有趣的是，由于 $(x-m_i)^T \Sigma_i^{-1}(x-m_i)$ 是从 x 到类 i 的 m_i 的均方 Mahalanobis 距离（记为 r_i^2），贝叶斯规则为

$$g^*(x) = \begin{cases} 1 & \text{if } r_1^2 < r_0^2 - 2\log((1-p)/p) + \log(\det(\Sigma_0)/\det(\Sigma_1)) \\ 0 & \text{otherwise.} \end{cases}$$

特别是，当 $p = 1/2, \Sigma_0 = \Sigma_1 = \Sigma$，我们有

$$g^*(x) = \begin{cases} 1 & \text{if } r_1^2 < r_0^2 \\ 0 & \text{otherwise ;} \end{cases}$$

仅根据均值按 Mahalanobis 距离最接近 x 的类进行分类。当 $\Sigma_0 = \Sigma_1 = \Sigma$，贝叶斯规则是线性的：

$$g^*(x) = \begin{cases} 1 & \text{if } a^T x + a_0 > 0 \\ 0 & \text{otherwise ,} \end{cases}$$

其中 $a = (m_1 - m_0)\Sigma^{-1}$，与 $a_0 = 2\log(p/(1-p)) + m_0^T\Sigma^{-1}m_0 - m_1^T\Sigma^{-1}m_1$。因此，线性判别规则是多元正态分布 Bayes 规则的特定情况。

我们直觉 a 在 $m_1 - m_0$ 方向是切割类的最优方向。以上叙述说明该直觉近乎是正确的。然而注意 a 一般并不垂直于到 m_0 和 m_1 等距的轨迹超平面。当 Σ 被替换成标准基于数据的估计时，我们实际上得到 Fisher 线性判别。另外，当 $\Sigma_1 \neq \Sigma_0$，则决策界一般不是线性的，Fisher 线性判别因此是次优的。

备注 4.2 在早期统计判别文献中，正态分布处于中心地位（Anderson（1958））。作为一个简介，我们参考 Duda 和 Hart（1973）。McLachlan（1992）介绍更多细节，Raudys（1972；1976）阐述了误差、维数和样本大小与正态（或近似正态）模型之间的关系。也见 Raudys 和 Pikelis（1980；1982）。

4.5 经验风险最小化

在本节中我们介绍一个分类器算法，其概率误差接近线性分类器能达到的最小误差概率 L，只要 X 具有一个密度。该算法从有限个——$2\binom{n}{d}$——线性分类器中选择一个使经验误差最小的分类器。对规则

$$\phi(x) = \begin{cases} 1 & \text{if } a^T x + a_0 > 0 \\ 0 & \text{otherwise,} \end{cases}$$

其经验误差为

$$L(\phi) = \mathbf{P}\{\phi(X) \neq Y\}.$$

$L(\phi)$ 可由经验损失

$$\widehat{L}_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\phi(X_i)\neq Y_i\}}$$

估计，即统计分类器 ϕ 出错的次数并对其标准化。

假设 X 具有一个密度，并考虑在 $\{X_1, \ldots, X_n\}$ 中 d 个任意数据点 $X_{i_1}, X_{i_2}, \ldots, X_{i_d}$。令 $a^T x + a_0 = 0$ 为包含这些点的超平面。因为已假设 X 的密度，故这 d 个点以概率 1 位于一般（general）位置，因此这一超平面是唯一的。该超平面确定了两个分类器：

$$\phi_1(x) = \begin{cases} 1 & \text{if } a^T x + a_0 > 0 \\ 0 & \text{otherwise,} \end{cases}$$

和

$$\phi_2(x) = \begin{cases} 1 & \text{if } a^T x + a_0 < 0 \\ 0 & \text{otherwise,} \end{cases}$$

其经验误差 $\widehat{L}_n(\phi_1)$ 和 $\widehat{L}_n(\phi_2)$ 是可求的。对数据点的任一 d 元组 $X_{i_1}, X_{i_2}, \ldots, X_{i_d}$，我们以这种方式生成两个分类器，故一共能生成 $2\binom{n}{d}$ 个分类器，分别记为 $\phi_1, \ldots, \phi_{2\binom{n}{d}}$。令 $\widehat{\phi}$ 为最小化 $\widehat{L}_n(\phi_i)$ 的线性分类器，$i = 1, \ldots, 2\binom{n}{d}$。

我们记在线性规则族中最小的误差概率为

$$L = \inf_{\phi} L(\phi),$$

并定义 $\phi^* = \arg\min_\phi L(\phi)$ 为最优线性规则。若存在几个线性分类器有 $L(\phi) = L$，则可以按任意方式选择一个 ϕ^*。接下来我们证明对应于 $\widehat{\phi}$ 的分类器确实是非常“好”的。

首先注意到不存在经验误差 $\widehat{L}_n(\phi)$ 小于 $\widehat{L}_n(\widehat{\phi}) - d/n$ 的线性分类器 ϕ。这是由于数据点都在一般位置上（回忆密度假设），因此对任一线性分类器我们可以找到其中一个，该分类器定义的超平面正好包含了 d 个数据点，使得这两个决策在除了 d 个点外的所有的点上是一致（agree on）的。见图 4.5。因此，我们把在有限集 $\{\phi_1, \ldots, \phi_{2\binom{n}{d}}\}$ 中最小化经验误差视为在无限线性分类器集上的近似最小化。在第 12 和 13 章中，我们将为经验损失最小化中的规则开发完全理论。定理4.5只是让你尝尝未来的味道。另外的证明——涉及更多，但也更具一般性——将参考 Vapnik 和 Chervonenkis（1971；1974c）。

定理 4.5 *假设 X 具有一个密度。若 $\widehat{\phi}$ 为使经验误差最小的分类器。则对任 (X,Y) 分布，若 $n \geq d$ 和 $2d/n \leq \epsilon \leq 1$，我们有*

$$\mathbf{P}\{L(\widehat{\phi}) > L + \epsilon\} \leq e^{2d\epsilon} \left(2\binom{n}{d} + 1\right) e^{-n\epsilon^2/2}$$

另外，若 $n \geq d$，则

$$\mathbf{E}\{L(\widehat{\phi}) - L\} \leq \sqrt{\frac{2}{n}((d+1)\log n + (2d+2))}$$

备注 4.3 注意，定理4.5和下面的定理4.6可以拓展，因此可以不使用密度假设。我们需要保证被选中的线性规则的经验误差接近（可达到的）最佳线性规则的误差。该属性对上述推出的分类器可能不成立，如果数据点不一定具有一般位置。该思想在第 12 章中进一步推广（见定理 12.2）。

证. [7] 我们先从如下不等式

7: 许有问题

$$\begin{aligned}
L(\widehat{\phi}) - L =& L(\widehat{\phi}) - \widehat{L}_n(\widehat{\phi}) + \widehat{L}_n(\widehat{\phi}) - L(\phi^*) \\
\leq& L(\widehat{\phi}) - \widehat{L}_n(\widehat{\phi}) + \widehat{L}_n(\phi^*) - L(\phi^*) + \frac{d}{n} \\
& \left(\text{由于} \widehat{L}_n(\widehat{\phi}) \leq \widehat{L}_n(\phi) + d/n \text{ 对任} \phi\right) \\
\leq& \max_{i=1,\ldots,2\binom{n}{d}} \left(L(\phi_i) - \widehat{L}_n(\phi_i)\right) + \widehat{L}_n(\phi^*) - L(\phi^*) + \frac{d}{n}.
\end{aligned}$$

因此，通过联合事件界（联合界），我们有

$$\begin{aligned}
&\mathbf{P}\{L(\widehat{\phi}) - L > \epsilon\} \\
&\leq \sum_{i=1}^{2\binom{n}{d}} \mathbf{P}\left\{L(\phi_i) - \widehat{L}_n(\phi_i) > \tfrac{\epsilon}{2}\right\} + \mathbf{P}\left\{\widehat{L}_n(\phi^*) - L(\phi^*) + \tfrac{d}{n} > \tfrac{\epsilon}{2}\right\}.
\end{aligned}$$

为了给出右边第二项的界，观察到 $n\widehat{L}_n(\phi^*)$ 是参数为 n 和 $L(\phi^*)$ 的二项分布。通过 Chernoff（1952）和 Okamoto（1958）中对二项尾分布的不等式，

$$\mathbf{P}\left\{\widehat{L}_n(\phi^*) - L(\phi^*) > \frac{\epsilon}{2} - \frac{d}{n}\right\} \leq e^{-2n\left(\frac{\epsilon}{2} - \frac{d}{n}\right)^2} \leq e^{2d\epsilon} \cdot e^{-n\epsilon^2/2}.$$

我们稍后证明该不等式（见定理**??**）。接下来我们给出右边和式的界。

注意由于对称性，和式中所有 $2\binom{n}{d}$ 项是相等的。假设分类器 ϕ_1 取决于前 d 个数据点构成的 d 元组 $X_1, \ldots, X_d$。我们记

$$\mathbf{P}\left\{L(\phi_1) - \widehat{L}_n(\phi_1) > \frac{\epsilon}{2}\right\} = \mathbf{E}\left\{\mathbf{P}\left\{L(\phi_1) - \widehat{L}_n(\phi_1) > \frac{\epsilon}{2} \mid X_1, \ldots, X_d\right\}\right\},$$

并给出里面条件概率的界。令 $\left(X_1'', Y_1''\right), \ldots, \left(X_d'', Y_d''\right)$ 为独立数据且与数据 $(X_1, Y_1), \ldots, (X_d, Y_d)$ 的分布相同。定义

$$\left(X_i', Y_i'\right) = \begin{cases} \left(X_i'', Y_i''\right) & \text{if } i \le d \\ (X_i, Y_i) & \text{if } i > d. \end{cases}$$

则

$$\begin{aligned}
&\mathbf{P}\left\{L(\phi_1) - \widehat{L}_n(\phi_1) > \frac{\epsilon}{2} \mid X_1, \ldots, X_d\right\} \\
&\le \mathbf{P}\left\{L(\phi_1) - \frac{1}{n}\sum_{i=d+1}^{n} I_{\{\phi_1(X_i) \ne Y_i\}} > \frac{\epsilon}{2} \mid X_1, \ldots, X_d\right\} \\
&\le \mathbf{P}\left\{L(\phi_1) - \frac{1}{n}\sum_{i=1}^{n} I_{\{\phi_1(X_i') \ne Y_i'\}} + \frac{d}{n} > \frac{\epsilon}{2} \mid X_1, \ldots, X_d\right\} \\
&= \mathbf{P}\left\{L(\phi_1) - \frac{1}{n}\,\text{Binomial}(n, L(\phi_1)) > \frac{\epsilon}{2} - \frac{d}{n} \mid X_1, \ldots, X_d\right\} \\
&\quad (\text{由于} L(\phi_1) \text{仅依赖于} X_1, \ldots, X_d \text{和} \left(X_1', Y_1'\right) \ldots, \left(X_d', Y_d'\right) \\
&\quad \text{独立于 } X_1, \ldots, X_d) \\
&\le e^{-2n\left(\frac{\epsilon}{2} - \frac{d}{n}\right)^2} \\
&\quad (\text{通过定理 8.1；使用} \epsilon \ge 2d/n) \\
&\le e^{-n\epsilon^2/2} e^{2d\epsilon}
\end{aligned}$$

期望值的不等式可由如下论述的概率不等式得到：通过 Cauchy-Schwarz 不等式，

$$(\mathbf{E}\{L(\widehat{\phi}) - L\})^2 \le \mathbf{E}\left\{(L(\widehat{\phi}) - L)^2\right\}.$$

记 $Z = (L(\widehat{\phi}) - L)^2$，对任 $u > 0$，

$$\begin{aligned}
\mathbf{E}\{Z\} &= \mathbf{E}\{Z \mid Z > u\}\mathbf{P}\{Z > u\} + \mathbf{E}\{Z \mid Z \le u\}\mathbf{P}\{Z \le u\} \\
&\le \mathbf{P}\{Z > u\} + u \\
&\le e^{2d}\left(2n^d + 1\right)e^{-nu/2} + u \quad \text{若} u \ge \left(\frac{2d}{n}\right)^2
\end{aligned}$$

通过使用概率不等式和 $\binom{n}{d} \le n^d$。选择 u 来最小化得到的表达式，能推出所希望的不等式：首先验证最小值为

$$u = \frac{2}{n}\log\frac{nc}{2},$$

其中 $c = e^{2d}\left(2n^d + 1\right)$。如果 $n \ge d$，则 $u \ge (2d/n)^2$。因此注意到界 $ce^{-nu/2} + u$ 等于

$$\frac{2}{n}\log\frac{nec}{2} \le \frac{2}{n}\log\left(e^{2d+2}n^{d+1}\right) = \frac{2}{n}((d+1)\log n + (2d+2)).$$

得证。 [证毕]

现在观察到 $\mathbf{P}\{L(\widehat{\phi}) > L + \epsilon\}$ 的界随 n 的增加急剧下降。要产生影响，n 必须小于小的 δ。粗略地说，对某些常数 c，

$$n \geq c \cdot \frac{d}{\epsilon^2}\left(\log\frac{d}{\epsilon^2} + \log\frac{1}{\delta}\right)$$

应成立。若维度 d 翻倍，则该最小样本大小也应翻一倍。

一个重要的特例时当分布为线性可分时，即 $L = 0$。在该情况下，上述的经验风险最小化会表现得更好，因为误差的大小从 $O(\sqrt{d\log n/n})$ 变成 $O(d\log n/n)$。显然，数据点也是线性可分的，即 $\widehat{L}_n(\phi^*) = 0$ 以概率 1 成立，因此 $\widehat{L}_n(\widehat{\phi}) \leq d/n$ 以概率 1 成立。

定理 4.6 *假设 X 具有一个密度，且最优的线性分类器误差具零概率 ($L = 0$)。则对定理4.5的经验风险最小化算法，任 $n > d$ 和任 $\epsilon \leq 1$，*

$$\mathbf{P}\{L(\widehat{\phi}) > \epsilon\} \leq 2\binom{n}{d}e^{-(n-d)\epsilon},$$

和

$$\mathbf{E}\{L(\widehat{\phi})\} \leq \frac{d\log n + 2}{n - d}$$

成立。

证. 通过联合界，

$$\begin{aligned}\mathbf{P}\{L(\widehat{\phi}) > \epsilon\} &\leq \mathbf{P}\left\{\max_{i=1,2,\ldots,2\binom{n}{d}:\widehat{L}_n(\phi_i)\leq\frac{d}{n}} L(\phi_i) > \epsilon\right\} \\ &\leq \sum_{i=1}^{2\binom{n}{d}} \mathbf{P}\left\{\widehat{L}_n(\phi_i) \leq \frac{d}{n}, L(\phi_i) > \epsilon\right\}.\end{aligned}$$

由对称性，这个和式等于

$$\begin{aligned}&2\binom{n}{d}\mathbf{P}\left\{\widehat{L}_n(\phi_1) \leq \frac{d}{n}, L(\phi_1) > \epsilon\right\} \\ &= 2\binom{n}{d}\mathbf{E}\left\{\mathbf{P}\left\{\widehat{L}_n(\phi_1) \leq \frac{d}{n}, L(\phi_1) > \epsilon \mid X_1, \ldots, X_d\right\}\right\}\end{aligned}$$

其中，如在定理4.5，ϕ_1 取决于数据点 $X_1, \ldots, X_d$。但是，

$$\begin{aligned}&\mathbf{P}\left\{\widehat{L}_n(\phi_1) \leq \frac{d}{n}, L(\phi_1) > \epsilon \mid X_1, \ldots, X_d\right\} \\ &\leq \mathbf{P}\left\{\phi_1(X_{d+1}) = Y_{d+1}, \ldots, \phi_1(X_n) = Y_n, L(\phi_1) > \epsilon \mid X_1, \ldots, X_d\right\} \\ &\text{(由于所有的（至多}d\text{个）误差均由}\phi_1\text{造成} \\ &\text{，在 } (X_1, Y_1), \ldots, (X_d, Y_d) \text{ 上)} \\ &\leq (1-\epsilon)^{n-d},\end{aligned}$$

由于所有 (X_i, Y_i) 均不落在集合 $\{(x, y) : \phi_1(x) \neq y\}$ 的概率小于 $(1 - \epsilon)^{n-d}$ 若该集合的概率大于 ϵ。概率不等式证明可通过 $1 - x \leq e^{-x}$ 完成

问:

。对期望误差概率，注意到对任 $u > 0$,

$$\begin{aligned}
\mathbf{E}\{L(\widehat{\phi})\} &= \int_0^\infty \mathbf{P}\{L(\widehat{\phi}) > t\}dt \\
&\le u + \int_u^\infty \mathbf{P}\{L(\widehat{\phi}) > t\}dt \\
&\le u + 2n^d \int_u^\infty e^{-(n-d)t}\,dt \\
&\quad \text{(通过概率不等式与 } \binom{n}{d} \le n^d\text{)} \\
&= u + \frac{2n^d}{n} e^{-(n-d)u}.
\end{aligned}$$

我们选择 u 以最小化获得的界，可以推出所求的不等式。 [证毕]

4.6 最小化其他准则

经验风险最小化需要大量的计算，因为 $\widehat{L}_n(\phi)$ 一般来说不是单峰函数（见问题 4.10 和 4.11）。另外，由于梯度几乎处处为零，梯度优化是相当困难的。事实上，给定 n 个已打标签的 $\mathcal{R}^d$ 上的点，找到最优线性二分（分类器）是 NP 难的（见 Johnson 和 Preparata（1978））。为帮助优化，有些研究者建议最小化经修改的经验误差，例如

$$\widehat{L}_n(\phi) = \frac{1}{n}\sum_{i=1}^n \Psi\left((2Y_i - 1) - a^T X_i - a_0\right) I_{\{Y_i \ne g_a(X_i)\}}$$

或

$$\widehat{L}_n(\phi) = \frac{1}{n}\sum_{i=1}^n \Psi\left((2Y_i - 1) - a^T X_i - a_0\right),$$

其中 Ψ 为正凸函数。这里重要的是均方误差准则 $\Psi(u) = u^2$（见 Widrow 和 Hoff（1960））。容易验证 $\widehat{L}_n(\phi)$ 具有一个关于（a, a_0）梯度，其有助于定位局部极小点。令 $\widehat{\phi}$ 记为最小化

$$\mathbf{E}\left\{\left((2Y - 1) - a^T X - a_0\right)^2\right\}$$

的线性判别规则，对任 a 和 a_0。问题 4.14 给出了解。

即使在一维情况，均方误差准则也会使问题变得模糊不清，并且也没有给出性能保证：

定理 4.7 *若 $sup_{(X,Y)}$ 记为 $\mathcal{R} \times \{0,1\}$ 上关于所有分布的上确界，则*

$$\sup_{(X,Y)}(L(\widehat{\phi}) - L) = 1,$$

其中 $\widehat{\phi}$ 为通过最小化在所有 a_1 和 a_0 上的

$$\mathbf{E}\left\{((2Y - 1) - a_1 X - a_0)^2\right\}$$

得到的线性判别。

备注 4.4 该定理建立了 $L(\widehat{\phi}) > 1-\epsilon$ 和 $L < \epsilon$ 对任意小的 $\epsilon > 0$ 同时成立时 (X,Y) 分布的存在性。因此除非有更多关于 (X,Y) 分布的信息，否则并不推荐最小化均方误差准则。

证. 令 $\epsilon > 0$ 和 $\theta > 0$。考虑 (X,Y) 的三元分布：

$$\mathbf{P}\{(X,Y) = (-\theta,1)\} = \mathbf{P}\{(X,Y) = (1,1)\} = \epsilon/2,$$

$$\mathbf{P}\{(X,Y) = (0,0)\} = 1-\epsilon.$$

对 $\epsilon < 1/2$，最优的线性规则将 $[-\theta/2,\infty)$ 上的点判定为类 0，其他点判为类 1，误差概率为 $L = \epsilon/2$。均方误差准则要求我们关于 $a_0 = v$ 和 $a_1 = u$ 最小化

$$L(\widehat{\phi}) = \left\{(1-\epsilon)(-1-v)^2 + \frac{\epsilon}{2}(1-u-v)^2 + \frac{\epsilon}{2}(1+u\theta-v)^2\right\}.$$

令关于 u 和 v 的导数为零可求

$$u = \frac{(v-1)\theta - v}{1+\theta^2}, \quad \text{和} \quad v = 2\epsilon - 1 + \frac{\epsilon}{2}u(\theta-1),$$

对

$$v = \frac{(2\epsilon-1)\left(1+\theta^2\right) - \frac{\epsilon}{2}\theta(\theta-1)}{1+\theta^2 - \frac{\epsilon}{2}(1-\theta)^2}.$$

若我们令 $\epsilon \downarrow 0$ 和 $\theta \uparrow \infty$，则 $v \sim 3\epsilon/2$。因此，对足够小的 ϵ 和足够大的 θ，仅考虑对 0 的决策，$L(\widehat{\phi}) \geq 1-\epsilon$，因为在 $x = 0$，$ux + v = v > 0$。因此，$L(\widehat{\phi}) - L \geq 1 - 3\epsilon/2$ 成立，对足够小的 ϵ 和足够大的 θ。 [证毕]

有其他研究者建议最小化

$$\sum_{i=1}^{n}\left(\sigma\left(a^T X_i + a_0\right) - Y_i\right)^2,$$

其中 $\sigma(u)$ 是 Sigmoid 函数，即从 0 到 1 的递增函数，例如 $1/(1+e^{-u})$，见 Wassel 和 Sklansky（1972）、Do Tu 和 Installe（1975）、Fritz 和 Györfi（1976）与 Sklansky 和 Wassel（1979）。显然，$\sigma(u) = I_{\{u\geq 0\}}$ 提供了经验误差概率。但是，这里的关键是使用光滑 Sigmoid 函数，因此梯度算法可被用来寻找最优点。其可被视为均方误差准则和经验误差最小化的一种权衡。同时异常的情况也可能发生，误差空间（error space）表现不好，具有许多局部极小点（Hertz，Krogh 和 Palmer（1991，p.108））。但是，见问题 4.16 和 4.17。

问：误差空间是什么？

4.7 问题与习题

问题 4.1

使用定理4.1的概念，证明一个一维理论 Stoller 切割的误差概率 L 满足

$$L \leq \frac{4p(1-p)}{1+p(1-p)\frac{(m_0-m_1)^2}{(1-p)\sigma_0^2+p\sigma_1^2}}.$$

（Györfi 和 Vajda（1980））

其是否比定理4.1的界更好？[提示：对任阈值规则 $g_c(x) = I_{\{x \geq c\}}$ 和 $u > 0$，记

$$\begin{aligned} L(g_c) &= \mathbf{P}\{X - c \geq 0, 2Y - 1 = -1\} + \mathbf{P}\{X - c < 0, 2Y - 1 = 1\} \\ &\leq \mathbf{P}\{|u(X - c) - (2Y - 1)| \geq 1\} \\ &\leq \mathbf{E}\left\{(u(X - c) - (2Y - 1))^2\right\} \end{aligned}$$

通过 Chebyshev 不等式。选择 u 和 c 以最小化上界。]

问题 4.2

令 $p = 1/2$。若 L 为一维理论 Stoller 切割的误差概率，证明

$$L \leq \frac{1}{2 + 2\frac{(m_0 - m_1)^2}{(\sigma_0 + \sigma_1)^2}}.$$

当 X 的类条件分布（即给定 $Y = 0$ 和 $Y = 1$）分别集中在两个点上时，其中一个点为两个类共有（Chernoff（1971），Becker（1968）），证明存在一些分布可达到该界。[8]

8: 许有问题：Show that the bound is achieved for some distribution when the class-conditional distributions of X (that is, given $Y = 0$ and $Y = 1$) are concentrated on two points each, one of which is shared by both classes (Chernoff (1971), Becker (1968).

问题 4.3

令 X 为单元随机变量。给定 $Y = 1$ 和 $Y = 0$ 时 X 的分布函数分别为 F_1 和 F_0。假设 X 的矩母函数存在，即 $\mathbf{E}\left\{e^{tX} \mid Y = 1\right\} = \psi_1(t), \mathbf{E}\left\{e^{tX} \mid Y = 0\right\} = \psi_0(t), t \in \mathcal{R}$，其中 ψ_1, ψ_0 对所有 t 均是有限的。按定理4.1的思想，为 L 推导出一个关于 ψ_1, ψ_0 的上界。应用该界到 F_1 和 F_0 均为正态分布的情况，其中该正态分布可能具有不同的均值和方差。

问题 4.4

加性高斯噪声中的信号

令 $s_0, s_1 \in \mathcal{R}^d$ 固定，并令 N 为具零均值和协方差矩阵 Σ 的多元高斯随机变量。另外使 $\mathbf{P}\{Y = 0\} = \mathbf{P}\{Y = 1\} = 1/2$，定义

$$X = \begin{cases} s_0 + N & \text{if } Y = 0 \\ s_1 + N & \text{if } Y = 1 \end{cases}$$

构造 Bayes 决策并计算 L^*。证明若 Σ 为单位矩阵，且 s_0 和 s_1 具常数分量，则随 $d \to \infty$，$L^* \to 0$ 按指数收敛。

问：“具”表示可能全部分量均为常数，也可能是部分分量为常数。

问题 4.5

在定理4.2的证明中的最后一个步骤，我们使用 Dvoretzky-Kiefer-Wolfowitz-Massart 不等式（定理**??**）。该结果表明若 $Z_1, \ldots, Z_n$ 是实数线上独立同分布（i.i.d）的随机变量，其分布函数为 $F(z) =$

$\mathbf{P}\{Z_1 \le z\}$ 和经验分布函数 $F_n(z) = (1/n)\sum_{i=1}^{n} I_{\{Z_i \le z\}}$，则

$$\mathbf{P}\left\{\sup_{z\in\mathcal{R}} |F(z) - F_n(z)| \ge \epsilon\right\} \le 2e^{-2n\epsilon^2}.$$

使用该不等式证明

$$\mathbf{P}\left\{\sup_{x} |v(C(x,1)) - v_n(C(x,1))| \ge \frac{\epsilon}{2}\right\} \le 2e^{-2n(\epsilon/2)^2}.$$

[提示：通过双射函数 $\psi : (\mathcal{R}\times\{0,1\}) \to \mathcal{R}$ 映射实数线上的点 (X,Y)，使得 $Z = \psi((X,Y)) < 0$ 当且仅当 $Y = 0$。对 Z 使用 Dvoretzky-Kiefer-Wolfowitz-Massart 不等式。]

问题 4.6

令 L 为最优球规则的误差概率，其中球规则将在球 $S_{x,r}$ 中的点划分至其中一个类，将球外的点划分至另一个类。记 x 表示球心，r 为半径。证明 $L = 1/2$ 当且仅当 $L^* = 1/2$，且 $L \le 1/2$ 成立。

问题 4.7

使用定理4.4的概念，证明最优线性判别的误差概率 L 满足

$$L \le \frac{4p(1-p)}{1 + p(1-p)\Delta^2},$$

其中

$$\Delta = \sqrt{(m_1 - m_0)^T \Sigma^{-1} (m_1 - m_0)}$$

为 Mahalanobis 距离（第 3 章），且 $\Sigma = p\Sigma_1 + (1-p)\Sigma_0$（Györfi 和 Vajda（1980））。有趣地是，上界正好是定理3.4的渐进最近邻误差界的两倍。因此，一个大的 Mahalanobis 距离不仅意味着 Bayes 误差是小的，而且简单线性分类器可以得到小的误差概率。

[提示：对一元随机变量 $X' = a^T X a = \Sigma^{-1}(m_1 - m_0)$ 应用问题4.1的不等式。]

问题 4.8

若 m_i 和 σ_i^2 分别为变量 $a^T X$ 的均值和变量，给定 $Y = i, i = 0,1$，其中 a 权重的列向量，则证明准则

$$J_1(a) = \frac{(m_1 - m_0)^2}{\sigma_1^2 + \sigma_0^2}$$

在 $a = (\Sigma_1 + \Sigma_0)^{-1}(M_1 - M_0)$ 处达到最小值，其中 M_i 和 Σ_i 为给定 $Y = i$ 时 X 的均值向量和协方差。另外，证明

$$J_2(a) = \frac{(m_1 - m_0)^2}{p\sigma_1^2 + (1-p)\sigma_0^2}$$

在 $a = (p\Sigma_1 + (1-p)\Sigma_0)^{-1}(M_1 - M_0)$ 处得到最小值，其中 $p, 1-p$ 为类概率。这个练习证明若在一维情况进行判别，我们可以考虑投影 $a^T X$，其中 a 最大化投影均值之间的权重距离。

问题 4.9

在具自由参数 a_0 的 Fisher 线性判别规则（4.1），证明对任 $\epsilon > 0$，存在 (X,Y) 的分布，$X \in \mathcal{R}^2$，[9] 具 $L = 0$ 与 $\mathbf{E}\{\|X\|^2\} < \infty$，使得 $\inf_{a_0} \mathbf{E}\{L(g_{a_0})\} > 1/2 - \epsilon$。另外，若令 a_0 最小化均方误差

$$\mathbf{E}\left\{\left((2Y-1) - a^T X - a_0\right)^2\right\},$$

则 $\mathbf{E}\{L(g_{a_0})\} > 1 - \epsilon$。

9: 许有问题

问题 4.10

找到 (X,Y) 的一个分布使得至少有一半的概率，$\widehat{L}_n(\phi)$ 不是关于权重向量 (a, a_0) 的单峰函数。

问题 4.11

如下的观察将有助于开发一个快速的算法以找到在某些情况下最优的线性分类器。假设贝叶斯规则为一个过原点的线性切割，即 $L^* = L(a^*)$ 对某个系数向量 $a^* \in \mathcal{R}^d$，其中 $L(a)$ 记为分类器

$$g_a(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^d a_i x^{(i)} \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

的误差概率，且 $a = (a_1, \ldots, a_d)$。证明 $L(a)$ 为 $a \in \mathcal{R}^d$ 的单峰函数，且 $L(a)$ 沿着 a^* 方向单调递增，即对任 $\lambda \in (0,1)$ 和 $a \in d$，$L(a) - L(\lambda a + (1-\lambda)a^*) \geq 0$（Fritz 和 Györfi（1976））。[提示：使用表达式

$$L(a) = 1/2 - \int (\eta(x) - 1/2)\,\mathrm{sign}\left(\sum_{i=1}^d a_i x^{(i)}\right)\mu(dx)$$

证明存在集合 $A \subset \mathcal{R}^d$ 使得 $L(a) - L(\lambda a + (1-\lambda)a^*) = \int_A |\eta(x) - 1/2|\mu(dx)$ 成立。]

问题 4.12

令 $a = (a_0, a_1)$ 与

$$\widehat{a} = \arg\min_a \mathbf{E}\left\{((2Y-1) - a_1 X - a_0)^2 I_{\{Y_i \neq g_a(X_i)\}}\right\},$$

与 $g_a(x) = I_{\{a_1 x + a_0 > 0\}}$。证明对任 $\epsilon > 0$，存在一个位于 $\mathcal{R} \times \{0,1\}$ 上的 (X,Y) 分布，使得 $L(\widehat{a}) - L \geq 1 - \epsilon$，其中 $L(\widehat{a})$ 为 $g_{\widehat{a}}$ 的误差概率。[提示：类似于在定理4.7证明中的论证。一个四元分布可以满足。]

问题 4.13

令

$$\widehat{a} = \arg\min_{a} \mathbf{E}\left\{\left|(2Y-1) - a_1 X - a_0\right|\right\},$$

重复论证上一道习题。

问题 4.14

令 ϕ^* 记为在所有 a, a_0 上最小化均方误差 $\mathbf{E}\left\{(2Y-1-a^T X - a_0)^2\right\}$ 的线性判别规则。当这一准则在 (a, a_0) 上是二次的,则它是单峰的。人们经常用 $\widehat{\phi}$ 近似 ϕ^*,即在所有 a, a_0 上最小化 $\sum_i (2Y_i - 1 - a^T X_i - a_0)^2$。证明最小列向量 (a, a_0) 由

$$\left(\sum_i X_i' X_i'^T\right)^{-1} \left(\sum_i (2Y_i - 1) X_i'\right)$$

给出，其中 $X_i' = (X_i, 1)$ 为 $(d+1)$ 维列向量。

问题 4.15

感知器准则为

$$J = \sum_{i:2Y_i - 1 \neq \operatorname{sign}\left(a^T X_i + a_0\right)} \left|a^T X_i + a_0\right|.$$

找到一个分布使得 $L^* = 0, L \leq 1/4$，但 $\liminf_{n\to\infty} \mathbf{E}\{L_n(\phi)\} \geq 1/2$，其中 ϕ 为使用 a, a_0 最小化 J 得到的线性判别规则。

问题 4.16

令 σ 为 $\mathcal{R}$ 上的单调非减函数, 满足 $\lim_{u\to-\infty} \sigma(u)$=0 和 $\lim_{u\to\infty} \sigma(u)$ = 1。对 $h > 0$，定义 $\sigma_h(u) = \sigma(hu)$。考虑选择参数 a, a_0，使对应的线性判别规则 $\widehat{\phi}$ 最小化

$$\sum_{i=1}^{n} \left(\sigma_h\left(a^T X_i + a_0\right) - Y_i\right)^2.$$

对每一固定的 $h > 0$ 和 $0 < \epsilon < 1$，找到一个分布使得 $L < \epsilon$ 和

$$\liminf_{n\to\infty} \mathbf{E}\left\{L_n(\widehat{\phi})\right\} > 1 - \epsilon.$$

另一方面, 证明若 h 依赖于样本大小 n, 使得当 $n \to \infty$ 时有 $h \to \infty$, 则对所有分布，$\mathbf{E}\left\{L_n(\widehat{\phi})\right\} \to L$ 成立。

问题 4.17

给定 $Y = i$，令 X 为均值 m_i 和协方差矩阵 Σ_i 的正态分布，$i = 0, 1$。

考虑基于最小化准则

$$\mathbf{E}\left\{\left(Y-\sigma\left(X^TAX+w^TX+c\right)\right)^2\right\}$$

的判别，该准则是关于变量 A, w 和 c 的函数，分别是 $d\times d$ 矩阵，$d\times 1$ 向量和常量。其中 $\sigma(u)=1/(1+e^{-u})$ 为标准 sigmoid 函数。证明最小化误差概率

$$\mathbf{P}\left\{2Y-1\neq \operatorname{sign}\left(X^TAX+w^TX+c\right)\right\}$$

的 A, w 和 c 使得准则达到最小值，并验证在该情况下，均方误差准则可以被用来获得 Bayes-最优分类器（Horne 和 Hush（1990））。

最近邻规则 5

5.1 引言

简单规则依然存在。k-最近邻规则，自从其于 1951 年和 1952 年开始构思（Fix 和 Hodges（1951；1952；1991a；1991b)），已经吸引了许多追随者并持续被研究。正式的说，我们定义k-NN 规则为

$$g_n(x) = \begin{cases} 1 & \text{若 } \sum_{i=1}^n w_{ni} I_{\{Y_i=1\}} > \sum_{i=1}^n w_{ni} I_{\{Y_i=0\}} \\ 0 & \text{否则,} \end{cases}$$

其中若 X_i 位于 x 的 k 个最近邻中，则 $w_{ni} = 1/k$，否则 $w_{ni} = 0$。X_i 被称为 x 的 k 最近邻，若距离 $\|x - X_i\|$ 在 $\|x - X_1\|, \ldots, \|x - X_n\|$ 中排第 k 小的。若距离出现相等的情况，则认为具更小下标的 X_i 与 x 更近。决策取决于多数投票。令 k 为奇数是方便的，可以避免出现正反票数相同的情况。下述几个问题值得讨论：

(A) 普遍一致性。若当 $n \to \infty$ 时有 $k \to \infty$ 和 $k/n \to 0$，建立至贝叶斯规则的收敛性。将在第 11 章处理。
(B) 有限 k 的性能。若我们固定 k 并令 n 趋向于无穷大，则情况如何?
(C) 权重向量 $(w_{n1}, \ldots, w_{nn})$ 的选择。具相同权重的 k 最近邻是否某种意义上比不同权重更优?
(D) 距离度量的选择。实现关于某种变换族（family）下的不变性。
(E) 数据大小的减少。当数据集大小被编辑或减少以降低存储加载时，我们是否可以获得良好的精度?

在前几节中，我们关心当 k 不随着 n 变化时，k 最近邻规则的收敛性问题。尤其是，我们将看到对所有分布，期望误差概率 $\mathbf{E}\{L_n\}$ 趋向于极限 $L_{k\text{NN}}$，该极限一般很接近但大于 L^*。获得该结果的方法就其本身而言是相当有趣的。$L_{k\text{NN}}$ 的表达式将随其后被讨论，几个关键不等式如 $L_{NN} \le 2L^*$（Cover 和 Hart（1967））与 $L_{k\text{NN}} \le L^*(1 + \sqrt{2/k})$ 将被证明与应用。其他上述提到的问题在剩余的几节中处理。对各个方面最近邻与相关方法的综述可见 Dasarathy（1991），Devijver（1980）或 Devroye 和 Wagner（1982）。

备注 5.1 *计算问题。*在数组存储 n 个数据对并搜索 k 最近邻，若不经过优化将花费与 nkd 成比例的时间——d 代表一次距离计算所耗费的时间。通过这三个因子的其中一个或多个，可降低该复杂度。一般来说，固定 k 和 d，最坏时间 $O\left(n^{1/d}\right)$（Papadimitriou 和 Bentley（1980））和期望时间 $O(\log n)$（Friedman，Bentley 和 Finkel（1977））是可达到的。切割空间与引导搜索的多维搜索树是相当有用的——对该方法，见 Fukunage 和 Narendra（1975），Friedman，Bentley 和 Finkel（1977），Niemann 和 Goppert（1988），Kim 和 Park（1986）和 Broder（1990）。我们引用了 Dasarathy（1991）的综述以获得更多参考文献。其他方法可见于 Yunck（1976），Friedman，Baskett 和 Shustek（1975），Vidal（1986），Sethi（1981）和 Faragó，Linder 和 Lugosi（1993）。一般来说，使用预处理，我们能大大减少与 n 和 d 相关的总复杂度。

5.2 概念与简单渐进性

我们固定 $x \in \mathscr{R}^d$，并根据 $\|X_i - x\|$ 的值按升序重排数据 (X_1, Y_1)，$\ldots, (X_n, Y_n)$。若不出现混淆情况，重排后的数据序列记为

$$\left(X_{(1)}(x), Y_{(1)}(x)\right), \ldots, \left(X_{(n)}(x), Y_{(n)}(x)\right) \text{ 或通过 } \left(X_{(1)}, Y_{(1)}\right), \ldots, \left(X_{(n)}, Y_{(n)}\right).$$

$X_{(k)}(x)$ 为 x 的 k 最近邻。

备注 5.2 这里我们注意，我们根据欧氏距离 $\|x - y\|$ 而不是任意地定义*邻居*。令人惊讶地是，本章推导地渐进性对很多度量均有效——渐进误差概率与距离度量无关（见问题5.1）。

记 μ 为 X 的概率测度，并令 $S_{x,\epsilon}$ 为以 x 为圆心，半径为 $\epsilon > 0$ 的闭球。所有满足 $\mu\left(S_{x,\epsilon}\right) > 0, \forall \epsilon > 0$ 的 x 的集合称为 X 或 μ 的*支撑*（support）。由于如下性质该集合相当重要。

引理 5.1 *若 $x \in support(u)$ 和 $\lim_{n\to\infty} k/n = 0$，则 $\left\|X_{(k)}(x) - x\right\| \to 0$ 以概率 1 成立。若 X 独立于数据且具概率测度 μ，则当 $k/n \to 0$，$\left\|X_{(k)}(X) - X\right\| \to 0$ 以概率 1 成立。*

证. 取 $\epsilon > 0$。通过定义，$x \in support(\mu)$ 蕴含着 $\mu(S_{x,\epsilon}) > 0$。观察到 $\left\|X_{(k)}(x) - x\right\| > \epsilon$ 当且仅当

$$\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \in S_{x,\epsilon}\}} < \frac{k}{n}.$$

通过强大数定律，不等式左边收敛至 $\mu(S_{x,\epsilon}) > 0$ 以概率 1 成立，然而由假设，右边趋于 0。因此 $\left\|X_{(k)}(x) - x\right\| \to 0$ 以概率 1 成立。

第二个证明也能按上述论证得到。首先注意到由附录中引理 A.1，$\mathbf{P}\{X \in \text{support}(\mu)\} = 1$，因此对任 $\epsilon > 0$，由控制收敛定理

$$\begin{aligned}&\mathbf{P}\left\{\left\|X_{(k)}(X) - X\right\| > \epsilon\right\}\\ &\quad = \mathbf{E}\left\{I_{\{X \in \text{ support }(\mu)\}}\mathbf{P}\left\{\left\|X_{(k)}(X) - X\right\| > \epsilon \mid X \in \text{support}(\mu)\right\}\right\}\end{aligned}$$

收敛至零点。证明了概率的收敛性。若 k 不随着 n 而改变，则对 $n \geq k$，$\left\|X_{(k)}(X) - X\right\|$ 单调递减；因此，它以概率 1 收敛。若允许 $k = k_n$ 随着 n 的增长而增长，使得 $k/n \to 0$，则使用概念 $X_{(k_n,n)}(X) = X_{(k)}(X)$，通过类似的论证，单调递增的随机变量序列

$$\sup_{m \geq n}\left\|X_{(k_m,m)}(X) - X\right\| \geq \left\|X_{(k_n,n)}(X) - X\right\|$$

概率收敛于 0，因此也以概率 1 成立。得证。 [证毕]

由于 η 是可测的（因此在一般意义下表现良好）与 $\left\|X_{(k)}(x) - x\right\|$ 是小的，故 $\eta\left(X_{(i)}(x)\right)$ 值应接近于 $\eta(x)$，对所有足够小的 i。我们现在介绍一个利用该事实的证明方法，其使得随后的分析变得非常简单——可以通过嵌入以一种新的方式观察数据样本[11] 。基本思想使定义一个额外的规则 g'_n，该规则将 $Y_{(i)}(x)$ 替换为 k 个独立同分布的参数为 $\eta(x)$ 的伯努利随机变量——按局部而言，$Y_{(i)}(x)$ 以该种方式出现（behave）。

11: 许有问题

容易证明这两个规则的误差概率使相近的，因此分析该附加规则的表现更加方便。

更加精确地说，假设我们给定独立同分布数据对 $(X_1,U_1),\ldots,(X_n,U_n)$，所有分布与 (X,U) 分布相同，其中 X 的概率测度 μ 定义在 $\mathscr{R}^d$ 的 Borel 集上，U 是 $[0,1]$ 上的均匀分布且与 X 相独立。若我们令 $Y_i = I_{\{U_i \le \eta(X_i)\}}$，则 $(X_1,Y_1),\ldots,(X_n,Y_n)$ 为独立同分布，其原型分布为 (X,Y)。那么何以为 U_i 的分布所烦扰？[12] 在嵌入参数中，我们将使用同一 U_i 来构建与原始数据序列高度相关（对偶）的第二个数据序列，该序列更加便于分析。例如，对固定的 $x \in \mathscr{R}^d$，我们定义

12: 许有问题

$$Y_i'(x) = I_{\{U_i \le \eta(x)\}}.$$

我们现在由一列独立同分布序列，其第 i 的向量记为 $X_i, Y_i, Y_i'(x), U_i$。根据 $\|X_i - x\|$ 的值按升序重排该数据序列得到新的序列，此时记第 i 个向量为 $X_{(i)}(x), Y_{(i)}(x), Y'_{(i)}(x), U_{(i)}(x)$。若不出现混淆，则参数（argument）$x$ 能被忽略。若对任 $n \ge k$，对某函数 ψ，g_n 具形式，

$$g_n(x) = \begin{cases} 1 & \text{若} \psi\left(x, Y_{(1)}(x),\ldots,Y_{(k)}(x)\right) > 0, \\ 0 & \text{否则,} \end{cases} \tag{5.1}$$

则称该规则为 *k-局部的*（local）。对于 $k-\mathrm{NN}$ 规则，我们有，例如

$$\psi\left(x, Y_{(1)},\ldots,Y_{(k)}\right) = \sum_{i=1}^{k} Y_{(i)}(x) - \frac{k}{2}.$$

换言之，g_n 在 x 的 k 个最近邻中取多数票，若票数相等，则视为类 0。

研究 g_n 因此变得等价于研究近似规则 g_n'：

$$g_n'(x) = \begin{cases} 1 & \text{若} \psi\left(x, Y'_{(1)}(x),\ldots,Y'_{(k)}(x)\right) > 0 \\ 0 & \text{否则.} \end{cases}$$

后一规则不具实用价值，因为它需要 $\eta(x)$ 的知识。但有趣地是它更容易求解，因为 $Y'_{(1)}(x),\ldots,Y'_{(k)}(x)$ 均为独立同分布的，而 $Y_{(1)}(x),\ldots,Y_{(k)}(x)$ 不是。应特别注意：

引理 5.2 *对任 x，$n \ge k$，*

$$\mathbf{P}\left\{\psi\left(x, Y_{(1)}(x),\ldots,Y_{(k)}(x)\right) \ne \psi\left(x, Y'_{(1)}(x),\ldots,Y'_{(k)}(x)\right)\right\} \le \sum_{i=1}^{k} \mathbf{E}\left\{\left|\eta(x) - \eta\left(X_{(i)}(x)\right)\right|\right\}$$

和

$$\mathbf{P}\left\{g_n(x) \ne g_n'(x)\right\} \le \sum_{i=1}^{k} \mathbf{E}\left\{\left|\eta(x) - \eta\left(X_{(i)}(x)\right)\right|\right\}.$$

证. 这两个表达式均可由如下观察直接推出：

$$\begin{aligned}
&\{\psi\ (x, Y_{(1)}(x), \ldots, Y_{(k)}(x)) \neq \psi\left(x, Y'_{(1)}(x), \ldots, Y'_{(k)}(x)\right)\} \\
&\quad \subseteq \left\{\left(Y_{(1)}(x), \ldots, Y_{(k)}(x)\right) \neq \left(Y'_{(1)}(x), \ldots, Y'_{(k)}(x)\right)\right\} \\
&\quad \subseteq \bigcup_{i=1}^{k} \left\{\eta\left(X_{(i)}(x)\right) \leq U_{(i)}(x) \leq \eta(x)\right\} \\
&\qquad \cup \bigcup_{i=1}^{k} \left\{\eta(x) \leq U_{(i)}(x) \leq \eta\left(X_{(i)}(x)\right)\right\},
\end{aligned}$$

使用联合界，并基于 $U_{(i)}(x)$ 是在 $[0,1]$ 上的均匀分布这一事实。 [证毕]

我们需要下述结果，其中 X 的分布如 X_1，但与数据序列相独立：

Stone（1977）

引理 5.3 *对任意可积函数 f，任意 n 和任意 $k \leq n$，*

$$\sum_{i=1}^{k} \mathbf{E}\left\{\left|f\left(X_{(i)}(X)\right)\right|\right\} \leq k\gamma_d \mathbf{E}\{|f(X)|\},$$

其中 $\gamma_d \leq (1 + 2/\sqrt{2-\sqrt{3}})^d - 1$ 仅依赖于维数 d。

该引理的证明是美妙的，但需要一点技术——其在单独一节中给出。这里仅说明如何应用它，以及为什么，对固定的 k，我们实际上将 $f\left(X_{(k)}(X)\right)$ 视为 $f(X)$。

引理 5.4 *对任意可积函数 f，*

$$\frac{1}{k}\sum_{i=1}^{k} \mathbf{E}\left\{\left|f(X) - f\left(X_{(i)}(X)\right)\right|\right\} \to 0$$

当 $n \to \infty$，只要 $k/n \to 0$。

证. 给定 $\epsilon > 0$，找到一个一致连续函数 g，其在有界集 A 上为 0，使得 $\mathbf{E}\{|g(X) - f(X)|\} < \epsilon$（见附录中的定理 A.8）。则对任 $\epsilon > 0$，存在 $\delta > 0$ 使得 $\|x - z\| < \delta$，意味着 $|g(x) - g(z)| < \epsilon$ 成立。因此，

$$\begin{aligned}
&\frac{1}{k}\sum_{i=1}^{k} \mathbf{E}\left\{\left|f(X) - f\left(X_{(i)}(X)\right)\right|\right\} \\
&\quad \leq \mathbf{E}\{|f(X) - g(X)|\} + \frac{1}{k}\sum_{i=1}^{k} \mathbf{E}\left\{\left|g(X) - g\left(X_{(i)}(X)\right)\right|\right\} \\
&\qquad + \frac{1}{k}\sum_{i=1}^{k} \mathbf{E}\left\{\left|g\left(X_{(i)}(X)\right) - f\left(X_{(i)}(X)\right)\right|\right\} \\
&\quad \leq (1 + \gamma_d)\, \mathbf{E}\{|f(X) - g(X)|\} + \epsilon + \|g\|_\infty \mathbf{P}\left\{\left\|X - X_{(k)}(X)\right\| > \delta\right\} \\
&\qquad (\text{通过引理5.3, 其中}\delta \text{ 仅依赖于}\epsilon) \\
&\quad \leq (2 + \gamma_d)\,\epsilon + o(1) \quad (\text{通过引理 5.1}).
\end{aligned}$$

[证毕]

5.3 证明 Stone 引理

在本节中我们证明引理 5.3。对 $\theta \in (0, \pi/2)$，圆锥 $C(x, \theta)$ 是所有 $y \in \mathcal{R}^d$ 的集合,其中 $angle(x, y) \leq \theta$。等价地,对向量概念,$x^T y / \|x\| \|y\| \geq \cos\theta$。集合 $z + C(x, \theta)$ 是 $C(x, \theta)$ 平移 z 后的结果。

若 $y, z \in C(x, \pi/6)$，和 $\|y\| < \|z\|$，则 $\|y - z\| < \|z\|$，如我们现在所见。确实

$$\begin{aligned}
\|y - z\|^2 &= \|y\|^2 + \|z\|^2 - 2\|y\|\|z\| \frac{y^T z}{\|y\|\|z\|} \\
&\leq \|y\|^2 + \|z\|^2 - 2\|y\|\|z\| \cos(\pi/3) \\
&= \|z\|^2 \left(1 + \frac{\|y\|^2}{\|z\|^2} - \frac{\|y\|}{\|z\|}\right) \\
&< \|z\|^2 \quad \text{（见图 5.3）}
\end{aligned}$$

在下文中需要如下覆盖引理:

引理 5.5 *固定 $\theta \in (0, \pi/2)$。则存在一个集合 $\{x_1, \ldots, x_{\gamma_d}\} \subset \mathcal{R}^d$，使得*

$$\mathcal{R}^d = \bigcup_{i=1}^{\gamma_d} C(x_i, \theta).$$

另外，总能取到

$$\gamma_d \leq \left(1 + \frac{1}{\sin(\theta/2)}\right)^d - 1.$$

对 $\theta = \pi/6$，我们有

$$\gamma_d \leq \left(1 + \frac{2}{\sqrt{2 - \sqrt{3}}}\right)^d - 1.$$

证. 不失一般性，我们假设 $\|x_i\| = 1$, $\forall i$。每个 x_i 为半径 $r = 2\sin(\theta/2)$ 的球 S_i 的圆心。球 S_i 具有性质

$$\{x : \|x\| = 1\} \cap S_i = \{x : \|x\| = 1\} \cap C(x_i, \theta).$$

我们仅关注使得 $\|x_i - x_j\| \geq r$, $\forall j \neq i$ 成立的 x_i。在这一情况下，$\bigcup C(x_i, \theta)$ 覆盖 $\mathcal{R}^d$ 当且仅当 $\bigcup S_i$ 覆盖 $\{x : \|x\| = 1\}$。则圆心在 x_i 处，半径为 $r/2$ 的球 S_i' 互不相交且 $\bigcup S_i' \subseteq S_{0,1+r/2} - S_{0,r/2}$（见图5.1）。

因此，若 $v_d = volume(S_{0,1})$，则

$$\gamma_d v_d \left(\frac{r}{2}\right)^d \leq v_d \left(1 + \frac{r}{2}\right)^d - v_d \left(\frac{r}{2}\right)^d$$

或

$$\gamma_d \leq \left(1 + \frac{2}{r}\right)^d - 1 = \left(1 + \frac{1}{\sin(\theta/2)}\right)^d - 1.$$

最后的不等式可由

$$\sin\frac{\pi}{12} = \sqrt{\frac{1 - \cos(\pi/6)}{2}} = \sqrt{\frac{2 - \sqrt{3}}{4}}$$

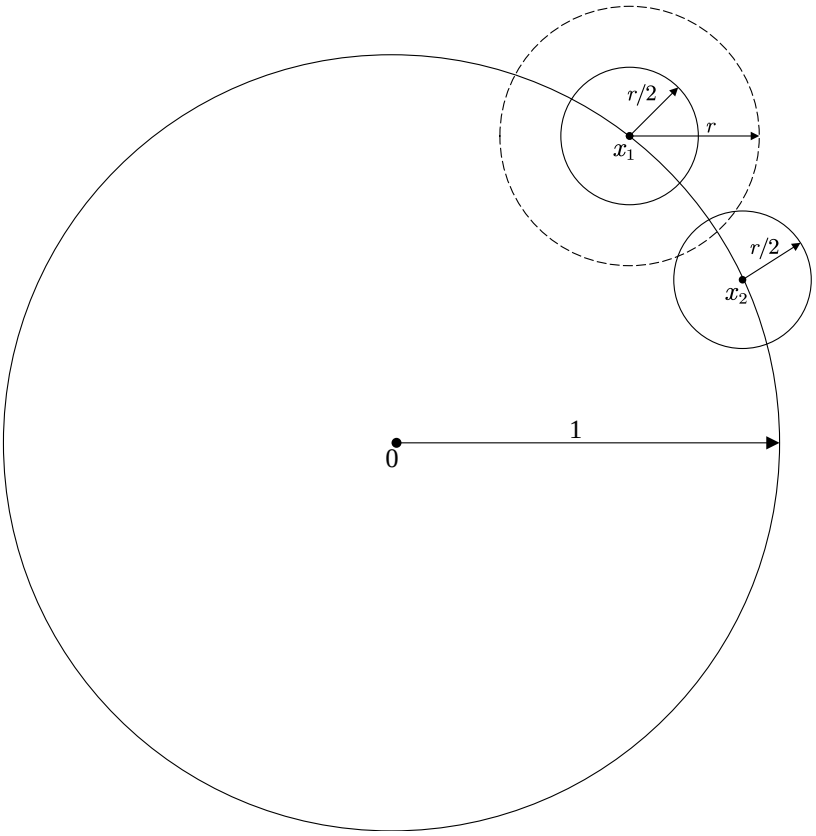

Figure 5.1: γ_d 的界

得到。 [证毕]

随着初步结果的出现，我们使用 γ_d 个圆锥 $X + C(x_j, \pi/6), 1 \le j \le \gamma_d$ 覆盖 $\mathcal{R}^d$，并在每个圆锥中标记最接近 X 的 X_i，如果这一 X_i 存在。若 X_i 属于 $X + C(x_j, \pi/6), 1 \le j \le \gamma_d$ 但没有被标记，则 X 不能是 X_i 在 $\{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}$ 的[13] 最近邻。类似地，我们将标记在每个圆锥中所有 X 的 k-最近邻（如果在一个圆锥中的点少于 k，则标记所有点）。同样，若 $X_i \in X + C(x_j, \pi/6)$ 未被标记，则 X 不是 X_i 在 $\{X_1, \ldots, X_{i-1}, X, X_{i+1}, \ldots, X_n\}$ 中的 k 最近邻。（若距离以大于零的概率出现相等，则这个点集的顺序相当重要，可以通过比较下标解决该问题。）因此，若 f 为非负函数，

13: 许有问题

$$
\begin{aligned}
&\sum_{i=1}^{k} \mathbf{E}\left\{f\left(X_{(i)}(X)\right)\right\} \\
&\quad = \mathbf{E}\left\{\sum_{i=1}^{n} I_{\left\{X_i \text{是} X \text{的} k-\text{最近邻在} \{X_1,\ldots,X_n\} \text{中}\right\}} f(X_i)\right\} \\
&\quad = \mathbf{E}\left\{f(X)\sum_{i=1}^{n} I_{\left\{X \text{是} X_i \text{的} k-\text{最近邻在} \{X_1,\ldots,X_{i-1},X,X_{i+1},\ldots,X_n\} \text{中}\right\}}\right\} \\
&\quad \text{(通过交换} X \text{和} X_i\text{)} \\
&\quad \le \mathbf{E}\left\{f(X)\sum_{i=1}^{n} I_{\{X_i \text{ is marked }\}}\right\} \\
&\quad \le k\gamma_d \mathbf{E}\{f(X)\},
\end{aligned}
$$

14: 许有问题

因为在每个圆锥中我们最多能标记 k 个点，且圆锥的数量最多为 γ_d 个

——见引理5.5。得证 Stone 引理。

5.4 渐进误差概率

我们继续讨论k-局部规则（特别是k-最近邻规则）。令 $D'_n = ((X_1,Y_1,U_1),\ldots,(X_n,Y_n,U_n))$ 为独立同分布数据，其由前述的均匀分布随机变量 $U_1,\ldots,U_n$ 扩增（augment）得到。对一基于 D_n 的决策 g_n，我们得到误差概率

$$\begin{aligned} L_n &= \mathbf{P}\left\{g_n(X) \neq Y \mid D'_n\right\} \\ &= \mathbf{P}\left\{\text{sign}\left(\psi\left(X, Y_{(1)}(X),\ldots,Y_{(k)}(X)\right)\right) \neq \text{sign}(2Y-1) \mid D'_n\right\}, \end{aligned}$$

其中函数 ψ 的符号决定 g_n；见式(5.1)。定义随机变量 $Y'_{(1)}(X),\ldots,Y'_{(k)}(X)$，并令

$$L'_n = \mathbf{P}\left\{\text{sign}\left(\psi\left(X, Y'_{(1)}(X),\ldots,Y'_{(k)}(X)\right)\right) \neq \text{sign}(2Y-1) \mid D'_n\right\}.$$

通过引理5.2和引理5.4,

$$\begin{aligned} &\mathbf{E}\left\{\left|L_n - L'_n\right|\right\} \\ &\leq \mathbf{P}\left\{\psi\left(X, Y_{(1)}(X),\ldots,Y_{(k)}(X)\right) \neq \psi\left(X, Y'_{(1)}(X),\ldots,Y'_{(k)}(X)\right)\right\} \\ &\leq \textstyle\sum_{i=1}^{k} \mathbf{E}\left\{\left|\eta(X) - \eta\left(X_{(i)}(X)\right)\right|\right\} \\ &= o(1). \end{aligned}$$

因为 $\lim_{n\to\infty}\left(\mathbf{E}L'_n - \mathbf{E}L_n\right) = 0$，我们仅仅需要研究规则

$$g'_n(x) = \begin{cases} 1 & \text{若}\psi\left(x, Z_1,\ldots,Z_k\right) > 0 \\ 0 & \text{否则} \end{cases}$$

（$Z_1,\ldots,Z_k$ 为独立同分布的 $Bernoulli(\eta(x))$ ）除非我们也关心 L_n 到 $\mathbf{E}L_n$ 的接近程度。

译注：伯努利分布

我们现在说明在 1-最近邻规则上一个重要的降低计算时间的方法。显然，$\psi\left(x, Z_1\right) = 2Z_1 - 1$，因此

$$\mathbf{E}\left\{L'_n\right\} = \mathbf{P}\left\{Z_1 \neq Y\right\} = \mathbf{E}\{2\eta(X)(1-\eta(X))\}.$$

毋须额外计算，我们有

定理 5.1 *对最近邻规则，对任 (X,Y) 分布，*

$$\lim_{n\to\infty} \mathbf{E}\left\{L_n\right\} = \mathbf{E}\{2\eta(X)(1-\eta(X))\} = L_{\text{NN}}.$$

在多个连续性条件下（X 具密度 f，且 f 与 η 都几乎处处连续）；这一结果来自 Cover 和 Hart (1967)。现在的表述一般沿用 Stone (1997)。另见 Devroye（1981c)。在第 3 章，我们已证明

$$L^* \leq L_{\text{NN}} \leq 2L^*\left(1-L^*\right) \leq 2L^*.$$

因此，上一个结果说明最近邻规则在最坏情况最多是贝叶斯规则的两倍——尤其是对于小的 L^*，该性质是相当有用的。

我们正式定义该量，当 k 为奇数，

$$L_{k\mathrm{NN}} = \mathbf{E}\left\{\sum_{j=0}^{k}\binom{k}{j}\eta^j(X)(1-\eta(X))^{k-j}\left(\eta(X)I_{\{j<k/2\}} + (1-\eta(X))I_{\{j>k/2\}}\right)\right\}$$

我们有如下结果：

定理 5.2 *令 k 为一固定奇数。则对 kNN，*

$$\lim_{n\to\infty}\mathbf{E}\{L_n\} = L_{k\mathrm{NN}}.$$

证. 我们注意到，可以证明 $\lim_{n\to\infty}\mathbf{E}\left\{L'_n\right\} = L_{k\mathrm{NN}}$（由前面介绍的概念）。但对任 n，

$$\begin{aligned}
&\mathbf{E}\left\{L'_n\right\}\\
&\quad= \mathbf{P}\left\{Z_1+\cdots+Z_k > \frac{k}{2}, Y=0\right\} + \mathbf{P}\left\{Z_1+\cdots+Z_k < \frac{k}{2}, Y=1\right\}\\
&\quad= \mathbf{P}\left\{Z_1+\cdots+Z_k > \frac{k}{2}, Z_0=0\right\} + \mathbf{P}\left\{Z_1+\cdots+Z_k < \frac{k}{2}, Z_0=1\right\}\\
&\quad(\text{其中}Z_0,\ldots,Z_k\text{为独立同分布 Bernoulli}(\eta(X))\text{随机变量}),
\end{aligned}$$

可直接得到所求结果。 [证毕]

如下几个 $L_{k\mathrm{NN}}$ 的表示将对后述分析有用。例如，我们有

$$\begin{aligned}
&L_{k\mathrm{NN}}\\
&\quad= \mathbf{E}\left\{\eta(X)\mathbf{P}\left\{\mathrm{Binomial}(k,\eta(X)) < \frac{k}{2}\middle| X\right\}\right\}\\
&\quad+ \mathbf{E}\left\{(1-\eta(X))\mathbf{P}\left\{\mathrm{Binomial}(k,\eta(X)) > \frac{k}{2}\middle| X\right\}\right\}\\
&\quad= \mathbf{E}\{\min(\eta(X),1-\eta(X))\}\\
&\qquad+ \mathbf{E}\left\{(1-2\min(\eta(X),1-\eta(X)))\mathbf{P}\left\{\mathrm{Binomial}(k,\eta(X)) > \frac{k}{2}\middle| X\right\}\right\}.
\end{aligned}$$

应强调定理5.2的极限结果与分布无关。极限 $L_{k\mathrm{NN}}$ 仅依赖于 η（或 $\min(\eta(X)),1-\eta(X))$。η 的连续性或非平滑性是无关紧要的——仅关注 $\mathbf{E}\{L_n\}$ 接近极限 $L_{k\mathrm{NN}}$ 的速度。

5.5 加权最近邻规则的渐进误差概率

Royall（1966）提出一个具权重 $w_1,\ldots,w_k$ 的加权最近邻规则，其根据

$$g_n(x) = \begin{cases} 1 & \text{若 } \sum_{i:Y_{(i)}(x)=1}^{k} w_i > \sum_{i:Y_{(i)}(x)=0}^{k} w_i \\ 0 & \text{否则.} \end{cases}$$

作出决策。在出现正反票数相同的情况下，该规则是非对称的。若出现该情况，我们将其修改为 $g_n(x) \stackrel{\mathrm{def}}{=} -1$。值“−1”应视为“无决策

(indecision)”。由上述论证，渐进误差概率为 $w_1, \ldots, w_k$ 的函数，即 $L(w_1, \ldots, w_k) = \mathbf{E}\{\alpha(\eta(X))\}$，其中

$$\begin{aligned}\alpha(p) =& \mathbf{P}\left\{\sum_{i=1}^{k} w_i Y_i' > \sum_{i=1}^{k} w_i \left(1 - Y_i'\right)\right\}(1-p) \\ &+ \mathbf{P}\left\{\sum_{i=1}^{k} w_i Y_i' \leq \sum_{i=1}^{k} w_i \left(1 - Y_i'\right)\right\} p,\end{aligned}$$

现在 $Y_1', \ldots, Y_k'$ 为独立同分布的参数 p 的伯努利分布。等价地，$Z_i = 2Y_i' - 1 \in \{-1, 1\}$，

$$\alpha(p) = (1-p)\mathbf{P}\left\{\sum_{i=1}^{k} w_i Z_i > 0\right\} + p\mathbf{P}\left\{\sum_{i=1}^{k} w_i Z_i \leq 0\right\}.$$

现在假设 $\mathbf{P}\left\{\sum_{i=1}^{k} w_i Z_i = 0\right\} = 0$。则，若 $p < 1/2$，

$$\alpha(p) = p + (1-2p)\mathbf{P}\left\{\sum_{i=1}^{k} w_i Z_i > 0\right\},$$

且当 $p > 1/2$，其反对称表达式成立。请注意如下内容。若我们令 N_l 为满足 $\sum I_{\{z_i=1\}} = l$ 和 $\sum w_i z_i > 0$ 的向量 $z = (z_1, \ldots, z_k) \in \{-1, 1\}^k$ 的个数，则 $N_l + N_{k-l} = \binom{k}{l}$。因此，

$$\begin{aligned}&\mathbf{P}\left\{\sum_{i=1}^{k} w_i Z_i > 0\right\} \\ &= \sum_{l=0}^{k} N_l p^l (1-p)^{k-l} \\ &= \sum_{l<k/2} \binom{k}{l} p^{k-l}(1-p)^l + \sum_{l<k/2} N_l \left(p^l(1-p)^{k-l} - p^{k-l}(1-p)^l\right) \\ &\quad + \frac{1}{2}\binom{k}{k/2} p^{k/2}(1-p)^{k/2} I_{\{k\text{为偶数}\}} \\ &= I + II + III.\end{aligned}$$

注意 $I + III$ 代表

$$\mathbf{P}\{\text{Binomial}(k, 1-p) \leq k/2\} = \mathbf{P}\{\text{Binomial}(k, p) \geq k/2\},$$

其不依赖于权重向量。最后，由于 $p \leq 1/2$，

$$\begin{aligned}II &= \sum_{l<k/2} N_l \left(p^l(1-p)^{k-l} - p^{k-l}(1-p)^l\right) \\ &= \sum_{l<k/2} N_l p^l (1-p)^l \left((1-p)^{k-2l} - p^{k-2l}\right) \\ &\geq 0.\end{aligned}$$

该项为 0，当且仅当 $N_l = 0, \forall l < k/2$。换言之，[15] 该项等式成立当且仅当不存在 w_i 的少数值（numerical minority），其可表示为多数值（majority）的和。（例如若权重集为 $(0.7, 0.2, 0.1)$，其中 0.7 为少数值，它大于所有其他权重）。但当 k 为奇数时，这种情况等价于普通的 k-最

15: 许有问题

近邻规则。当 k 为偶数时，我们为一个 w_i 添加小的权重

$$\left(\frac{1+\epsilon}{k}, \frac{1-\epsilon/(k-1)}{k}, \ldots, \frac{1-\epsilon/(k-1)}{k}\right),$$

对小 $\epsilon > 0$，则不存在 w_i 的少数值能大于其他权重，因此我们得到一个最优规则（$II = 0$）。故有：

Bailey 和 Jain（1978）

定理 5.3 *令 $L(w_1, \ldots, w_k)$ 为具权重 $w_1, \ldots, w_k$ 的加权 $k-$NN 规则的渐进误差概率。当 k 为奇数时，则 $k-$NN 规则可定义为 $(1/k, 1/k, \ldots, 1/k)$。当 k 为偶数时，则可定义为 $\forall \epsilon \in (0,1)$，*

$$(1/k, 1/k, \ldots, 1/k) + \epsilon(1, -1/(k-1), -1/(k-1), \ldots, -1/(k-1)).$$

记后者的渐进误差概率为 $L_{k\mathrm{NN}}$，我们有

$$L(w_1, \ldots, w_k) \geq L_{k\mathrm{NN}}.$$

若 $\mathbf{P}\{\eta(X) = 1/2\} < 1$，则等式成立当且仅当 w_i 的每个少数值比所有权重和的一半小。

该结果表明在渐进意义上，普通 k-最近邻规则是更优的。这并不意味着对一特定样本大小，我们应避免不均匀的权重分布。事实上，若允许 k 随 n 的变化而变化，则不均匀权重反而更为有利（Royall（1966））。

考虑到所有满足 $w_i \geq 0, \sum_{i=1}^{k} w_i = 1$ 的权重向量 $(w_1, \ldots, w_k)$ 构成的空间 $\mathscr{W}$。$L(w_1, \ldots, w_k)$ 是否是全序的？为回答该问题，我们需再次回到 $\alpha(p)$。权重向量仅影响 II 项。例如，考虑到权重向量

$$(0.3, 0.22, 0.13, 0.12, 0.071, 0.071, 0.071, 0.017)$$
$$\text{和} \quad (0.26, 0.26, 0.13, 0.12, 0.071, 0.071, 0.071, 0.017).$$

少数值由一，二或三个分量组成。对这两个权重向量，$N_1 = 0, N_2 = 1$。但是，前者 $N_3 = 6 + 4$，而后者 $N_3 = 6 + 2$。故后者的项 II 对所有 $p < 1/2$ 一致更小，因此对*所有*分布，第二个权重向量更优。当 N_l 不严格嵌套（nested）时，这种对比 是无法进行的。，见问题5.8中的示例。因此，$\mathscr{W}$ 仅是偏序的。

全体对比

不知不觉中我们已证明了如下定理：

定理 5.4 *对任意分布，*

$$L^* \leq \cdots \leq L_{(2k+1)\mathrm{NN}} \leq L_{(2k-1)\mathrm{NN}} \leq \cdots \leq L_{3\mathrm{NN}} \leq L_{\mathrm{NN}} \leq 2L^*.$$

证. 我们再次关注 $\alpha(p)$。对于 $(2k+1)-$NN 考虑权重向量 $w_1 = \ldots = w_{2k+1} = 1$（忽略标准化）。因为 $N_0 = N_1 = \cdots = N_k = 0$，故项 II 为 0。但是，$(2k-1)-$NN 规则具权重 $w_1 = \cdots = w_{2k-1} = 1, w_{2k} = w_{2k+1} = 0$，具有一个非零项，因为 $N_0 = \ldots = N_{k-1} = 0$，而 $N_k = \binom{2k-1}{k} > 0$。因此 $L_{(2k+1)\mathrm{NN}} \leq L_{(2k-1)\mathrm{NN}}$。 [证毕]

备注 5.3 当 $\mathbf{P}\{\eta(X) \notin \{0, 1, 1/2\}\} > 0$，我们有严格不等式 $L_{(2k+1)\mathrm{NN}} \leq L_{(2k-1)\mathrm{NN}}$。当 $L^* = 0$，我们由 $L_{\mathrm{NN}} = L_{3\mathrm{NN}} = L_{5\mathrm{NN}} = \cdots = 0$。

5.6 k-最近邻规则：偶数 k

至今我们始终假设 k 为奇数，因此能避免正反票数相同的情况。对 $2k$-最近邻规则（偶数情况），我们用如下定义

$$g_n(x) = \begin{cases} 1 & \text{if } \sum_{i=1}^{2k} Y_{(i)}(x) > k \\ 0 & \text{if } \sum_{i=1}^{2k} Y_{(i)}(x) < k \\ Y_{(1)}(x) & \text{if } \sum_{i=1}^{2k} Y_{(i)}(x) = k \end{cases}$$

解决票数相同问题。正式地说,这等价于一个具权重向量 $(3,2,2,2,\ldots,2,2)$ 的加权 $2k$-最近邻规则。根据定理5.3容易验证该权重是渐进最优权重向量。偶数情况并未降低误差概率。具体来说，我们有：

定理 5.5 *对任意分布，任一整数 k，*

$$L_{(2k-1)\text{NN}} = L_{(2k)\text{NN}}$$

Devijver（1978）

证. 回忆 $L_{k\text{NN}}$ 可表示为 $L_{k\text{NN}} = \mathbf{E}\{\alpha(\eta(X))\}$，其中

$$\alpha(\eta(x)) = \lim_{n\to\infty} \mathbf{P}\left\{g_n^{(k)}(X) \neq Y \mid X = x\right\}$$

为 $k-NN$ 规则 $g_n^{(k)}$ 的逐点（pointwise）渐进误差概率。为方便讨论，将 $Z_1,...,Z_{2k}$ 看作独立同分布的 $\{-1,1\}$ 值随机变量，且 $\mathbf{P}\{Z_i = 1\} = p = \eta(x)$，另外把 $\sum_{i=1}^{2k} Z_i$ 的符号作为最终的决策值。由加权最近邻规则的一般公式（general formula），$(2k)-\text{NN}$ 规则的逐点渐进误差概率为

$$\begin{aligned}
&\lim_{n\to\infty} \mathbf{P}\left\{g_n^{(2k)}(X) \neq Y \mid X = x\right\} \\
&\quad = p\mathbf{P}\left\{\sum_{i=1}^{2k} Z_i < 0\right\} + p\mathbf{P}\left\{\sum_{i=1}^{2k} Z_i = 0, Z_1 < 0\right\} \\
&\qquad + (1-p)\mathbf{P}\left\{\sum_{i=1}^{2k} Z_i > 0\right\} + (1-p)\mathbf{P}\left\{\sum_{i=1}^{2k} Z_i = 0, Z_1 > 0\right\} \\
&\quad = p\mathbf{P}\left\{\sum_{i=2}^{2k} Z_i < 0\right\} + (1-p)\mathbf{P}\left\{\sum_{i=2}^{2k} Z_i > 0\right\} \\
&\quad = \lim_{n\to\infty} \mathbf{P}\left\{g_n^{(2k-1)}(X) \neq Y \mid X = x\right\}
\end{aligned}$$

因此，$L_{(2k)\text{NN}} = L_{(2k-1)\text{NN}}$。 [证毕]

5.7 误差概率不等式

我们回到 k 为奇数的情况。回忆到

$$L_{k\text{NN}} = \mathbf{E}\left\{\alpha_k(\eta(X))\right\},$$

其中

$$\alpha_k(p) = \min(p, 1-p) + |2p-1|\mathbf{P}\left\{\text{Binomial}(k, \min(p,1-p)) > \frac{k}{2}\right\}.$$

因为 $L^* = \mathbf{E}\{\min(\eta(X), 1-\eta(X))\}$，我们利用该表示得到 $L_{k\text{NN}} - L^*$ 的几个不等式。我们先介绍其中一个容易证明但可能不那么强的不等式。

定理 5.6 *对任意奇数 k 与所有分布，*

$$L_{k\text{NN}} \le L^* + \frac{1}{\sqrt{ke}}.$$

证. 由上述表示，

$$\begin{aligned}
L_{k\text{NN}} - L^* &\le \sup_{0\le p\le 1/2} (1-2p)\mathbf{P}\left\{B > \frac{k}{2}\right\} \\
&\quad (B\text{是Binomial}(k,p)) \\
&= \sup_{0\le p\le 1/2} (1-2p)\mathbf{P}\left\{\frac{B-kp}{k} > \frac{1}{2}-p\right\} \\
&\le \sup_{0\le p\le 1/2} (1-2p)e^{-2k(1/2-p)^2} \\
&\quad (\text{通过 Okamoto-Hoeffding 不等式-定理??}) \\
&= \sup_{0\le u\le 1} ue^{-ku^2/2} \\
&= \frac{1}{\sqrt{ke}}.
\end{aligned}$$

[证毕]

Györfi 和 Györfi（1978）

定理 5.7 *对任意奇数 k 与所有分布，*

$$L_{k\text{NN}} \le L^* + \sqrt{\frac{2L_{\text{NN}}}{k}}.$$

证. 我们注意到对 $p \le 1/2$，B 具分布 $binomial(k,p)$，

$$\begin{aligned}
\mathbf{P}\left\{B > \frac{k}{2}\right\} &= \mathbf{P}\left\{B - kp > k\left(\frac{1}{2}-p\right)\right\} \\
&\le \frac{\mathbf{E}\{|B-kp|\}}{k(1/2-p)} \quad \text{(Markov 不等式)} \\
&\le \frac{\sqrt{\text{Var}\{B\}}}{k(1/2-p)} \quad \text{(Cauchy-Schwarz 不等式)} \\
&= \frac{2\sqrt{p(1-p)}}{\sqrt{k}(1-2p)}.
\end{aligned}$$

因此，

$$\begin{aligned}
L_{k\text{NN}} - L^* &\le \mathbf{E}\left\{\frac{2}{\sqrt{k}}\sqrt{\eta(X)(1-\eta(X))}\right\} \\
&\le \frac{2}{\sqrt{k}}\sqrt{\mathbf{E}\{\eta(X)(1-\eta(X))\}} \quad \text{(Jensen 不等式)} \\
&= \frac{2}{\sqrt{k}}\sqrt{\frac{L_{\text{NN}}}{2}} \\
&= \sqrt{\frac{2L_{\text{NN}}}{k}}.
\end{aligned}$$

得证。 [证毕]

备注 5.4 对大的 k，B 的分布近似为 $normal(k, p(1-p))$，因此 $\mathbf{E}\{|B-kp|\} \approx \sqrt{kp(1-p)}\sqrt{2/\pi}$，因为一个正态随机变量的第一绝对矩为 $\sqrt{2/\pi}$（见问题5.11）。由此推出一个 $\sqrt{L_{\mathrm{NN}}/(\pi k)}$ 的近似界。该界与 $\sqrt{L_{\mathrm{NN}}}$ 成比例。如果不使用 Markov 不等式求其界，而是直接用如下方法近似 $\mathbf{P}\{B-kp > k(1/2-p)\}$，则该界可被推广至 L^*。

定理 5.8 *对任意分布与任奇数 $k \geq 3$,* Devroye（1981a）

$$L_{k\mathrm{NN}} \leq L^*\left(1+\frac{\gamma}{\sqrt{k}}\left(1+O\left(k^{-1/6}\right)\right)\right),$$

其中 $\gamma = \sup_{r>0} 2r\mathbf{P}\{N > r\} = 0.33994241\ldots$, N 是 $normal(0,1)$ 分布，且 $O(\cdot)$ 指 $k \to \infty$[16] *。（显式变量在证明中给出。）*

16: 许有问题

证明中的变量 γ 无法进一步改进。一个稍微弱的界由 Devijver（1979）获得：

$$\begin{aligned} L_{k\mathrm{NN}} &\leq L^* + \frac{1}{2^{2k'}}\binom{2k'}{k'}L_{\mathrm{NN}} \quad (\text{其中} k' = \lceil k/2 \rceil) \\ &= L^* + L_{\mathrm{NN}}\sqrt{\frac{2}{\pi k'}}(1+o(1)) \quad (\text{当} k \to \infty, \text{见引理 A.3}) \end{aligned}$$

另见 Devijver 和 Kittler（1982，102 页）。

引理 5.6 *对 $p \leq 1/2$ 与奇数 $k > 3$,* Devroye（1981B）

$$\begin{aligned} \mathbf{P}\left\{\mathrm{Binomial}(k,p) > \frac{k}{2}\right\} &= \frac{k!}{\left(\frac{k-1}{2}\right)!\left(\frac{k-1}{2}\right)!}\int_0^p (x(1-x))^{(k-1)/2}dx \\ &\leq A\int_{(1-2p)\sqrt{k-1}}^{\sqrt{k-1}} e^{-z^2/2}dz, \end{aligned}$$

其中 $A \leq \frac{1}{\sqrt{2\pi}}\left(1+\frac{2}{k}+\frac{3}{4k^2}\right)$。

证. 考虑 k 个在 $[0,1]$ 上的独立同分布均匀随机变量。所有值位于区间 $[0,p]$ 的变量个数为 $binomial(k,p)$ 分布 。个数超过 $k/2$ 当且仅当均匀云（uniform cloud）的第 $(k+1)/2$ 个次序统计量的值最大为 p。后者为 $beta((k+1)/2,(k+1)/2)$ 分布，解释了第一个等式（问题5.32）。注意我们将离散和写成积分——在某些情况下，该技巧会相当有用。为证明接下来的不等式成立，替换 x 为 $\frac{1}{2}\left(1-\frac{z}{\sqrt{k-1}}\right)$，并使用不等式 $1-u \leq e^{-u}$ 获得一个界： 二项分布

$$A = \frac{1}{2^k\sqrt{k-1}} \times \frac{k!}{\left(\frac{k-1}{2}\right)!\left(\frac{k-1}{2}\right)!}.$$

最后，

$$\begin{aligned}
A &= \mathbf{P}\left\{B = \frac{k+1}{2}\right\}\frac{k+1}{2\sqrt{k-3}} \quad (B\text{为}binomial(k,1/2)\text{分布}) \\
&\le \sqrt{\frac{k}{2\pi\frac{k+1}{2}\frac{k-1}{2}}}\frac{k+1}{2\sqrt{k-1}} \quad (\text{问题5.17}) \\
&= \frac{1}{\sqrt{2\pi}}\frac{\sqrt{k(k+1)}}{k-1} \\
&\le \frac{1}{\sqrt{2\pi}}\left(1+\frac{2}{k}+\frac{3}{4k^2}\right) \quad (\text{问题5.18}).
\end{aligned}$$

[证毕]

证 (定理5.8). 由前述有，

$$\begin{aligned}
L_{k\mathrm{NN}} - L^* &= \mathbf{E}\left\{\alpha_k(\eta(X)) - \min(\eta(X), 1-\eta(X))\right\} \\
&= \mathbf{E}\left\{\left(\frac{\alpha_k(\eta(X))}{\min(\eta(X),1-\eta(X))} - 1\right)\min(\eta(X),1-\eta(X))\right\} \\
&\le \left(\sup_{0<p<1/2}\frac{1-2p}{p}\mathbf{P}\left\{B > \frac{k}{2}\right\}\right)L^* \quad (B\text{为}binomial(k,p)\text{分布}).
\end{aligned}$$

我们仅给出括号中因子的界。显然，由引理5.6，

$$L_{k\mathrm{NN}} - L^* \le L^*\left(\sup_{0<p<1/2}\frac{1-2p}{p}A\int_{(1-2p)\sqrt{k-1}}^{\sqrt{k-1}} e^{-z^2/2}dz\right).$$

取 $a < 1$ 作为 $\left(3/\left(ea^2\right)\right)^{3/2}\frac{1}{\sqrt{2\pi(k-1)}} = \gamma$ 的解，若 $k-1 > \frac{1}{2\pi}\frac{1}{\gamma^2}\left(\frac{3}{e}\right)^6 = 2.4886858\ldots$，该解存在。令 $v = (1-2p)\sqrt{k-1}$，我们有

$$\begin{aligned}
&\sup_{0<p<1/2}\frac{1-2p}{p}\int_{(1-2p)\sqrt{k-1}}^{\sqrt{k-1}}\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz \\
&\quad\le \max\left(\sup_{0<v\le a\sqrt{k-1}}\frac{2v/\sqrt{k-1}}{1-v/\sqrt{k-1}}\int_v^\infty\frac{e^{-z^2/2}}{\sqrt{2\pi}}dz\right. \\
&\quad\quad\left.\sup_{a\sqrt{k-1}\le v<\sqrt{k-1}}\frac{2p\sqrt{k-1}}{1-v/\sqrt{k-1}}\frac{e^{-v^2/2}}{\sqrt{2\pi}}\right) \\
&\quad\le \max\left(\frac{\gamma}{(1-a)\sqrt{k-1}}, \frac{\sqrt{k-1}}{\sqrt{2\pi}}e^{-a^2(k-1)/2}\right) \\
&\quad\le \max\left(\frac{\gamma}{(1-a)\sqrt{k-1}}, \left(\frac{3}{ea^2}\right)^{3/2}\frac{1}{(k-1)\sqrt{2\pi}}\right) \\
&\quad\quad(\text{使用 } u^{3/2}e^{-cu} \le (3/(2ce))^{3/2}\forall, u>0\Big) \\
&\quad= \frac{\gamma}{(1-a)\sqrt{k-1}}.
\end{aligned}$$

使用这些界并注意 $a = O\left(k^{-1/6}\right)$，可得证。 [证毕]

5.8 当 L^* 很小时的表现

在本节，我们更进一步研究当 L^* 很小时 $L_{k\mathrm{NN}}$ 的表现。回忆前面有 $L_{k\mathrm{NN}} = \mathbf{E}\{\alpha_k(\eta(X))\}$ 且对任意奇数 k 有

$$\alpha_k(p) = \min(p, 1-p) + |1 - 2\min(p, 1-p)| \times \mathbf{P}\left\{\mathrm{Binomial}(k, \min(p, 1-p)) > \frac{k}{2}\right\},$$

容易验证对[17] 某些函数 ξ_k，$L_{k\mathrm{NN}} = \mathbf{E}\{\xi_k(\min(\eta(X), 1-\eta(X)))\}$ 成立。因为

17: 许有问题。存在

$$\min(p, 1-p) = \frac{1 - \sqrt{1-4p(1-p)}}{2},$$

对某些其他函数 ψ_k，我们也有 $L_{k\mathrm{NN}} = \mathbf{E}\{\psi_k(\eta(X)(1-\eta(X)))\}$。$L_{k\mathrm{NN}}$ 的解形式（workedout forms）包括

$$\begin{aligned} L_{k\mathrm{NN}} =& \mathbf{E}\left\{\sum_{j<k/2}\binom{k}{j}\eta(X)^{j+1}(1-\eta(X))^{k-j} \right. \\ &\left. + \sum_{j>k/2}\binom{k}{j}\eta(X)^{j}(1-\eta(X))^{k-j+1}\right\} \\ =& \sum_{j<k/2}\binom{k}{j}\mathbf{E}\left\{(\eta(X)(1-\eta(X)))^{j+1}\left((1-\eta(X))^{k-2j-1} + \eta(X)^{k-2j-1}\right)\right\}. \end{aligned}$$

因为 $p^a + (1-p)^a$ 是 $p(1-p)$ 对整数 a 的函数，所以能进一步简化。例如

$$\begin{aligned} L_{\mathrm{NN}} =& \mathbf{E}\{2\eta(X)(1-\eta(X))\}, \\ L_{3\mathrm{NN}} =& \mathbf{E}\{\eta(X)(1-\eta(X))\} + 4\mathbf{E}\left\{(\eta(X)(1-\eta(X)))^2\right\}, \\ L_{5\mathrm{NN}} =& \mathbf{E}\{\eta(X)(1-\eta(X))\} + \mathbf{E}\left\{(\eta(X)(1-\eta(X)))^2\right\} \\ &+ 12\mathbf{E}\left\{(\eta(X)(1-\eta(X)))^3\right\}. \end{aligned}$$

α_k 接近于 0 时的表现极具信息量。当 $p \downarrow 0$，我们有

$$\begin{aligned} &\alpha_1(p) = 2p(1-p) \sim 2p, \\ &\alpha_3(p) = p(1-p)(1+4p) \sim p + 3p^2, \\ &\alpha_5(p) \sim p + 10p^3, \end{aligned}$$

同时对贝叶斯误差，$L^* = \mathbf{E}\{\min(\eta(X), 1-\eta(X))\} = \mathbf{E}\{\alpha_\infty(\eta(X))\}$，其中 $\alpha_\infty = \min(p, 1-p) \sim p$ 当 $p \downarrow 0$。假设 $\eta(x) = p$, $\forall x$。则当 $p \downarrow 0$,

$$L_{\mathrm{NN}} \sim 2L^* \quad \text{和} \quad L_{3\mathrm{NN}} \sim L^*.$$

另外，$L_{\mathrm{NN}} - L^* \sim L^*, L_{3\mathrm{NN}} - L^* \sim 3L^{*2}$。假设 $L^* = p = 0.01$。则 $L_1 \approx 0.02$，然而 $L_{3\mathrm{NN}} - L^* \approx 0.0003$。仅从实用角度，$3-\mathrm{NN}$ 规则几乎是完美的。因为这一理由，相当推荐 $3-\mathrm{NN}$ 规则。当 p 很小时，考虑 $5-\mathrm{NN}$ 规则几乎没有得到什么增益，因为 $L_{5\mathrm{NN}} - L^* \approx 0.00001$。

令 a_k 为使得 $\alpha_k(p) \le a_k \min(p, 1-p), \forall p$ 成立的最小数（图**??**的切线）。则

$$\begin{aligned} L_{k\mathrm{NN}} = \mathbf{E}\{\alpha_k(\eta(X))\} &\le a_k \mathbf{E}\{\min(\eta(X), (1-\eta(X))\} \\ &= a_k L^*. \end{aligned}$$

这正好是定理5.6和5.8的不等式的基础，结果表明 $a_k = 1 + O(1/\sqrt{k})$。

5.9 当 $L^* = 0$ 时的最近邻规则

从定理5.4我们有，若 $L^* = 0$ 则 $L_{k\text{NN}} = 0$, $\forall k$。实际上，则对任一固定 k，k-最近邻规则是一致的（consistent）。Cover 给出一个美妙的例子说明这一非凡的事实。$L^* = 0$ 意味着 $\eta(x) \in \{0,1\}$, $\forall x$，因此类是可分的。但这并不意味着给定 $Y = 0$ 时 X 的支撑与给定 $Y = 1$ 时的支撑不相同。举个例子，取一个 $[0,1]$ 上的随机有理数（例如，独立随机地按几何分布从集合 $\{1,2,3,\ldots\}$ 生成 I, J，并令 $X = \min(I,J)/\max(I,J)$）。每个 $[0,1]$ 上的有理数具有正概率。给定 $Y = 1$，X 如上；给定 $Y = 0$，X 为 $[0,1]$ 上的均匀分布。令 $\mathbf{P}\{Y = 1\} = \mathbf{P}\{Y = 0\} = 1/2$。$X$ 的支撑在这两种情况下是相等的。当

$$\eta(x) = \begin{cases} 1 & \text{若}x\text{为有理数} \\ 0 & \text{若}x\text{为无理数}, \end{cases}$$

我们有 $L^* = 0$，且最近邻规则是一致的。若从同一个分布的数据中取数 X，则我们可以仅仅根据 X 的最近邻的有理性与否判断 X 的有理性。尽管我们还未证明这一点，但若给定任一 $x \in [0,1]$，其同样成立：

18: 许有问题

$$\begin{aligned} &\lim_{n\to\infty} \mathbf{P}\{x\text{为有理数}^{18}\ Y_{(1)}(x) = 0\ (X_{(1)}(x)\text{不是有理数})\} \\ &\quad = \lim_{n\to\infty} \mathbf{P}\left\{x\text{不是有理数}Y_{(1)}(x) = 1\ (X_{(1)}(x)\text{是有理数})\right\} \\ &\quad = 0 \quad \text{(见问题5.38).} \end{aligned}$$

5.10 最近邻规则的容许性

第 11 章的一致性定理表明我们应在 $k-\text{NN}$ 规则中取 $k = k_n \to \infty$。$L_{k\text{NN}}$ 的递减性证明了这一点。然而，存在分布使得 $\forall n$，$1-\text{NN}$ 规则优于其他 $k-\text{NN}$ 规则，$k \geq 3$。该发现首先由 Cover 和 Hart（1967）提出，其依赖于一类例子。令 S_0 和 S_1 为两个半径均为 1，圆心分别为 a, b 的圆，其中 $\|a-b\| > 2$。给定 $Y = 1$ 时，X 服从在 S_1 上的均匀分布；给定 $Y = 0$ 时，X 服从在 S_0 上的均匀分布；另外 $\mathbf{P}\{Y = 1\} = \mathbf{P}\{Y = 0\} = 1/2$。我们注意到给定 n 个观测点，使用 $1-\text{NN}$ 规则，有

$$\mathbf{E}\{L_n\} = \mathbf{P}\{Y = 0, Y_1 = \cdots = Y_n = 1\} + \mathbf{P}\{Y = 1, Y_1 = \cdots = Y_n = 0\} = \frac{1}{2^n}.$$

对 $k-\text{NN}$ 规则，k 为奇数，有

$$\begin{aligned} \mathbf{E}\{L_n\} =& \mathbf{P}\left\{Y = 0, \sum_{i=1}^n I_{\{Y_i=0\}} \leq \lfloor k/2 \rfloor\right\} \\ &+ \mathbf{P}\left\{Y = 1, \sum_{i=1}^n I_{\{Y_i=1\}} \leq \lfloor k/2 \rfloor\right\} \\ =& \mathbf{P}\{\text{Binomial}(n,1/2) \leq \lfloor k/2 \rfloor\} \\ =& \frac{1}{2^n} \sum_{j=0}^{\lfloor k/2 \rfloor} \binom{n}{j} > \frac{1}{2^n} \quad \text{when } k \geq 3. \end{aligned}$$

因此，对任 n，当分布如上给定时，$k-\text{NN}$ 规则比 $1-\text{NN}$ 表现更差。我们参考了一些有趣的关于 $k-\text{NN}$ 规则容许性问题的练习。

5.11 (k,l)-最近邻规则

在 1970 年，Hellman（1970）提出了 (k,l)-最近邻规则，其与 k-最近邻规则完全相同，但只在存在至少 $l > k/2$ 个观测点位于同一个类的情况下做出决策。具体地说，我们令

$$g_n(x) = \begin{cases} 1 & \text{若 } \sum_{i=1}^{k} Y_{(i)}(x) \geq l \\ 0 & \text{若 } \sum_{i=1}^{k} Y_{(i)}(x) \leq k-l \\ -1 & \text{否则（无决策）}. \end{cases}$$

定义误差伪概率为：

$$L_n = \mathbf{P}\{g_n(X) \notin \{-1, Y\} \mid D_n\},$$

即,我们得到一个正确分类 X 的决策的概率。显然,$L_n \leq \mathbf{P}\{g_n(X) \neq Y \mid D_n\}$（标准误差概率）。后一不等式仅是看起来很有趣，因为 L_n 并未考虑无决策时的概率。我们扩展定理5.2以证明如下命题（问题5.35）：

定理 5.9 *对 (k,l)-最近邻规则，误差伪概率 L_n 满足*

$$\begin{aligned} \lim_{n\to\infty} \mathbf{E}\{L_n\} =& \mathbf{E}\{\eta(X)\mathbf{P}\{\text{Binomial}(k,\eta(X)) \leq k-l \mid X\} \\ &+ (1-\eta(X))\mathbf{P}\{\text{Binomial}(k,\eta(X)) \geq l \mid X\}\} \\ \stackrel{def}{=}& L_{k,l}. \end{aligned}$$

上述结果是与分布无关的。注意 k 为奇数的 k-最近邻规则对应于 $L_{k,(k+1)/2}$。极限 $L_{k,l}$ 本身是无趣的，但 Devijver（1979）证明 $L_{k,l}$ 具有关于贝叶斯误差 L^* 的信息。

定理 5.10 *对所有分布，且 k 为奇数，*

Devijver（1979）

$$L_{k,k} \leq L_{k,k-1} \leq \cdots \leq L_{k,\lceil k/2\rceil+1} \leq L^* \leq L_{k,\lceil k/2\rceil} = L_{k\text{NN}}.$$

另外，

$$\frac{L_{k\text{NN}} + L_{k,\lceil k/2\rceil+1}}{2} \leq L^* \leq L_{k\text{NN}}.$$

该定理（参考问题5.34）证明 L^* 被夹在 k-最近邻规则的渐进误差概率 $L_{k\text{NN}}$ 和“网球”规则之间，其中“网球”规则要求在 k 个最近邻中两个不同类的票数差值至少为 2。若 L_n 接近其极限，且若我们可以估计 L_n（见误差估计一章），则我们可以使用 Devijver 不等式来得到贝叶斯误差 L^* 的估计。更多结果可见 Loizou 和 Maybank（1987）。

作为 Devijver 不等式的推论，注意

$$L_{k\text{NN}} - L^* \leq \frac{L_{k\text{NN}} - L_{k,\lceil k/2\rceil+1}}{2}.$$

我们有

$$\begin{aligned} L_{k,l} =& L^* + \mathbf{E}\{|1-2\min(\eta(X),1-\eta(X))| \\ & \times \mathbf{P}\{\text{Binomial}(k,\min(\eta(X),1-\eta(X))) \geq l \mid X\}\}, \end{aligned}$$

因此

$$\begin{aligned} & L_{k,l} - L_{k,l+1} \\ & = \mathbf{E}\{|1-2\min(\eta(X),1-\eta(X))| \\ & \times \mathbf{P}\{\text{Binomial}(k,\min(\eta(X),1-\eta(X))) = l \mid X\}\} \\ & = \mathbf{E}\{|1-2\min(\eta(X),1-\eta(X))| \\ & \times \binom{k}{l} \min(\eta(X),1-\eta(X))^l (1-\min(\eta(X),1-\eta(X)))^{k-l}\Big\} \\ & \leq \binom{k}{l} \frac{l^l (k-l)^{k-l}}{k^k} \\ & \text{(因为} u^l(1-u)^{k-l} \text{在}[0,1]\text{的}u = l/k \text{点处达到最大值)} \\ & \leq \sqrt{\frac{k}{12\pi l(k-l)}} \\ & \text{(使用} \binom{k}{l} \leq \frac{k^k}{l^l(k-l)^{k-l}} \frac{1}{\sqrt{2\pi}} \sqrt{\frac{k}{l(k-l)}}\text{，通过 Stirling 公式).} \end{aligned}$$

若[19] $l = \lceil k/2 \rceil$，因此我们得到

$$\begin{aligned} L_{k\text{NN}} - L^* &\leq \frac{1}{2\sqrt{2\pi}} \sqrt{\frac{k}{\lceil k/2 \rceil \lfloor k/2 \rfloor}} \\ &= \sqrt{\frac{k}{2\pi\left(k^2-1\right)}} \approx \frac{0.398942}{\sqrt{k}}, \end{aligned}$$

其是定理5.6的改进。许多其他不等式也可通过这种方式推出。

19: 许有问题（with）

5.12 问题与练习

问题 5.1

令 $\|\cdot\|$ 为 $\mathcal{R}^d$ 的任意范数，并根据距离 $\rho(x,z) = \|x-z\|$ 定义 k-最近邻规则。证明定理5.1和5.2仍成立。

[提示：仅 Stone 引理需要进行调整。证明中使用的圆锥 $C(x,\pi/6)$ 现在换成如下性质的集合：x 和 z 属于同一个集合当且仅当

$$\left\| \frac{x}{\|x\|} - \frac{z}{\|z\|} \right\| < 1.]$$

问题 5.2

最近邻规则是否存在一个分布，使得 $\sup_{n\geq 1} \mathbf{E}\{L_n\} > 1/2$?

问题 5.3

证明 $L_{3\text{NN}} \le 1.32L^*$ 与 $L_{5\text{NN}} \le 1.22L^*$。

问题 5.4

证明若 C^* 为 $\mathcal{R}^d$ 的紧集，且 C 为概率测度 μ 的支撑集，

$$\sup_{x \in C \cap C^*} \|X_{(1)} - x\| \to 0$$

以概率 1 成立，其中 $X_{(1)}$ 为 $X_1, \ldots, X_n$ 中 x 的最近邻。

Wagner（1971）

问题 5.5

令 μ 为给定 $Y = 0$ 时 X 的概率测度，并令 ν 为给定 $Y = 1$ 时的测度。假设 X 为实值且 $\mathbf{P}\{Y = 0\} = \mathbf{P}\{Y = 1\} = 1$。找到一对 (ν, μ) 使得：

(1) $\text{support}(\mu) = \text{support}(\nu)$;
(2) $L^* = 0$。

这说明了 $L^* = 0$ 并未给出关于 μ 和 ν 支撑集的太多信息。

问题 5.6

考虑一特定 (X, Y) 分布下的 $(2k+1)$-最近邻规则，该分布具 $\eta(x) \equiv p$（常数），且 Y 与 X 相独立。本练习寻求当 $p \downarrow 0$ 时 $L_{(2k+1)\text{NN}}$ 的表现。

(1) 对固定整数 $l > 0$，当 $p \downarrow 0$，证明

$$\mathbf{P}\{\text{Binomial}(2k, p) \ge l\} \sim \binom{2k}{l} p^l.$$

(2) 使用 $L_{(2k+1)\text{NN}}$ 的一个方便表示来证明随着 $p \downarrow 0$，

$$L_{(2k+1)\text{NN}} = p + \left(\binom{2k}{k} + \binom{2k}{k+1} \right) p^{k+1} + o\left(p^{k+1}\right).$$

问题 5.7

Das Gupta 和 Lin（1980）对数据 $X \in \mathcal{R}$ 提出了如下规则。假设 X 是非原子的（nonatomic）。首先，根据升序重排 $X_1, \ldots, X_n, X$，并记 $X_{(1)} < X_{(2)} < \cdots < X_{(i)} < X < X_{(i+1)} < \cdots < X_{(n)}$ 为有序集。置换 Y_i 使得 $Y_{(j)}$ 为 $X_{(j)}$ 的标签。在 $\{Y_{(i)}, Y_{(i+1)}\}, \{Y_{(i-1)}, Y_{(i+2)}\}, \ldots$ 中进行投票直至首次出现 $(Y_{(i-j)} = Y_{(i+j+1)})$，这时我们可判定类为 $g_n(X) = Y_{(i-j)} = Y_{(i+j+1)}$。该规则在 x-轴的单调变换下是不变的。

(1) 若记 L 为渐进期望误差概率，证明对任非原子 X，

$$L = \mathbf{E}\left\{ \frac{\eta(X)(1-\eta(X))}{1 - 2\eta(X)(1-\eta(X))} \right\}.$$

(2) 证明 L 等同于如下规则，其中 $X \in \mathcal{R}^d$，我们轮流考虑 $2-\text{NN}$，

$4-\mathrm{NN}$，$6-\mathrm{NN}$ 等规则，直至首次出现一个没有正负票数相同的 $2k-\mathrm{NN}$ 规则才停止。为了方便，假设 X 具有一个密度（若具有一个好的距离——规则可避免出现票数相同情况，则可以忽略此假设）[20]。

(3) 证明 $L - L_{\mathrm{NN}} \geq (1/2)(L_{3\mathrm{NN}} - L_{\mathrm{NN}})$，因此 $L \geq (L_{\mathrm{NN}} + L_{3\mathrm{NN}})/2$。

(4) 证明 $L \leq L_{\mathrm{NN}}$。因此，该规则的表现介于 $2-\mathrm{NN}$ 规则与 $3-\mathrm{NN}$ 规则之间。

20: 许有问题：(with a good distance-tie breaking rule, this may be dropped)

问题 5.8

令 Y 与 X 相互独立，且 $\eta(x) \equiv p$（常数）。考虑一个具权重 $(2m+1, 1, 1, \ldots, 1)$($2k$ 个“1”)加权 $(2k+1)$-最近邻规则,其中 $k-1 \geq m \geq 0$。对 $m=0$，我们得到 $(2k+1)-\mathrm{NN}$ 规则。令 $L(k,m)$ 为渐进误差概率。

(1) 使用问题5.6的结果，证明随着 $p \downarrow 0$,

$$L(k,m) = p + \binom{2k}{k-m} p^{k-m+1} + \binom{2k}{k+m+1} p^{k+m+1} + o\left(p^{k-m+1}\right)$$

成立。说明了在这族规则中，对小的 p，规则优越性的度量取决于 $k-m$。

(2) 令 $\delta > 0$ 足够小,且 $p = 1/2 - \delta$。证明若 X 为 $binomial(2k, 1/2-\delta)$ 随机变量且 Z 为 $binomial(2k, 1/2)$ 随机变量，则对固定的 l，随着 $\delta \downarrow 0$,

$$\mathbf{P}\{X \geq l\} = \mathbf{P}\{Z \geq l\} - 2k\delta \mathbf{P}\{Z = l\} + o\left(\delta^2\right),$$

与

$$\mathbf{P}\{X \leq l\} = \mathbf{P}\{Z \leq l\} + 2k\delta \mathbf{P}\{Z = l+1\} + o\left(\delta^2\right)$$

成立。

(3) 证明对固定的 k，m，随着 $\delta \downarrow 0$，有

$$\begin{aligned} L(k,m) = &\frac{1}{2} - 2\delta^2 (k\mathbf{P}\{Z = k+m\} + k\mathbf{P}\{Z = k+m+1\} \\ &+ \mathbf{P}\{Z \leq k+m\} - \mathbf{P}\{Z \geq k+m+1\}) + o\left(\delta^2\right). \end{aligned}$$

(4) 取权重向量 w，k 固定且 $m = \lfloor 10\sqrt{k} \rfloor$，将其与另一权重向量 w' 进行比较，后者具有 $k/2$ 个分量且当 $p \downarrow 0$ 与 $p \uparrow 1/2$ 时，$m = \lfloor \sqrt{k/2} \rfloor$ 成立。假设 k 值非常大（但固定）。证明当 $p \downarrow 0$ 时，w 相对较优；当 $p \uparrow 1/2$ 时，w' 相对较优。在后一情况，对一固定 $c > 0$，由中心极限定理可得

$$k\mathbf{P}\{Z = k+m\} + k\mathbf{P}\{Z = k+m+1\} \sim 8\sqrt{k} \frac{1}{\sqrt{2\pi}} e^{-2c^2},$$

当 $k \to \infty$ 时。

(5) 证明存在不同的权重向量 w，w'，使得存在一对 (X,Y) 分布，这两个分布的渐进误差概率顺序不同（differently ordered）。因此，关于误差概率的 $\mathscr{W}$ 不是全序的（totally ordered）。[21]

21: 许有问题：Conclude that there exist different weight vectors w, w' for which there exists a pair of distributions of (X,Y) such that their asymptotic error probabilities are differently ordered. Thus, W is not totally ordered with respect to the probability of error.

问题 5.9

Patrick 和 Fisher（1970）在两类中的每个类找到第 k 个最近邻，并根据哪一个最近（nearest）来进行分类。证明该规则等价于一个 $(2k-1)$-最近邻规则。[22]

22: 许有问题：Patrick and Fisher (1970) find the k-th nearest neighbor in each of the two classes and classify according to which is nearest. 其中“which is nearest”指的是哪些点?

问题 5.10

Rabiner 等人（1979）推广问题5.9中的规则，使得其根据到每个类第 k 个最近邻的平均距离进行分类。假设 X 具有一个密度。对固定 k，求渐进误差概率。

问题 5.11

若 N 为均值为 0，方差为 1 的正态分布随机变量（$normal(0,1)$），则 $\mathbf{E}\{|N|\} = \sqrt{2/\pi}$。证明它。

问题 5.12

证明若 $L_{9\text{NN}} = L_{(11)\text{NN}}$，则 $L_{(99)\text{NN}} = L_{(111)\text{NN}}$。

问题 5.13

证明对任意分布，

$$L_{3\text{NN}} \leq \left(\frac{7\sqrt{7}+17}{27\sqrt{3}}+1\right) L^* \approx 1.3155\ldots L^*$$

成立。

[提示：求最小常数 a 使得 $L_{3\text{NN}} \leq L^*(1+a)$，使用 $L_{3\text{NN}}$ 的二项式尾分布表示。[23]]

Devroye（1981b）

23: 许有问题：Find the smallest constant a such that $L_{3\text{NN}} \leq L^*(1+a)$ using the representation of $L_{3\text{NN}}$ in terms of the binomial tail.

问题 5.14

证明若 X 具有密度 f，则对任 $u > 0$，

$$\lim_{n\to\infty} \mathbf{P}\left\{n^{1/d}\left\|X_{(1)}(X) - X\right\| > u \mid X\right\} = e^{-f(X)vu^d}$$

按概率 1 成立，其中 $v = \int_{S_{0,1}} dx$ 为空间 $\mathcal{R}^d$ 中单位球的体积。

问题 5.15

考虑一个规则在所有 Y_i 上取多数票，对其 $\|X_i - x\| \leq (c/vn)^{1/d}$，其中 $v = \int_{S_{0,1}} dx$ 为单位球的体积，且 $c > 0$ 固定。若出现票数相同，则令 $g_n(x) = 0$。

(1) 若 X 具有密度 f，证明 $\liminf_{n\to\infty} \mathbf{E}\{L_n\} \geq \mathbf{E}\left\{\eta(X)e^{-cf(X)}\right\}$。[提示：使用不等式 $\mathbf{E}\{L_n\} \geq \mathbf{P}\left\{Y=1, \mu_n\left(S_{X,c/vn}\right)=0\right\}$。]

(2) 若 Y 与 X 相互独立，且 $\eta \equiv p > 1/2$，则当 $p \uparrow 1$，证明

$$\frac{\mathbf{E}\left\{\eta(X)e^{-cf(X)}\right\}}{L^*} = \mathbf{E}\left\{e^{-cf(X)}\right\}\frac{p}{1-p} \uparrow \infty.$$

(3) 证明

$$\sup_{(X,Y):L^*>0} \frac{\liminf_{n\to\infty} \mathbf{E}\{L_n\}}{L^*} = \infty,$$

因此对这些简单规则，由 k-最近邻估计得到的与分布无关且具有形式 $\lim_{n\to\infty} \mathbf{E}\{L_n\} \leq c'L^*$ 的界并不存在。

Devroye（1981a）

问题 5.16

举例 $\eta(X) \equiv 1/2 - 1/(2\sqrt{k})$，证明界 $L_{k\mathrm{NN}} - L^* \leq 1/\sqrt{ke}$ 对大的 k，不能再进一步提升。即存在一个分布序列（记下标为 k），当 $k \to \infty$ 时，有

$$L_{k\mathrm{NN}} - L^* \geq \frac{1-o(1)}{\sqrt{k}}\mathbf{P}\{N \geq 1\},$$

其中 N 为 $normal(0,1)$ 随机变量。

问题 5.17

若 B 为 $binomial(n,p)$，则

$$\sup_p \mathbf{P}\{B=i\} \leq \sqrt{\frac{n}{2\pi i(n-i)}}, 0 < i < n.$$

问题 5.18

证明对 $k \geq 3$，

$$\frac{\sqrt{k(k+1)}}{k-1} \leq \left(1+\frac{1}{2k}\right)\left(1+\frac{3}{2k}\right) = 1 + \frac{2}{k} + \frac{3}{4k^2}.$$

问题 5.19

证明存在一个 (X,Y) 分布序列（记下标为 k），其中 Y 与 X 相互独立且 $\eta(x) \equiv p$（p 仅依赖于 k）使得

$$\liminf_{n\to\infty}\left(\frac{L_{k\mathrm{NN}} - L^*}{L^*}\right)\sqrt{k} \geq \gamma = 0.339942\ldots,$$

其中 γ 为定理5.8的常数（Devroye（(1981b)）。

[提示：验证定理5.8的证明，但使用如下不等式给出界。Slud 不等式（见附录中引理 A.6）在这或许有用。]

问题 5.20

考虑一个具权重 $1, \rho, \rho^2, \rho^3, \ldots$ 的加权最近邻规则，其中 $\rho < 1$。证明对任意分布，期望误差概率趋向极限 $L(\rho)$。

[提示：在 k 处截断，其中 k 值很大且固定。并验证尾部具有渐进可忽略权重。[24]]

24: 许有问题：Truncate at k fixed but large, and argue that the tail has asymptotically negligible weight.

问题 5.21

（接续）$L(\rho)$ 如上题，证明当 $\rho < 1/2$ 时，$L(\rho) = L_{\mathrm{NN}}$。

问题 5.22

（接续）证明或反证：当 ρ 从 1/2 递增至 1，$L(\rho)$ 从 L_{NN} 至 L^* 单调递减。

本问题很难

问题 5.23

证明对具权重 $(1, \rho, \rho^2)$, $0 < \rho < 1$ 的加权 NN 规则，若 $\rho < (\sqrt{5}-1)/2$，则渐进误差概率为 L_{NN}；若 $\rho > (\sqrt{5}-1)/2$，则渐进误差概率为 $L_{3\mathrm{NN}}$。

问题 5.24

是否存在 k（可多个），使得 $k-\mathrm{NN}$ 规则是容许的？即是否存在一个 (X,Y) 分布使得 $k-\mathrm{NN}$ 规则的 $\mathbf{E}\{L_n\}$ 小于任一 $k'-\mathrm{NN}$ 规则的 $\mathbf{E}\{L_n\}$，其中 $k' \neq k$，$\forall n$？

[提示：注意若其对任 n 成立，则对其极限也必定成立。由此推断，对任意这种分布，$\eta(x) \in \{0, 1/2, 1\}$ 按概率 1 成立。[25]]

本题很难

25: 许有问题：Note that if this is to hold for all n, then it must hold for the limits. From this, deduce that with probability one, $\eta(x) \in \{0, 1/2, 1\}$ for any such distribution.

问题 5.25

对每一固定 n 和奇数 k（$n > 1000k$），求一 (X,Y) 分布使得 $k-\mathrm{NN}$ 规则的 $\mathbf{E}\{L_n\}$ 小于任一 $k'-\mathrm{NN}$ 规则的 $\mathbf{E}\{L_n\}$，其中 $k' \neq k$，k' 为奇数。因此，对一给定 n，不存在可以考虑先验忽略的 k 值。

问题 5.26

令 X 为 $[0,1]$ 上的均匀分布，$\eta(x) \equiv x$，并 $\mathbf{P}\{Y=0\} = \mathbf{P}\{Y=1\} = 1/2$。证明对最近邻规则，有

$$\mathbf{E}\{L_n\} = \frac{1}{3} + \frac{3n+5}{2(n+1)(n+2)(n+3)}.$$

Cover 和 Hart（1967）；Peterson（1970）

问题 5.27

Cover（1968a）

对最近邻规则，若 X 具有一个密度，则

$$\mathbf{E}\{L_n\} = \frac{1}{3} + \frac{3n+5}{2(n+1)(n+2)(n+3)}.$$

Cover（1968a）

问题 5.28

令 X 具有一个在 $[0,1]$ 上的密度 $f \geq c > 0$，并假设 f_0''' 和 f_1''' 存在且一直有界。证明对最近邻规则，$\mathbf{E}\{L_n\} = L_{\text{NN}} + O\left(1/n^2\right)$。对 d 维问题，该结果由 Psaltis、Snapp 和 Venkatesh（1994）推广。

问题 5.29

证明在具 $L_{k\text{NN}} \leq (1 + a/\sqrt{k})L^*$ 形式的界中，$L_{k\text{NN}} \leq (1 + \sqrt{2/k})L^*$ 是最优的，其对所有 $k \geq 1$ 同时成立。

问题 5.30

Devroye（1981b）

证明对任 $k \geq 3$，$L_{k\text{NN}} \leq (1 + \sqrt{1/k})L^*$ 成立。

问题 5.31

26: 许有问题：Show that this is asymptotically not better than if we had used $(x(1), x(2))$.

令 $x = (x(1), x(2)) \in \mathcal{R}^2$。考虑基于向量 $\left(x^3(1), x^7(2), x(1)x(2)\right)$ 的最近邻规则。证明在渐近意义上这并不比使用 $(x(1), x(2))$ 好。[26]通过例子证明 $\left(x^2(1), x^3(2), x^6(1)x(2)\right)$ 可能产生一个比用 $(x(1), x(2)$ 更差的渐进误差概率。

问题 5.32

统一次序统计量（Uniform Order Statistics

令 $U_{(1)} < \cdots < U_{(n)}$ 为 n 个在 $[0,1]$ 上的独立同分布均匀随机变量的次序统计量。证明下述命题：

(1) $U_{(k)}$ 为 $beta(k, n+1-k)$，即 $U_{(k)}$ 具密度

$$f(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}, 0 \leq x \leq 1.$$

(2)

$$\mathbf{E}\left\{U_{(k)}^a\right\} = \frac{\Gamma(k+a)\Gamma(n+1)}{\Gamma(k)\Gamma(n+1+a)}, \quad \text{for any } a > 0.$$

(3) 对 $a \geq 1$,

$$1 - \frac{a}{n} \leq \frac{\mathbf{E}\left\{U_{(k)}^a\right\}}{(k/n)^a} \leq 1 + \frac{\psi(a)}{k},$$

其中 $\psi(a)$ 仅为 a 的函数。

Royall（1966）

问题 5.33

Dudani 规则

Dudani（1976）提出一个加权 $k-\text{NN}$ 规则，其中 $Y_{(i)}(x)$ 具权重

$$\left\|X_{(k)}(x)-x\right\|-\left\|X_{(i)}(x)-x\right\|, \quad 1 \le i \le k.$$

若 X 具有一个密度，为什么可以粗略认为该规则与一个对第 i 个最近邻具权重 $1-(i/k)^{1/d}$ 的加权最近邻规则等价?

[提示：若 μ 为 X 的概率测度，则

$$\mu\left(S_{x,\|X_{(1)}(x)-x\|}\right), \ldots, \mu\left(S_{x,\|X_{(k)}(x)-x\|}\right)$$

的分布像 $U_{(1)}, \ldots, U_{(k)}$，其中 $U_{(1)} < \cdots < U_{(n)}$ 为 $[0,1]$ 上 n 个独立同分布随机变量的次序统计量。用一个好的局部预测替换 μ，并使用上一习题的结果。]

问题 5.34

分两部分证明 Devijver 定理（定理5.10）：首先为“网球规则”构建不等式 $L^* \ge L_{k,\lceil k/2\rceil-1}$；然后证明单调性。

问题 5.35

对 (k,l) 最近邻规则证明定理5.9。

问题 5.36

令 R 为 $(2,2)$-最近邻规则的渐进误差概率。证明 $R = \mathbf{E}\{2\eta(X)(1-\eta(X))\} = L_{\text{NN}}$。

问题 5.37

Devijver 和 Kittler（1982）

对最近邻规则，证明对任意分布，

$$\begin{aligned}\lim_{n\to\infty} \mathbf{P}\{g_n(X)=0, Y=1\} &= \lim_{n\to\infty} \mathbf{P}\{g_n(X)=1, Y=0\} \\ &= \mathbf{E}\{\eta(X)(1-\eta(X))\}.\end{aligned}$$

因此，两种类型的误差的可能性是相同的。[27]

27: Thus, errors of both kinds are equally likely.

问题 5.38

令 $\mathbf{P}\{Y=1\} = \mathbf{P}\{Y=0\} = 1/2$，若 $Y=1$，令 X 为随机有理数（如在第 5.9 节的定义）使得每个有理数具有正概率；若 $Y=1$，令 X 服从 $[0,1]$ 上的均匀分布。证明对任意无理数 $x \in [0,1]$，当 $n \to \infty$ 时，$\mathbf{P}\left\{Y_{(1)}(x)=1\right\} \to 0$ 成立；而对任有理数 $x \in [0,1]$，$\mathbf{P}\left\{Y_{(1)}(x)=0\right\} \to 0$ 成立。

问题 5.39

令 $X_1,\ldots,X_n$ 为独立同分布随机变量且具有相同的密度。证明对固定的 $k>0$,

$$n\mathbf{P}\left\{X_3 \text{ is among the } k \text{ nearest neighbors of } X_1 \text{ and } X_2 \text{ in } \{X_3,\ldots,X_n\}\right\} \to 0.$$

证明当 k 随 n 变化且有 $k/\sqrt{n}\to 0$ 时，该结果仍成立。

问题 5.40

不完全训练（Lugosi（1992）

令 $(X,Y,Z),(X_1,Y_1,Z_1),\ldots,(X_n,Y_n,Z_n)$ 为 $\mathcal{R}^d\times\{0,1\}\times\{0,1\}$ 中的独立同分布的三元组序列，其中 $\mathbf{P}\{Y=1\mid X=x\}=\eta(x)$ 和 $\mathbf{P}\{Z=1\mid X=x\}=\eta'(x)$。若 X_j 为 $X_1,\ldots,X_n$ 中 X 的最近邻，令 $Z_{(1)}(X)$ 为 Z_j。证明

$$\lim_{n\to\infty}\mathbf{P}\left\{Z_{(1)}(X)\neq Y\right\}=\mathbf{E}\left\{\eta(X)+\eta'(X)-2\eta(X)\eta'(X)\right\}.$$

问题 5.41

改进引理5.3中的界为 $\gamma_d\le 3^d-1$。

问题 5.42

证明若 $\left\{C\left(x_1,\pi/6\right),\ldots,C\left(x_{\gamma_d},\pi/6\right)\right\}$ 为覆盖 $\mathcal{R}^d$ 的圆锥的集合，则 $\gamma_d\ge 2^d$。

问题 5.43

回忆 $L_{k\text{NN}}=\mathbf{E}\left\{\alpha_k(\eta(X))\right\}$，其中

$$\alpha_k(p)=\min(p,1-p)+|1-2\min(p,1-p)|\times\mathbf{P}\left\{\text{Binomial}(k,\min(p,1-p))>\frac{k}{2}\right\},$$

证明对任固定的 p, 当 $k\to\infty$, 有 $\mathbf{P}\left\{\text{Binomial}(k,\min(p,1-p))>\frac{k}{2}\right\}\downarrow 0$（单调性更加难以证明）。接着如何证明 $\lim_{k\to\infty}L_{k\text{NN}}=L^*$?

问题 5.44

证明判别 $g_n(x)=Y_{(8)}(x)$ 的规则与 $g_n(x)=Y_{(3)}(x)$ 的规则的渐进误差概率相等。

问题 5.45

证明对所有分布，$L_{5\text{NN}}=\mathbf{E}\left\{\psi_5(\eta(X)(1-\eta(X))\right\}$ 成立，其中 $\psi_5(u)=u+u^2+12u^3$。

问题 5.46

证明对任意分布，

$$L_{5\text{NN}} \geq \frac{L_{\text{NN}}}{2} + \frac{L_{\text{NN}}^2}{4} + \frac{3L_{\text{NN}}^3}{2}$$

且

$$L_{3\text{NN}} \geq \frac{L_{\text{NN}}}{2} + L_{\text{NN}}^2.$$

问题 5.47

令 $X_{(1)}$ 为 x 在 $X_1, \ldots, X_n$ 中的最近邻。构建一个例子使得对任意 $x \in \mathcal{R}^d$，有 $\mathbf{E}\{\|X_{(1)} - x\|\} = \infty$。（因此，我们必须避免引理5.1中的均值收敛。）令 $X_{(1)}$ 为 X 在 $X_2, \ldots, X_n$ 中的最近邻。构建一个分布使得 $\mathbf{E}\{\|X_{(1)} - X_1\|\} = \infty, \forall n$。

问题 5.48

考虑一个具权重 $(w_1, \ldots, w_k)$ 的加权最近邻规则。定义一个新的权重向量 $(w_1, w_2, \ldots, w_{k-1}, v_1, \ldots, v_l)$，其中 $\sum_{i=1}^{l} v_i = w_k$。因此，通过“切割”操作，权重向量被进行部分排序。假设所有权重是非负的。令二者的渐进期望误差概率分别为 L 和 L'。判断如下命题真假：对任意 (X, Y) 分布，$L' \leq L$。

问题 5.49

给定 $X_1, \ldots, X_n \in \mathcal{R}^d$，若圆心为 $(X_i + X_j)/2$ 且半径为 $\|X_i - X_j\|/2$ 的圆不包含 X_k，$k \neq i, j$，我们称 X_i 与 X_j 为 Gabriel 邻。显然，若 X_j 为 X_i 的最近邻，则 X_i 和 X_j 为 Gabriel 邻。证明若 X 具有一个密度且 $X_1, \ldots, X_n$ 为独立同分布的随机变量，且其分布均来自 X，则 X_1 的 Gabriel 邻的个数趋向于 2^d，当 $n \to \infty$。

Gabriel 邻

Gabriel 和 Sokal（1969）；Matula 和 Sokal（1980）。

Devroye（1988c）

问题 5.50

简单定义 Gabriel 邻规则为在所有 Y_i 上对 X 在 $X_1, \ldots, X_n$ 中的 Gabriel 邻取多数票。通过掷硬币的方式解决出现票数相同的情况。令 L_n 为 Gabriel 规则的条件误差概率。使用上一习题的结果证明若 L^* 为贝叶斯误差，则

(1) $\lim_{n\to\infty} \mathbf{E}\{L_n\} = 0$ 若 $L^* = 0$；
(2) $\limsup_{n\to\infty} \mathbf{E}\{L_n\} < L_{\text{NN}}$ 若 $L^* > 0, d > 1$；
(3) $\limsup_{n\to\infty} \mathbf{E}\{L_n\} \leq cL^*$ 对某 $c < 2$，若 $d > 1$。

对（3），确定最佳可能的 c 值。

[提示：使用定理5.8并尝试获得 $d = 2$ 时对 $\mathbf{P}\{N_X \geq 3\}$ 的下界，其中 N_X 为 X 中 $X_1, \ldots, X_n$ 中 Gabriel 邻的个数。]

Gabriel 邻规则

6 一致性

6.1 普遍一致性

若我们给定一个训练数据序列 $D_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$，则最优的分类函数可得到贝叶斯误差概率 L^*。一般来说，我们不期望得到一个正好实现贝叶斯误差概率的函数，但可以构建一个分类函数序列 $\{g_n\}$，即分类规则，使得误差概率

$$L_n = L(g_n) = \mathbf{P}\{g_n(X, D_n) \neq Y \mid D_n\}$$

以大的概率（即对“大多数”D_n）可任意接近 L^*。该思想可形式化为概念一致性：

弱一致性与强一致性

定义 6.1 *若*

$$\mathbf{E}L_n = \mathbf{P}\{g_n(X, D_n) \neq Y\} \to L^* \quad \text{当} n \to \infty,$$

则称分类规则对某一 (X,Y) 分布是弱一致的（或渐进贝叶斯风险有效的；若

$$\lim_{n\to\infty} L_n = L^*$$

按概率 1 成立，则称为强一致的。

备注 6.1 一致性被定义为 L_n 的期望值收敛至 L^*。由于 L_n 是界于 L^* 和 1 之间的随机变量，该收敛性等价于 L_n 至 L^* 的概率收敛性。具体来说，对任 $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\{L_n - L^* > \epsilon\} = 0 \text{。}$$

显然，由于几乎处处收敛蕴含着按概率收敛，因此强一致性蕴含着弱一致性。

一致性规则保证通过增加数据量，误差概率在最优可达的极小距离下的概率可任意接近 1。[29]直观地说，规则最终以很高地概率可从大量训练数据学到最优决策。强一致性意味着通过使用更多数据，除了一组概率全为零的序列外，对每一训练序列，误差概率任意接近最优值。

29: 许有问题：A consistent rule guarantees that by increasing the amount of data the probability that the error probability is within a very small distance of the optimal achievable gets arbitrarily close to one.

决策规则对某一类 (X,Y) 分布可能是一致的，但对其他分布可能不是一致的。显然我们期望存在一个规则对一大类分布是一致的。因为在许多情况下我们并未具有任何关于分布的先验知识，因此存在一个对所有分布均表现出优越性能的规则是至关重要的。这个非常强的普遍优越性假设可形式化为：

普遍一致性

定义 6.2 *一个决策规则序列被称为普遍（强）一致的，如果它对任意 (X,Y) 对的分布是（强）一致的。*

在本章中我们证明该普遍一致性分类规则确实存在。首先，对一些非常“陌生”且看起来难以学习的分布，该结论可能令人惊讶。例如，令 X 为在 $[0,1]$ 上具概率 $1/2$ 的均匀分布，且令 X 按概率 $1/2$ 在有理数上是原子的（atomic）。[30]举例来说，若将所有有理数列举为 $r_1, r_1, , r_3, \ldots$，则 $\mathbf{P}\{X = r_i\} = 1/2^{i+1}$。若 X 为有理数，则令 $Y = 1$，否则令 $Y = 0$。显然，$L^* = 0$。若一分类规则 g_n 为一致的，则预测 X 的有理性出现错误的概率趋向于 0。这里应注意我们并未“检查” X 是否为有理数，而是应该仅仅基于给定的数据 D_n 做出决策。一个一致性规则的例子如下：若 X_k 为 $X_1, \ldots, X_n$ 中最接近 x 的点，则 $g_n(x, D_n) = Y_k$。有理数在 $[0,1]$ 上是稠密的，因此使得该结论更加令人讶异。见问题6.3。

30: 许有问题：let X be atomic on the rationals with probability $1/2$.

6.2 分类与回归估计

在这节中我们证明分类器规则的一致性可以从一致回归估计中推导而出。在许多情况下，后验概率 $\eta(x)$ 是从训练数据 D_n 中由某一函数 $\eta_n(x) = \eta_n(x, D_n)$ 估计得到。在这种情况下，插件规则

$$g_n(x) = \begin{cases} 0 & \text{若} \eta_n(x) \le 1/2 \\ 1 & \text{否则} \end{cases}$$

的误差概率 $L(g_n) = \mathbf{P}\{g_n(X) \neq Y \mid D_n\}$ 是一个随机变量。因此定理2.2的一个简单推论如下所示：

推论 6.1 *按如上定义的分类器 $g_n(x)$ 的误差概率满足不等式*

$$L(g_n) - L^* \le 2 \int_{\mathcal{R}^d} |\eta(x) - \eta_n(x)| \, \mu(dx) = 2\mathbf{E}\{|\eta(X) - \eta_n(X)| \mid D_n\} \text{。}$$

下一推论可由 Cauchy-Schwarz 不等式得到。

推论 6.2 *若*

$$g_n(x) = \begin{cases} 0 & \text{若} \eta_n(x) \le 1/2 \\ 1 & \text{否则,} \end{cases}$$

则它的误差概率满足

$$\mathbf{P}\{g_n(X) \neq Y \mid D_n\} - L^* \le 2\sqrt{\int_{\mathcal{R}^d} |\eta(x) - \eta_n(x)|^2 \, \mu(dx)} \text{。}$$

显然，$\eta(x) = \mathbf{P}\{Y = 1 \mid X = x\} = \mathbf{E}\{Y \mid X = x\}$ 仅是 Y 在 X 上的回归函数。因此，定理2.2最有趣的结果是只要回归函数估计 $\eta_n(x)$ 存在，其中 $\eta_n(x)$ 使得

$$\int |\eta(x) - \eta_n(x)|^2 \, \mu(dx) \to 0$$

按概率成立或按概率 1 成立，即蕴含着插件决策规则 g_n 分别是一致的或强一致的。

31: 许有问题：are estimated from data separately by some $\eta_{0,n}$ and $\eta_{1,n}$, respectively.

显然，从定理2.3，当概率 $\eta_0(x) = \mathbf{P}\{Y = 0 \mid X = x\}$ 和 $\eta_1(x) = \mathbf{P}\{Y = 1 \mid X = x\}$ 分别根据数据 D_n，由某 $\eta_{0,n}$ 和 $\eta_{1,n}$ 进行估计，[31]我们可以得到一个类似推论6.1的结论。一般来说，证明分类器规则的关键部分是重写规则为插件形式，并证明该近似函数到后验概率的 L_1-收敛

性。重写规则有一点自由性，因为对任意正函数 $\tau_n(x)$，我们可以有

$$g_n(x) = \begin{cases} 0 & \text{若} \eta_{1,n}(x) \le \eta_{0,n}(x) \\ 1 & \text{否则,} \end{cases} = \begin{cases} 0 & \text{若} \frac{\eta_{1,n}(x)}{\tau_n(x)} \le \frac{\eta_{0,n}(x)}{\tau_n(x)} \\ 1 & \text{否则。} \end{cases}$$

6.3 划分规则

许多重要的分类规则划分 $\mathcal{R}^d$ 为不相交元胞 $A_1, A_2, \ldots$ 并在每个元胞中根据多票原则（落在同一元胞内 X_i 的标签作为表决票）进行分类。更具体地说，

$$g_n(x) = \begin{cases} 0 & \text{若} \sum_{i=1}^n I_{\{Y_i=1\}} I_{\{X_i \in A(x)\}} \le \sum_{i=1}^n I_{\{Y_i=0\}} I_{\{X_i \in A(x)\}} \\ 1 & \text{否则,} \end{cases}$$

其中 $A(x)$ 为包含 x 的元胞。如果在包含 X 的元胞内标签 1 的数量没有超过标签 0 的数量，则该决策为 0；反之亦然。本节中考虑的划分可随 n 变化，它们也可能依赖于点 $X_1, \ldots, X_n$，但我们假设标签在构造划分时并不重要。下一定理是对这些划分规则的一般一致性结果。它要求划分应具有两个性质：首先，元胞应该尽量小使得分布的局部变化能被识别；其次，元胞也应该足够大以包含大量的点，使得标签平均是有效的。记 $diam(A)$ 为集合 A 的直径，即

$$\text{diam}(A) = \sup_{x,y \in A} \|x - y\| \text{。}$$

令

$$N(x) = n\mu_n(A(x)) = \sum_{i=1}^n I_{\{X_i \in A(x)\}}$$

为与 x 同落在同一元胞的 X_i 的个数。下述定理的条件要求随机元胞——根据 X 的分布选择——具有小的直径，并且大概率包含许多点。

定理 6.1 *考虑如上定义的划分分类规则。若*

(1) $diam(A(X)) \to 0$ 按概率收敛,
(2) 按概率 $N(X) \to \infty$ 成立,

则

$$\mathbf{E}\{L_n\} \to L^* \text{。}$$

证. 定义 $\eta(x) = \mathbf{P}\{Y = 1 \mid X = x\}$。从引论6.1知我们仅需证明 $\mathbf{E}\{|\widehat{\eta}_n(X) - \eta(X)|\} \to 0$，其中

$$\widehat{\eta}_n(x) = \frac{1}{N(x)} \sum_{i:X_i \in A(x)} Y_i \text{。}$$

引入 $\bar{\eta}(x) = \mathbf{E}\{\eta(X) \mid X \in A(x)\}$。通过三角不等式，有

$$\mathbf{E}\{|\widehat{\eta}_n(X) - \eta(X)|\} \le \mathbf{E}\{|\widehat{\eta}_n(X) - \bar{\eta}(X)|\} + \mathbf{E}\{|\bar{\eta}(X) - \eta(X)|\} \text{。}$$

通过对随机变量 $N(x)$ 加条件,容易看到 $N(x)\widehat{\eta}_n(x)$ 的分布为 $B(N(x), \bar{\eta}(x))$，一个具参数 $N(x)$ 和 $\bar{\eta}(x)$ 的二项分布随机变量。因此，由 Cauchy-

Schwarz 不等式有

$$\begin{aligned}
&\mathbf{E}\left\{|\ \widehat{\eta}_n(\ X)-\bar{\eta}(X)||X, I_{\{X_1\in A(X)\}},\ldots,I_{\{X_n\in A(X)\}}\right\}\\
&\quad\leq \mathbf{E}\left\{\left|\frac{B(N(X),\bar{\eta}(X))}{N(X)}-\bar{\eta}(X)\right| I_{\{N(X)>0\}}\mid X, I_{\{X_1\in A(X)\}},\ldots,I_{\{X_n\in A(X)\}}\right\}\\
&\qquad + I_{\{N(X)=0\}}\\
&\quad\leq \mathbf{E}\left\{\sqrt{\frac{\bar{\eta}(X)(1-\bar{\eta}(X))}{N(X)}} I_{\{N(X)>0\}}\mid X, I_{\{X_1\in A(X)\}},\ldots,I_{\{X_n\in A(X)\}}\right\}\\
&\qquad + I_{\{N(X)=0\}}\ 。
\end{aligned}$$

取期望，我们可见对任意 k，

$$\begin{aligned}
\mathbf{E}\left\{|\widehat{\eta}_n(X)-\bar{\eta}(X)|\right\} &\leq \mathbf{E}\left\{\frac{1}{2\sqrt{N(X)}} I_{[N(X)>0\}}\right\}+\mathbf{P}\{N(X)=0\}\\
&\leq \frac{1}{2}\mathbf{P}\{N(X)\leq k\}+\frac{1}{2\sqrt{k}}+\mathbf{P}\{N(X)=0\}\ ,
\end{aligned}$$

通过使 k 足够大和使用条件（2），期望值可变得很小。

对 $\epsilon>0$，找到一个一致连续的 $[0,1]$-值函数 η_ϵ，该函数在有界集 C 中取非零值，在 C 之外取 0，使得 $\mathbf{E}\left\{|\eta_\epsilon(X)-\eta(X)|\right\}<\epsilon$。接着，我们利用三角不等式：

$$\begin{aligned}
\mathbf{E}\{|\bar{\eta}(X)-\eta(X)|\} \leq &\mathbf{E}\left\{|\bar{\eta}(X)-\bar{\eta}_\epsilon(X)|\right\}\\
&+\mathbf{E}\left\{|\bar{\eta}_\epsilon(X)-\eta_\epsilon(X)|\right\}\\
&+\mathbf{E}\left\{|\eta_\epsilon(X)-\eta(X)|\right\}\\
&=I+II+III\ ,
\end{aligned}$$

其中 $\bar{\eta}_\epsilon(x)=\mathbf{E}\left\{\eta_\epsilon(X)\mid X\in A(x)\right\}$。显然通过选定 η_ϵ，$III<\epsilon$。由于 η_ϵ 是一致连续的，我们可以找到一个 $\theta=\theta(\epsilon)>0$ 使得

$$II\leq\epsilon+\mathbf{P}\{\mathrm{diam}(A(X))>\theta\}\ 。$$

因此，通过条件（1），对足够大的 n，$II<2\epsilon$ 成立。最后，$I\leq III<\epsilon$。将这些步骤结合起来可证该定理。 [证毕]

6.4 直方图规则

32: 许有问题：and makes a decision according to the majority vote among the Y_i's such that the corresponding X_i falls in the same cube as X.

在本节中我们介绍立体直方图规则并通过检查定理6.1的条件证明其普遍一致性。该规则将 $\mathcal{R}^d$ 划分为相同大小的立方体，并使用 Y_i 根据多票原则进行决策。其中 Y_i 对应的 X_i 与 X 落在同一个立方体中。[32]正式的说，令 $\mathcal{P}_n=\{A_{n1},A_{n2},\ldots\}$ 为一个对 $\mathcal{R}^d$ 的划分，其中每个立方体大小为 $h_n>0$。换言之，该划分将 $\mathcal{R}^d$ 划分为类型 $\prod_{i=1}^d[k_ih_n,(k_i+1)h_n)$ 的集合，其中 k_i 是整数。对每一个 $x\in\mathcal{R}^d$，若 $x\in A_{ni}$，令 $A_n(x)=A_{ni}$。直方图规则定义为

$$g_n(x)=\begin{cases}0 & \text{若 } \sum_{i=1}^n I_{\{Y_i=1\}}I_{\{X_i\in A_n(x)\}}\leq\sum_{i=1}^n I_{\{Y_i=0\}}I_{\{X_i\in A_n(x)\}}\\ 1 & \text{否则。}\end{cases}$$

直方图规则的一致性由 Glick（1973）基于一些附加条件提出。随

之 Gordom 和 Olshen（1978），（1980）的结果引出普遍一致性。强普遍一致性的直接证明在第 9 章中给出。

下一定理建立了某种立体直方图规则的普遍一致性。

定理 6.2 *若当 $n \to \infty$ 时有 $h_n \to 0$ 和 $nh_n^d \to \infty$，则立体直方图规则是普遍一致的。*

证. 我们检查定理6.1的两个简单条件。显然，每个元胞的直径为 $\sqrt{d}h^d$。因此条件（1）明显成立（follows trivially）。为证明条件（2），我们需证明对任 $M < \infty$，$\mathbf{P}\{N(X) \le M\} \to 0$。令 S 为圆心位于原点的任意球。则与 S 相交的元胞的数量不超过 $c_1 + c_2/h^d$，对某正常数 c_1，c_2。则

$$
\begin{aligned}
&\mathbf{P}\{N(X) \le M\} \\
&\le \sum_{j:A_{nj}\cap S\neq\emptyset} \mathbf{P}\left\{X \in A_{nj}, N(X) \le M\right\} + \mathbf{P}\left\{X \in S^c\right\} \\
&\le \sum_{\substack{j:A_{nj}\cap S\neq\emptyset \\ \mu\left(A_{nj}\right)\le 2M/n}} \mu\left(A_n j\right) + \sum_{\substack{j:A_{nj}\cap >\neq\emptyset \\ \mu\left(A_{nj}\right)>2M/n}} \mu\left(A_{nj}\right) \mathbf{P}\left\{n\mu_n\left(A_{nj}\right) \le M\right\} + \mu\left(S^c\right) \\
&\le \frac{2M}{n}\left(c_1 + \frac{c_2}{h^d}\right) \\
&\quad + \sum_{\substack{j:A_{nj}\cap\neq\emptyset \\ \mu}} \mu\left(A_{nj}\right) \mathbf{P}\left\{\mu_n\left(A_{nj}\right) - \mathbf{E}\left\{\mu_n\left(A_{nj}\right)\right\} \le M/n - \mu\left(A_{nj}\right)\right\} + \mu\left(S^c\right) \\
&\le \frac{2M}{n}\left(c_1 + \frac{c_2}{h^d}\right) \\
&\quad + \sum_{\substack{j:A_{nj}\cap\neq\emptyset \\ \left(A_{nj}\right)>2M/n}} \mu\left(A_{nj}\right) \mathbf{P}\left\{\mu_n\left(A_{nj}\right) - \mathbf{E}\left\{\mu_n\left(A_{nj}\right)\right\} \le \frac{-\mu\left(A_{nj}\right)}{2}\right\} + \mu\left(S^c\right) \\
&\le \frac{2M}{n}\left(c_1 + \frac{c_2}{h^d}\right) + \sum_{\substack{j:A_{nj}\cap S\neq\emptyset \\ \mu\left(A_{nj}\right)>2M/n}} 4\mu\left(A_{nj}\right) \frac{\operatorname{Var}\left\{\mu_n\left(A_{nj}\right)\right\}}{\left(\mu\left(A_{nj}\right)\right)^2} + \mu\left(S^c\right) \\
&\text{(通过 Chebyshev 不等式)} \\
&\le \frac{2M}{n}\left(c_1 + \frac{c_2}{h^d}\right) + \sum_{\substack{j:A_{nj}\cap S\neq\emptyset \\ \mu\left(A_{nj}\right)>2M/n}} 4\mu\left(A_{nj}\right) \frac{1}{n\mu\left(A_{nj}\right)} + \mu\left(S^c\right) \\
&\le \frac{2M+4}{n}\left(c_1 + \frac{c_2}{h^d}\right) + \mu\left(S^c\right) \\
&\to \mu\left(S^c\right) ,
\end{aligned}
$$

因为 $nh^d \to \infty$。由于 S 是任意的，即完成定理的证明。 [证毕]

6.5 Stone 定理

Stone（1977）给出的一般定理（general theorem）允许我们推导出几个分类规则的普遍一致性。考虑一个基于后验概率 η 估计的规则，其中 η 具有形式

$$
\eta_n(x) = \sum_{i=1}^{n} I_{\{Y_i=1\}} W_{ni}(x) = \sum_{i=1}^{n} Y_i W_{ni}(x) ,
$$

记 $W_{ni}(x) = W_{ni}(x, X_1, \ldots, X_n)$ 为非负权重且所有权重之和为 1：

$$\sum_{i=1}^{n} W_{ni}(x) = 1 \text{。}$$

定义分类规则为

$$\begin{aligned} g_n(x) &= \begin{cases} 0 & \text{若 } \sum_{i=1}^n I_{\{Y_i=1\}} W_{ni}(x) \le \sum_{i=1}^n I_{\{Y_i=0\}} W_{ni}(x) \\ 1 & \text{否则,} \end{cases} \\ &= \begin{cases} 0 & \text{若 } \sum_{i=1}^n Y_i W_{ni}(x) \le 1/2 \\ 1 & \text{否则。} \end{cases} \end{aligned}$$

η_n 是 η 的加权平均估计量。直觉上，一个 X_i 足够接近 x 的点对 (X_i, Y_i) 应该比那些离 x 远的点对提供更多的关于 $\eta(x)$ 的信息。因此，X 邻居的权重通常都相对更大，故 η_n 粗略地说是一个在 X 的所有邻近点中具标签 1 的 X_i 的（加权）相对频率。因此，η_n 可被视为局部平均估计量，且 g_n 为局部（加权）多数票（规则）[33]。这类规则的具体例子包括直方图，核与最近邻规则。随后将进一步研究这些规则。

33: 许有问题：a local (weighted) majority vote.

定理 6.3 *（Stone（1977））假设对任意 X 的分布，权重满足如下三个条件：*

(i) 存在一个常量 c 使得对任意非负可测函数 f 满足 $\mathbf{E}f(X) < \infty$，和

$$\mathbf{E}\left\{\sum_{i=1}^{n} W_{ni}(X) f(X_i)\right\} \le c\mathbf{E}f(X) \text{。}$$

(ii) 对所有 $a > 0$，

$$\lim_{n\to\infty} \mathbf{E}\left\{\sum_{i=1}^{n} W_{ni}(X) I_{\{\|X_i - X\| > a\}}\right\} = 0 \text{。}$$

(iii)

$$\lim_{n\to\infty} \mathbf{E}\left\{\max_{1\le i\le n} W_{ni}(X)\right\} = 0 \text{。}$$

则 g_n 为普遍一致的。

备注 6.2 条件（*ii*）要求所有 X_i 的权重必须趋于 0，其中 X_i 位于一个以 X 为圆心，半径固定的任意圆球之外。换言之，在取平均时应仅考虑 X 缩小邻域（shrinking neighborhood）中的点。条件（*iii*）要求不应存在一个对估计具有过大影响的点 X_i。因此，平均化的点的数量必须趋向于无穷大。[34]条件（*i*）是技术上的。

34: 应有问题：Hence, the number of points encountered in the averaging must tend to infinity.

证. 通过推论6.2足以证明对任意 (X, Y) 分布，

$$\lim_{n\to\infty} \mathbf{E}\left\{(\eta(X) - \eta_n(X))^2\right\} = 0 \text{。}$$

引入概念

$$\widehat{\eta}_n(x) = \sum_{i=1}^{n} \eta(X_i) W_{ni}(x) \text{。}$$

则通过一个简单不等式 $(a+b)^2 \le 2\left(a^2+b^2\right)$，有

$$\begin{aligned}
&\mathbf{E}\left\{(\eta(X)-\eta_n(X))^2\right\} \\
&\quad = \mathbf{E}\left\{((\eta(X)-\widehat{\eta}_n(X))+(\widehat{\eta}_n(X)-\eta_n(X)))^2\right\} \\
&\quad \le 2\left(\mathbf{E}\left\{(\eta(X)-\widehat{\eta}_n(X))^2\right\}+\mathbf{E}\left\{(\widehat{\eta}_n(X)-\eta_n(X))^2\right\}\right) 。
\end{aligned} \tag{6.1}$$

因此，现在可以证明不等号右边两项均趋于 0。因为 W_{ni} 均为非负且和为 1，由 Jensen 不等式，第一项为

$$\begin{aligned}
\mathbf{E}\left\{(\eta(X)-\widehat{\eta}_n(X))^2\right\} &= \mathbf{E}\left\{\left(\sum_{i=1}^n W_{ni}(X)\left(\eta(X)-\eta\left(X_i\right)\right)\right)^2\right\} \\
&\le \mathbf{E}\left\{\sum_{i=1}^n W_{ni}(X)\left(\eta(X)-\eta\left(X_i\right)\right)^2\right\} 。
\end{aligned}$$

若函数 $0 \le \eta^* \le 1$ 是连续的且其支撑是有界的，则 η^* 也是一致连续的：对任意 $\epsilon > 0$，存在一个 $a > 0$ 使得 $\|x_1 - x\| < a$，$\left|\eta^*\left(x_1\right)-\eta^*(x)\right|^2 < \epsilon$。其中 $\|x\|$ 表示向量 $x \in \mathscr{R}^d$ 的 Euclidean 范数（欧几里得范数）。因此，由于 $\left|\eta^*\left(x_1\right)-\eta^*(x)\right| \le 1$，根据条件（$ii$）有

$$\begin{aligned}
&\mathbf{E}\left\{\sum_{i=1}^n W_{ni}(X)\left(\eta^*(X)-\eta^*\left(X_i\right)\right)^2\right\} \\
&\le \mathbf{E}\left\{\sum_{i=1}^n W_{ni}(X) I_{\{\|X-X_i\| \ge a\}}\right\}+\mathbf{E}\left\{\sum_{i=1}^n W_{ni}(X)\epsilon\right\} \to \epsilon 。
\end{aligned}$$

由于具有界支撑的连续函数集在 $L_2(\mu)$ 中是稠密的，对任 $\epsilon > 0$ 我们可选择 η^* 使得

$$\mathbf{E}\left\{(\eta(X)-\eta^*(X))^2\right\} < \epsilon 。$$

通过该选择，使用不等式 $(a+b+c)^2 \le 3\left(a^2+b^2+c^2\right)$（来自 Cauchy-Schwarz 不等式），

$$\begin{aligned}
&\mathbf{E}\left\{(\eta(X)-\widehat{\eta}_n(X))^2\right\} \\
&\le \quad \mathbf{E}\left\{\sum_{i=1}^n W_{ni}(X)\left(\eta(X)-\eta\left(X_i\right)\right)^2\right\} \\
&\le \quad 3\mathbf{E}\left\{\sum_{i=1}^n W_{ni}(X)\left((\eta(X)-\eta^*(X))^2+(\eta^*(X)-\eta^*\left(X_i\right))^2\right.\right. \\
&\qquad \left.\left.+\left(\eta^*\left(X_i\right)-\eta\left(X_i\right)\right)^2\right)\right\} \\
&\le \quad 3\mathbf{E}\left\{(\eta(X)-\eta^*(X))^2\right\} \\
&\qquad +3\mathbf{E}\left\{\sum_{i=1}^n W_{ni}(X)\left(\eta^*(X)-\eta^*\left(X_i\right)\right)^2\right\}+3c\mathbf{E}\left\{(\eta(X)-\eta^*(X))^2\right\} ,
\end{aligned}$$

这里我们使用条件（i）。因此，有

$$\limsup_{n\to\infty} \mathbf{E}\left\{(\eta(X)-\widehat{\eta}_n(X))^2\right\} \le 3\epsilon(1+1+c) 。$$

为处理式（6.1）右边第二项，由独立性观察到有

$$\mathbf{E}\left\{(Y_i-\eta(X_i))\left(Y_j-\eta\left(X_j\right)\right)\mid X,X_1,\ldots,X_n\right\}=0 \text{ 对任} i\neq j \text{ 。}$$

因此由条件（*iii*），有

$$\begin{aligned}
&\mathbf{E}\left\{\widehat{\eta}_n(X)-\eta_n(X))^2\right\}\\
&=\mathbf{E}\left\{\left(\sum_{i=1}^n W_{ni}(X)\left(\eta\left(X_i\right)-Y_i\right)\right)^2\right\}\\
&=\sum_{i=1}^n\sum_{j=1}^n\mathbf{E}\left\{W_{ni}(X)\left(\eta\left(X_i\right)-Y_i\right)W_{nj}(X)\left(\eta\left(X_j\right)-Y_j\right)\right\}\\
&=\sum_{i=1}^n\mathbf{E}\left\{W_{ni}^2(X)\left(\eta\left(X_i\right)-Y_i\right)^2\right\}\\
&\leq\mathbf{E}\left\{\sum_{i=1}^n W_{ni}^2(X)\right\}\leq\mathbf{E}\left\{\max_{1\leq i\leq n}W_{ni}(X)\sum_{j=1}^n W_{nj}(X)\right\}\\
&=\mathbf{E}\left\{\max_{1\leq i\leq n}W_{ni}(X)\right\}\to 0\text{ 。}
\end{aligned}$$

定理得证。 [证毕]

6.6 k-最近邻规则

在第 5 章中我们讨论了当样本大小 n 增加时 k 保持固定时，k-最近邻规则的渐进性。在这种情况下，期望误差概率在 L^* 和 $2L^*$ 构成的区间内收敛。在本节中，我们证明若允许 k 随着 n 的增加而变化，使得 $k/n\to 0$，则该规则是弱普遍一致的。该定理的证明是 Stone 定理的简单应用。该结果首次发表在 Stone 的论文（1977）中，是第一个对任意规则的普遍一致性结果。强一致性，和许多其他不同视角的 $k-\mathrm{NN}$ 规则将在第 11 章和 26 章进行介绍。

回忆 k-最近邻规则的定义：首先根据 X_i 距 x 的 Euclidean 距离按升序对数据进行排序：

$$\left(X_{(1)}(x),Y_{(1)}(x)\right),\ldots,\left(X_{(n)}(x),Y_{(n)}(x)\right)\ ,$$

即 $X_{(i)}(x)$ 为数据点 $X_1,\ldots,\ Xn$ 中 x 的第 i 个最近邻。当出现距离相等的情况，可通过比较下标进行处理，即若 $\|X_i-x\|=\|X_j-x\|$，$i<j$，则可认为 X_i 更“接近”x。

$k-\mathrm{NN}$ 分类规则可定义为

$$g_n(x)=\begin{cases}0 & \text{若}\sum_{i=1}^k I_{\{Y_{(i)}(x)=1\}}\leq\sum_{i=1}^k I_{\{Y_{(i)}(x)=0\}}\\ 1 & \text{否则。}\end{cases}$$

换言之，$g_n(x)$ 是在 x 的 k 个最近邻对应的标签集合中的多数票。

Stone（1977）

定理 6.4 *若 $k\to\infty$ 且 $k/n\to 0$，则对任意分布 $\mathbf{E}L_n\to L^*$。*

证. 我们通过检查 Stone 弱收敛定理（定理6.3）的条件进行证明。定理6.3中的权重 $W_{ni}(X)$ 等于 $1/k$ 当且仅当 X_i 是 X 的 k 个最近邻之一；否则等于 0.

条件（*iii*）明显成立，因为 $k \to \infty$。对条件（*ii*），观察到当

$$\mathbf{P}\left\{\left\|X_{(k)}(X) - X\right\| > \epsilon\right\} \to 0$$

时

$$\mathbf{E}\left\{\sum_{i=1}^{n} W_{ni}(X) I_{\{\|X_i - X\| > \epsilon\}}\right\} \to 0$$

成立。其中 $X_{(k)}(x)$ 表示 x 在 $X_1, \ldots, X_n$ 中的第 k 个最近邻。但我们从引理5.1中知道当 $k/n \to 0$ 时对任意 $\epsilon > 0$，该式为真。[35]

35: 些许疑惑

最后，我们我们考虑条件 (i)。我们需要证明对任意满足 $\mathbf{E}\{f(X)\} < \infty$ 的非负可测函数 f，存在常数 c，有

$$\mathbf{E}\left\{\sum_{i=1}^{n} \frac{1}{k} I_{\left\{X_i \text{ 是} X \text{ 的 } k \text{ 最近邻之一}\right\}} f(X_i)\right\} \leq \mathbf{E}\{cf(X)\}\text{。}$$

但我们已在引理5.3中证明当 $c = \gamma_d$ 时该不等式总会成立。因此，条件 (i) 得证。 [证毕]

6.7 分类比回归函数估计更简单

再次假设我们的决策基于一些后验概率函数 η 的估计 η_n，即

$$g_n(x) = \begin{cases} 0 & \text{if } \eta_n(x) \leq 1/2 \\ 1 & \text{otherwise}\text{。} \end{cases}$$

定理2.2，2.3和推论6.2中的界指出若 η_n 是 η 的一致估计，则导出规则（resulting rule）也是一致的。例如，记 $L_n = \mathbf{P}\{g_n(X) \neq Y \mid D_n\}$，我们有

$$\mathbf{E}L_n - L^* \leq 2\sqrt{\mathbf{E}\left\{(\eta_n(X) - \eta(X))^2\right\}}\,,$$

即，回归函数 η 的 L_2-一致估计蕴含了（lead to）一致分类。实际上，这是定理6.3证明中使用的主要工具。然而所示的界对证明一致性是有用的，但在研究收敛速率时几乎无用武之地。下述的定理6.5证明，对一致性规则，$\mathbf{P}\{g_n(X) \neq Y\}$ 到 L^* 的收敛速率总是比 $\sqrt{\mathbf{E}\left\{(\eta(X) - \eta_n(X))^2\right\}}$ 到零的收敛速率要好几个数量级。

因此模式识别比回归函数估计更容易。这将是一个反复出现的问题——在模式识别中取得可接受的结果，给定少数样本时，相比回归函数估计，模式识别可以做得更多。这是因为模式识别要求的更少这一事实。它也证实了我们的信念：模式识别与回归函数估计是大不相同的，且模式识别在统计学社区值得单独进行处理（deserves separate treatment）。

定理 6.5 *令 η_n 为弱一致回归估计，即*

$$\lim_{n\to\infty} \mathbf{E}\left\{(\eta_n(X) - \eta(X))^2\right\} = 0\text{。}$$

定义

$$g_n(x) = \begin{cases} 0 & \text{若} \eta_n(x) \le 1/2 \\ 1 & \text{否则。} \end{cases}$$

则

$$\lim_{n\to\infty} \frac{\mathbf{E}L_n - L^*}{\sqrt{\mathbf{E}\left\{(\eta_n(X) - \eta(X))^2\right\}}} = 0\,,$$

即，$\mathbf{E}L_n - L^*$ 收敛至 0 的速度比回归估计的 L_2-误差的收敛速率更快。

证. 我们从定理2.2的等式开始:

$$\mathbf{E}L_n - L^* = 2\mathbf{E}\left\{|\eta(X) - 1/2| I_{\{g_n(X)\neq g^*(X)\}}\right\} \text{。}$$

固定 $\epsilon > 0$。我们可由

$$\begin{aligned} &\mathbf{E}\left\{|\eta(X) - 1/2| I_{\{g_n(X)\neq g^*(X)\}}\right\} \\ &\quad\le \quad \mathbf{E}\left\{I_{\{\eta(X)\neq 1/2\}} |\eta(X) - \eta_n(X)| I_{\{g_n(X)\neq g^*(X)\}}\right\} \\ &\quad= \quad \mathbf{E}\left\{|\eta(X) - \eta_n(X)| I_{\{g_n(X)\neq g^*(X)\}} I_{\{|\eta(X)-1/2|\le\epsilon\}} I_{\{\eta(X)\neq 1/2\}}\right\} \\ &\qquad + \mathbf{E}\left\{|\eta(X) - \eta_n(X)| I_{\{g_n(X)\neq g^*(X)\}} I_{\{|\eta(X)-1/2|>\epsilon\}}\right\} \\ &\quad\le \quad \sqrt{\mathbf{E}\left\{(\eta_n(X) - \eta(X))^2\right\}} \\ &\qquad \times \Big(\sqrt{\mathbf{P}\{|\eta(X) - 1/2| \le \epsilon, \eta(X) \neq 1/2\}} \\ &\qquad + \sqrt{\mathbf{P}\left\{g_n(X) \neq g^*(X), |\eta(X) - 1/2| > \epsilon\right\}}\Big) \\ &\qquad \text{(由 Cauchy-Schwarz 不等式)} \end{aligned}$$

给出最后那个因子的界。由于 $g_n(X) \neq g^*(X)$ 且 $|\eta(X) - 1/2| > \epsilon$ 蕴含着 $|\eta_n(X) - \eta(X)| > \epsilon$，回归估计的一致性蕴含着对任固定 $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbf{P}\left\{g_n(X) \neq g^*(X), |\eta(X) - 1/2| > \epsilon\right\} = 0 \text{。}$$

换言之,

$$\mathbf{P}\{|\eta(X) - 1/2| \le \epsilon, \eta(X) \neq 1/2\} \to 0 \text{当} \epsilon \to 0\,,$$

得证定理。 [证毕]

比例

$$\rho_n = \frac{\mathbf{E}L_n - L^*}{\sqrt{\mathbf{E}\left\{(\eta_n(X) - \eta(X))^2\right\}}}$$

的真实值不具普遍界。实质上，ρ_n 可任意慢的趋于 0（见问题**??**）。另一方面，ρ_n 也能非常快地趋于 0。在问题**??**、**??**和下述定理中，给出的 ρ_n 上界能被用来推导收敛速率。具体地说，定理6.6证明当 $L^* = 0$ 时，$\mathbf{E}L_n - L^*$ 趋于 0 的速度与回归估计（即 $\mathbf{E}\left\{(\eta_n(X) - \eta(X))^2\right\}$）的 L_2 误差的平方收敛速率一样快。ρ_n 趋于 0 的速度依赖于两点：（1）η_n 至 η 的收敛速率；（2）当 $\epsilon \downarrow 0$ 时，概率 $\mathbf{P}\{|\eta(X) - 1/2| \le \epsilon, \eta(X) \neq 1/2\}$ 作为 ϵ 的函数的表现（即 $\eta(x)$ 在那些使得 $\eta(x) \approx 1/2$ 的 x 处的表现）。

定理 6.6 *假设 $L^* = 0$，并考虑决策*

$$g_n(x) = \begin{cases} 0 & 若\eta_n(x) \le 1/2 \\ 1 & 否则。 \end{cases}$$

则

$$\mathbf{P}\{g_n(X) \neq Y\} \le 4\mathbf{E}\left\{(\eta_n(X) - \eta(X))^2\right\} 。$$

证. 由定理2.2,

$$\begin{aligned} \mathbf{P}\{g_n(X) \neq Y\} =& 2\mathbf{E}\left\{|\eta(X) - 1/2| I_{\{g_n(X) \neq g^*(X)\}}\right\} \\ =& 2\mathbf{E}\left\{|\eta(X) - 1/2| I_{\{g_n(X) \neq Y\}}\right\} \\ & (\text{由于} g^*(X) = Y \text{ 通过假设} L^* = 0) \\ \le& 2\sqrt{\mathbf{E}\left\{(\eta_n(X) - \eta(X))^2\right\}} \sqrt{\mathbf{P}\{g_n(X) \neq Y\}} \\ & (\text{由 Cauchy-Schwarz 不等式})。 \end{aligned}$$

将两边同时除以 $\sqrt{\mathbf{P}\{g_n(X) \neq Y\}}$ 即可得证。 [证毕]

上述定理的结果证明定理2.2、2.3和推论**??**的界可以任意松的，且误差概率收敛至 L^* 的速率比回归估计的 L_2-误差收敛于 0 的速率更快。在一些情况下，$\mathbf{E}|\eta_n(X) - \eta(X)|$ 不收敛于 0，但一致性存在。例如考虑一个*严格可分*的分布，即该分布使得存在两个集合 A，$B \subset \mathcal{R}^d$，对某 $\delta > 0$ 满足

$$\inf_{x \in A, y \in B} \|x - y\| \ge \delta > 0 ,$$

且具有性质

$$\mathbf{P}\{X \in A \mid Y = 1\} = \mathbf{P}\{X \in B \mid Y = 0\} = 1 。$$

在该情况下，存在一个版本的 η，若 x 在集合 A 中，则 $\eta(x) = 1$；若 x 在集合 B 中，则 $\eta(x) = 0$。我们说“版本”是因为 η 在测度为 0 的集合上是未定义的。对这些严格可分的分布，$L^* = 0$。令 $\tilde{\eta}$ 在 B 上取 $1/2 - \epsilon$，在 A 上取 $1/2 + \epsilon$。则，根据

$$g(x) = \begin{cases} 0 & 若\tilde{\eta}(x) \le 1/2 \\ 1 & 否则, \end{cases} = \begin{cases} 0 & 若x \in B \\ 1 & 若x \in A , \end{cases}$$

我们有 $\mathbf{P}\{g(X) \neq Y\} = L^* = 0$。然而,

$$2\mathbf{E}|\eta(X) - \tilde{\eta}(X)| = 1 - 2\epsilon$$

可任意接近 1。

在更现实的例子中，我们考虑*核规则*（kernel rule）（见第 10 章),

$$g_n(x) = \begin{cases} 0 & 若\eta_n(x) \le 1/2 \\ 1 & 否则, \end{cases}$$

其中

$$\eta_n(x) = \frac{\sum_{i=1}^n Y_i K(x - X_i)}{\sum_{i=1}^n K(x - X_i)} ,$$

K 为 $\mathcal{R}^d$ 中的标准正态密度：

$$K(u) = \frac{1}{(2\pi)^{d/2}} e^{-\|u\|^2/2} 。$$

假设 A 和 B 各自包含一个点，其相互距离为 δ——即 X 的分布集中在两个点上。若 $\mathbf{P}\{Y=0\} = \mathbf{P}\{Y=1\} = 1/2$，我们可见当 $x \in A \cup B$ 时，由大数定律

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} K(x - X_i) = \frac{K(0) + K(\delta)}{2} \quad \text{按概率 1 成立。}$$

另外，

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i K(x - X_i) = \begin{cases} K(0)/2 & \text{若} x \in A \\ K(\delta)/2 & \text{若} x \in B \end{cases}$$

按概率 1 成立。因此，

$$\lim_{n\to\infty} \eta_n(x) = \begin{cases} \frac{K(0)}{K(0)+K(\delta)} & \text{若} x \in A \\ \frac{K(\delta)}{K(0)+K(\delta)} & \text{若} x \in B \end{cases}$$

按概率 1 成立。因此，当在 A 上 $\eta(x) = 1$，在 B 上 $\eta(x) = 0$ 时，

$$\begin{aligned} &\lim_{n\to\infty} 2\mathbf{E}\,|\eta(X) - \eta_n(X)| \\ &\quad = 2\frac{1}{2}\left(\left|1 - \frac{K(0)}{K(0)+K(\delta)}\right| + \left|0 - \frac{K(\delta)}{K(0)+K(\delta)}\right|\right) \\ &\quad = \frac{2K(\delta)}{K(0)+K(\delta)} 。 \end{aligned}$$

然而，$L^* = 0$ 且 $\mathbf{P}\{g_n(X) \neq Y\} \to 0$。实际上，若 D_n 表示训练数据，

$$\lim_{n\to\infty} \mathbf{P}\{g_n(X) \neq Y \mid D_n\} = L^*$$

按概率 1 成立，且

$$\lim_{n\to\infty} 2\mathbf{E}\{|\eta_n(X) - \eta(X)| \mid D_n\} = \frac{2K(\delta)}{K(0)+K(\delta)}$$

按概率 1 成立。这强力证明了对任意 $\delta > 0$，对许多现实的分类规则，我们完全不需要 η_n 至 η 的收敛性！

因为第 6 章到 11 章的所有一致性证明均依赖于 η_n 至 η 的收敛性，我们将为一些分布创建非必要条件，尽管总是可以找到需要该条件的 (X,Y) 分布——在后一种意义上，这些普遍一致性结果的条件是不可改进的。[36]

36: 许有问题：As all the consistency proofs in Chapters 6 through 11 rely on the convergence of η_n to η, we will create unnecessary conditions for some distributions, although it will always be possible to find distributions of (X, Y) for which the conditions are needed-in the latter sense, the conditions of these universal consistency results are not improvable.

6.8 聪明规则

规则是映射 $g_n : \mathcal{R}^d \times (\mathcal{R}^d \times \{0,1\})^n \to \{0,1\}$ 的序列。当 n 增加时，大多数规则期望能表现得更好。因此，我们称一个规则是聪明的，如果对任意 (X,Y) 分布，$\mathbf{E}\{L(g_n)\}$ 是非递减的，其中

$$L(g_n) = \mathbf{P}\{g_n(X, D_n) \neq Y \mid D_n\} 。$$

一些“愚蠢规则”（dumb rules）是聪明的，例如对一“无用规则”，它对任 n，忽略 X_i 直接在所有 Y_i 上取多数票。由于

$$\mathbf{P}\left\{\sum_{i=1}^{n}(2Y_i-1)>0, Y=0 \text{ 或 } \sum_{i=1}^{n}(2Y_i-1)\le 0, Y=1\right\}$$

是对 n 单调的，故该规则是聪明的。这是二项分布的一个性质（见问题**??**）。具固定划分的直方图规则是聪明的（问题**??**）。1-最近邻规则不是聪明的。要看到这一点，令 (X,Y) 在点 $(0,1)$ 和 $(Z,0)$ 处的概率分别为 p 和 $1-p$，其中 Z 是在 $[-1000,1000]$ 上的均匀变量。验证对 $n=1, \mathbf{E}L_n = 2p(1-p)$，同时对 $n=2$，

$$\begin{aligned}\mathbf{E}L_n &= 2p(1-p)^2\left(\frac{1}{2}+\frac{\mathbf{E}|Z|}{4000}\right)+p^2(1-p)+(1-p)^2p\\ &= 2p(1-p)\left(\frac{5(1-p)}{8}+\frac{1}{2}\right),\end{aligned}$$

当 $p\in(0,1/5)$ 时，$\mathbf{E}L_n$ 大于 $2p(1-p)$。这证明了在所有这些情况下，$n=1$ 比 $n=2$ 更好。类似地，对固定 h 的标准核规则——在第 10 章讨论——不是聪明的（见问题**??**，**??**）。

上述举例的聪明规则的误差概率并未随 n 显著变化。但是，变化是保证贝叶斯风险一致性的必要条件。在变化的地方——例如在直方图规则中当 h_n 跳转至一个新值时——单调性可能会丧失。这导致一个猜想：*不存在*聪明的普遍一致性规则。

些许疑惑

6.9 问题与练习

问题 6.1

记 $\mathcal{R}^d$ 中的独立同分布变量 $X_1,\ldots,X_n$ 的密度为 f。通过 f_n 估计 f，其中 f_n 为 x 和 $X_1,\ldots,X_n$ 的函数。并假设 $\int|f_n(x)-f(x)|\,dx\to 0$ 按概率成立（或按概率 1 成立）。则当条件密度 f_0 和 f_1 存在时，证明存在一个一致（或强一致）分类规则。

问题 6.2

令 $X_1,\ldots,X_n$ 为 $\mathcal{R}^d$ 中的独立同分布随机变量，其密度为 f。令 $\mathcal{P}_n$ 是切割 $\mathcal{R}^d$ 成多个大小为 h_n 的立方体的一个划分，并定义直方图密度估计为

直方图密度估计

$$f_n(x)=\frac{1}{nh_n^d}\sum_{i=1}^{n}I_{\{X_i\in A_n(x)\}},$$

其中 $A_n(x)$ 为 $\mathcal{P}_n$ 中包含 x 的集合。证明若当 $n\to\infty$ 时有 $h_n\to 0$ 和 $nh_n^d\to\infty$，则该估计在 L_1 中是*普遍一致的*。即对任 f，估计的 L_1 误差 $\int|f_n(x)-f(x)|\,dx$ 按概率收敛于 0，或等价地，

$$\lim_{n\to\infty}\mathbf{E}\left\{\int|f_n(x)-f(x)|\,dx\right\}=0\,。$$

[提示：下述建议可能有用

(1) $\mathbf{E}\left\{\int |f_n - f|\right\} \leq \mathbf{E}\left\{\int |f_n - \mathbf{E}f_n|\right\} + \int |\mathbf{E}f_n - f|$。
(2) $\mathbf{E}\left\{\int |f_n - \mathbf{E}f_n|\right\} = \sum_j \left|\mu\left(A_{nj}\right) - \mu_n\left(A_{nj}\right)\right|$。
(3) 首先证明 $\int |\mathbf{E}f_n - f| \to 0$ 对一致连续函数 f 成立，接着将其扩展至任意密度函数。

]

问题 6.3

令 X 在 $[0,1]$ 上按概率 1/2 服从均匀分布，并令 X 在有理数上按概率 1/2 是原子的（例如，若所有有理数排列为 $r_1, r_2, r_3, \ldots$，则 $\mathbf{P}\{X = r_i\} = 1/2^{i+1}$）。若 X 为有理数则令 $Y = 1$，否则令 $Y = 0$。给出 1-最近邻规则一致性的直接证明。[提示：给定 $Y = 1$，X 的条件分布是离散的。因此，对任 $\epsilon > 0$，存在一个整数 k 使得当给定 $Y = 1$ 时，X 等于概率至少为 $1 - \epsilon$ 的 k 个有理数的其中之一。现在，若 n 足够大，该集合中的每个点可以大的概率捕获标签 1 的数据点。另外，对大的 n，这些点之间的空间被具标签 0 的数据点填满。]

问题 6.4

通过检查 Stone 定理的条件证明立体直方图规则的一致性。[提示：检查条件（i），首先给出 $W_{ni}(x)$ 的界

$$I_{\{X_i \in A_n(x)\}} / \sum_{j=1}^{n} I_{\{X_j \in A_n(x)\}} + 1/n \text{。}$$

因为

$$\mathbf{E}\left\{\sum_{i=1}^{n} \frac{1}{n} f\left(X_i\right)\right\} = \mathbf{E}f(X)\,,$$

故足可证明存在常数 $c' > 0$ 使得对任意具 $\mathbf{E}f(x) < \infty$ 的非负函数 f，有

$$\mathbf{E}\left\{\sum_{i=1}^{n} f\left(X_i\right) \frac{I_{\{X_i \in A_n(X)\}}}{\sum_{j=1}^{n} I_{\{X_j \in A_n(X)\}}}\right\} \leq c'\mathbf{E}f(X) \text{。}$$

要完成该证明，你可能需要使用引理 $A.2$（i）。为证明条件（iii）成立，重写

$$\mathbf{E}\left\{\max_{1 \leq i \leq n} W_{ni}(X)\right\}$$
$$\leq \frac{1}{n} + \mathbf{P}\left\{X \in S^c\right\} + \sum_{j: A_{nj} \cap S \neq \emptyset} \mathbf{E}\left\{I_{\{X \in A_{nj}\}} \frac{1}{n\mu_n\left(A_{nj}\right)} I_{\{\mu_n(A_{nj}) > 0\}}\right\}$$

] 并使用引理 $A.2$（ii）。