概率机器学习

Sayan Mukherjee

概率机器学习

Sayan Mukherjee

[1]统计科学、计算机科学和数学系，杜克大学，
杜罕，27708。
电子邮件地址　　：sayan@stat.duke.edu.

2015年11月19日

讲座 1
课程准备

机器学习这个术语可以追溯到Arthur Samuels和他的计算机下棋算法。 1959年，Samuels将机器学习描述为：“一种使计算机具有学习能力而无需明确编程的研究领域。”

机器学习被认为是人工智能的一个子领域，学习机器的概念出现在AlanTuring于1950年在《心智：心理学和哲学的季刊》中发表的《计算机与智能》一文中。本文的第一句提出的问题是“机器能思考吗？”。

在这门机器学习课程中，我们将考虑使用算法和概率方法来“从数据中学习”。这门课程涉及统计学、计算机科学的交叉领域，机器学习的一个夸张描述是计算机科学家在做统计学。
机器学习通常也与“大数据”这个术语联系在一起，它通常指的是对非常大的数据集进行统计分析，在这里，计算挑战与推断问题一样严重。

广义上说，我们将讨论的方法可以分为两类：程序员：这将涵盖频率统计学和机器学习的算法方法。这种方法基于找到适用于数据的良好程序。 良好的意思是某个长期概率的过程，例如分类中出错的长期概率很小。

贝叶斯：一种基于数据推断参数或模型后验概率的一致公理方法。 在某些情况下，贝叶斯推断可能不可行或不实际。

1.1. 复习

我们将从统计学的基础知识开始复习。 我们将使用贝叶斯和频率主义分析来研究一个统计问题。 以下形式将在两个模型中进行量化。

$$\begin{aligned} \mathrm{P}(M \mid D) &= \frac{\mathrm{P}(D \mid M)\mathrm{P}(M)}{\mathrm{P}(D)} \\ &\propto \mathrm{P}(D \mid M)\mathrm{P}(M), \end{aligned}$$

其中$\mathrm{P}(M \mid D)$是给定数据 D的模型 M的证据，$\mathrm{P}(D \mid M)$是给定模型 M的数据D的证据，$\mathrm{P}(M)$是模型 M的概率，$\mathrm{P}(D)$是数据的概率。这些对象的标准统计术语为

$\mathrm{P}(D \mid M) \equiv \mathrm{Lik}(D; M)$，给定模型 M的数据的似然，$\mathrm{P}(M \mid D) \equiv \mathrm{Po}$
$\mathrm{st}(D; M)$，给定数据的模型 M的后验证据，$\mathrm{P}(M) \equiv \pi(M)$，模型$M$的
先验概率（在观察数据之前）。

示例**1**：模式估计
我们考虑一个随机变量 X，它是从一个包含 k= 4个字母的字母表中抽取的$\{A, C, T, G\}$，其中我们表示 $A \equiv 1,\ C \equiv 2,\ T \equiv 3, \text{and } G \equiv 4$. 我们将概率分布设置为以下多项式分布，注意我们

正在模拟抽取四个字母

$$\begin{aligned} \mathrm{P}(n_1, n_2, n_3, n_4 \mid p_1, p_2, p_3, p_4) &\equiv \mathrm{Multi}(p_1, p_2, p_3, p_4) \\ &\propto \prod_{j=1}^{4} p_j^{n_j}, \quad \sum_{j=1}^{4} p_j = 1, p_j \geq 0 \ \forall j = 1, ..., 4, \end{aligned}$$

其中 p_i是观察到第i个字母（$\{A, C, T, G\}$在字母表中
并且 n_i表示第i个字母观察到的次数（要么是1，要么是0）。上述内容
是多项式分布的一个例子。

随机变量 X 是一个序列中的字符串，我们可以将随机字符串 $Z=($ $X_1, ..., X_m$) 视为长度为 m的字符串，每个X_i从分布中独立同分布地抽取。 这是一个字符串的例子，让我们称这些字符串为模式。
数据由一系列字符串组成， $D = \{Z_1, ..., Z\ n\}$每个字符串 Z_i独立同分布地抽取
。

我们首先陈述观察到的数据 D的可能性

$$\begin{aligned} \mathrm{P}(D \mid M) &= \mathrm{Lik}(D \mid p_1, ..., p_4) \\ \mathrm{Lik}(D \mid p_1, ..., p_4) &\propto \prod_{i=1}^{n} \left[\prod_{\ell=1}^{m} \left(\prod_{j=1}^{k} p_j^{n_{i\ell j}} \right) \right] \\ &\propto \prod_{\ell=1}^{m} \left[\prod_{i=1}^{n} \left(\prod_{j=1}^{k} p_j^{n_{i\ell j}} \right) \right] \\ &\propto \prod_{\ell=1}^{m} \left[\prod_{j=1}^{k} p_j^{\tilde{n}_{\ell j}} \right], \end{aligned}$$

其中 $n_{i\ell j}$是观察 i中在位置 ℓ上观察到字母 j的次数（这个次数为0或1），而$\tilde{n}_{\ell j} = \sum_i n_{i\ell j}$是在序列中观察到字母 j在位置 ℓ的次数。

估计 $p_1, .., p_k$的经典方法是最大似然公式-

$$\begin{aligned} \{\hat{p}_1, ..., \hat{p}_k\} &= \arg \max_{p_1,...,p_k} \left[\mathrm{Lik}(D \mid p_1, ..., p_k) \right], \\ &\text{受限于} \quad \sum_{j=1}^{k} p_j = 1, p_j \geq 0 \ \forall j = 1, ..., k. \end{aligned}$$

要理解如何进行上述优化，请了解拉格朗日乘数法的方法。 这是一个非常合理的方法，但它有一个问题，如何估计 $\{\hat{p}_1, ...,\hat{p}_k\}$的估计不确定性呢?

我们可以使用贝叶斯规则正式地对不确定性进行建模。

$$\mathrm{P}(M \mid D) \propto \mathrm{P}(D \mid M)\mathrm{P}(M),$$

如果我们可以在模型空间上放置一个概率分布，即$(p_1, ..., p_k)$。
所有点 $\mathbf{p}= (p_1, ..., p_k)$的空间，使得 $\sum_j p_k = 1$ and $p_k \geq 0$ for all $j = 1, ..., k$被称为单纯形。 我们现在介绍一个经典的分布在

单纯形上称为狄利克雷分布

$$\begin{aligned} f(p_1, ..., p_k \mid \alpha_1, ..., \alpha_k) &\equiv \mathrm{Dir}(\alpha_1, ..., \alpha_k) \\ &\propto \prod_{j=1}^{k} p_j^{\alpha_j - 1}, \quad \alpha_j \geq 0 \, \forall j, \, \alpha_j \in \mathbb{N}, \end{aligned}$$

其中 $\mathbb{N}$是自然数，我们可以将 $\{\alpha_1, ..., \alpha_k\}$参数看作是计数。 我们可以使用狄利克雷分布作为先验 $\pi(M)$，其中均匀先验为$\mathrm{Dir}(\alpha_1 = 1, ..., \alpha_k = 1)$.我们现在陈述后验概率$\mathrm{P}(M \mid D)$

$$\begin{aligned} &\propto \mathrm{Lik}(D \mid p_1, ..., p_4) \times \pi(p_1, ..., p_4) \\ &\propto \prod_{i=1}^{n} \left[\prod_{\ell=1}^{m} \left(\prod_{j=1}^{k} p_j^{n_{i\ell j}} \right) \right] \times \prod_{j=1}^{k} p_j^{\alpha_j - 1} \\ &\propto \prod_{\ell=1}^{m} \left[\prod_{i=1}^{n} \left(\prod_{j=1}^{k} p_j^{n_{i\ell j}} \right) \right] \times \prod_{j=1}^{k} p_j^{\alpha_j - 1} \\ &\propto \prod_{\ell=1}^{m} \left[\prod_{j=1}^{k} p_j^{\tilde{n}_{\ell j}} \right] \times \prod_{j=1}^{k} p_j^{\alpha_j - 1} \\ &\propto \prod_{\ell=1}^{m} \left[\prod_{j=1}^{k} p_j^{\tilde{n}_{\ell j}} \right] \times \prod_{j=1}^{k} p_j^{\alpha_j - 1} \\ &\propto \left[\prod_{j=1}^{k} p_j^{\breve{n}_j} \right] \times \prod_{j=1}^{k} p_j^{\alpha_j - 1} \\ &\propto \left[\prod_{j=1}^{k} p_j^{\breve{n}_j + \alpha_j - 1} \right] \\ &= \mathrm{Dir}(\breve{n}_1 + \alpha_1, ..., \breve{n}_k + \alpha_k), \end{aligned}$$

where $\breve{n}_j = \sum$这种估计过程的优势在于，我们不仅得到了一个点估计 $\{\hat{p}_1, ...,\hat{p}_k\}$，就像MLE方法中那样，而且我们得到了一个后验分布。 我们可以使用最高概率值作为$(p_1, ..., p_k)$的估计值，或者使用后验分布的均值。 这个例子之所以如此容易，是因为多项式和狄利克雷分布是共轭的。这意味着

$$\mathrm{Multi}(p_1, ..., p_k) \times \mathrm{Dir}(\alpha_1, ..., \alpha_k) = \mathrm{Dir}(p_1 + \alpha_1, ..., p_k + \alpha_k).$$

讲座 2
线性回归程序员的方法

2.1. 标准多元线性回归

回归问题通常被陈述为

$$Y = f(X) + \varepsilon, \quad \varepsilon \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$$

多元随机变量 $X \subseteq \mathrm{IR}^p$被称为协变量，而单变量随机变量 $Y \subseteq \mathrm{IR}$ 被称为响应变量，函数属于一个类of functions $f \in \mathcal{F}$. 随机变量X, Y与之相关的分布为联合分布: ρ

$$_{X,Y}(x, y), \quad \text{边缘分布: } \rho_{X}(x), \rho_{Y}(y), \quad \text{条件: } \rho(y \mid x).$$

目前，函数 $\mathcal{F}$的类将是线性函数

$$f(x) = \beta^T x, \quad \beta \in \mathrm{IR}^p.$$

我们所给出的数据包括 n个观测 $D = \{(x_i, y_i)\}^{ni}_{=1} \overset{iid}{\sim} \rho(x, y)$,我们称之为样本，样本大小为 n. 我们将假设我们观察到的数据与以下模型一致

$$Y_i = \beta^T X_i + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2).$$

我们的目标是给定数据 D解决以下问题:

(1) 参数推断: 对于 β，什么是一个合理的估计值，我们将其称为估计值 $\hat{\beta}$

(2) 预测: 给定一个新的 x_*，对应的 y_*是什么，尝试 $y_* = \hat{\beta}^T x_*$。

(3) 估计条件分布: 当 $X = x$时， Y |X=x是什么。一个回到高斯时代的估计 $\hat{\beta}$ 的想法是最小化最小二乘误差。

$$\arg\min_{\beta \in \mathrm{IR}^p} \left[\frac{1}{n} \sum_i (y_i - \beta^T x_i)^2, \right].$$

可以从以下概率模型推导出上述估计量

$$\mathrm{Lik}(D; \beta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2} \right),$$

通过对 β最大化上述似然函数

$$\arg\max_{\beta} \mathrm{Lik}(D;\beta) \equiv \arg\min_{\beta\in\mathrm{IR}^p} \left[\frac{1}{n}\sum_i (y_i - \beta^T x_i)^2,\right].$$

我们可以将对数似然的负数表示为

$$L = \sum_{i=1}^{n} (y_i - x_i^T\beta)^2.$$

我们将以上内容改写为矩阵表示。 在这样做的过程中，我们定义了一个矩阵 $\mathbf{X}$它是 $n \times p$的，并且矩阵的每一行都是一个数据点 x_i。 我们还定义了一个列向量 Y （$p \times 1$），其中 y_i是 y的第 i个元素。 类似地， β是一个具有 p行的列向量。 我们可以将误差最小化重写为$\arg\min$

$$_{\beta} \left[L = (Y - \mathbf{X}\beta)^T (Y - \mathbf{X}\beta)\right],$$

对 θ求导并将其设置为零（对 β求导意味着对 β中的每个元素求导）dL

$$\begin{aligned} \overline{d\beta} &= -2\mathbf{X}^T(Y - \mathbf{X}\beta) = 0 \\ &= \mathbf{X}^T(Y - \mathbf{X}\beta) = 0. \end{aligned}$$

这意味着

$$\begin{aligned} \mathbf{X}^T Y &= \mathbf{X}^T\mathbf{X}\beta \\ \hat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y. \end{aligned}$$

如果我们仔细观察上面的公式，会发现存在严重的数值问题 $-(\mathbf{X}^T\mathbf{X})^{-1}$. 矩阵 $\mathbf{X}^T\mathbf{X}$是一个 $p \times p$的秩为 n的矩阵，其中 $p \gg n$. 这意味着 $\mathbf{X}^T\mathbf{X}$不能被求逆，因此我们无法通过矩阵求逆来计算估计值$\hat{\beta}\cdot$ 有一些数值方法可以解决这个问题，但解决方案可能不是唯一的或稳定的. 在估计问题中，一个普遍的规则是数值问题通常与估计参数的统计误差或估计的方差相一致.

2.2. Stein估计器

上述数值问题与一个令人惊奇的结果有关，这个结果首次由Charles Stein在1956年观察到。Stein提出的问题是，如果给定一个多元正态分布的观测值

$$(x_i)_{i=1}^n \overset{iid}{\sim} \mathrm{N}(\theta,\, \sigma^2\mathbf{I}),$$

什么是 θ的最佳估计器。 在统计学中，如果存在一个比你正在使用的估计器更好的估计器，那么你的估计器被称为不可接受的。 Stein发现，对于 $p \geq 3$，样本均值

$$\hat{\theta} = \frac{1}{n}\sum_i x_i,$$

不可接受。 一个更好的估计器，称为James-Stein估计器，如下所示

$$\hat{\theta} = \left(1 - \frac{(p-2)\frac{\sigma^2}{n}}{\|\bar{x}\|^2}\right)\bar{x}.$$

关于上述估计器的直觉是将样本均值 $\bar{x}$向零点移动一点，这被称为收缩或收缩向零点。 现在什么是最优估计器，它是最小化的估计器

$$\arg\min_{\beta\in\mathrm{IR}^p}\left[\mathrm{IE}_{X,Y}\left[(y-\beta^T x)^2\right] = \int_{Y,X}(y-\beta^T x)^2\rho(x,\,y)\;\mathrm{d}x\;\mathrm{d}y\right],$$

上面的想法是在未见数据上最小化误差，在课程后面我们会称之为泛化误差$I[\hat{f}] = \mathrm{IE}_{X,Y}$

$$\left[(y-\hat{f}(x))^2\right], \quad \hat{f}(x) = \hat{\beta}^T x,$$

并提供一些理论来证明我们将提出的估计器。

2.3. 交叉验证

由于没有访问生成分布 $\rho(x,y)$，因此无法估计泛化误差。 泛化误差的常见代理是留一交叉验证误差

$$I[cv] \equiv \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_{D^{\setminus i}}(x_i))^2,$$

其中 $D^{\setminus i}$是删除第 i个样本后的数据集， $\hat{f}_{D^{\setminus i}}$是在删除第 i个样本时估计的函数。 这个想法是删除第 i个样本，用剩下的样本估计一个函数，然后测试在第 i个样本上产生的误差，并将其平均化到所有的 n 个观测值上。 当然，不一定要留出一个观测值，可以留出 k个观测值。 留一估计器是（几乎）无偏的。

2.4. 收缩模型

我们现在可以将James-Stein估计器的思想应用到线性回归问题上。 我们将最小化以下损失函数

$$\arg\min_{\beta\in\mathrm{IR}^p}\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|^2,$$

其中 $\lambda > 0$是一个参数， $\|\beta\|^2 = \sum_{i=1}^{p}\beta_i^2$.

$$\begin{aligned}\frac{dL}{d\beta} &= -2\mathbf{X}^T(Y - \mathbf{X}\beta) + 2\lambda n\beta = 0\\ &= \mathbf{X}^T(\mathbf{X}\beta - Y) + \lambda n\beta.\end{aligned}$$

这意味着

$$\begin{aligned}\mathbf{X}^T Y &= \mathbf{X}^T\mathbf{X}\beta + \lambda n\beta\\ \mathbf{X}^T Y &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})n\beta\\ \hat{\beta} &= (\mathbf{X}^T\mathbf{X} + \lambda n\mathbf{I})^{-1}\mathbf{X}^T Y,\end{aligned}$$

其中 $\mathbf{I}$是 $p \times p$的单位矩阵。 矩阵 $(\mathbf{X}^T\mathbf{X} + \lambda n\mathbf{I})$ 是可逆的，这种惩罚损失函数的方法在变量比观测值多的问题中取得了巨大的成功。

讲座 3

贝叶斯动机的程序主义方法

通常有两种方法来验证统计估计过程
(1) 证明该过程是一致的。
(2) 证明存在一个贝叶斯过程产生相同的结果。

3.1. 一致性

在回归的背景下，一致估计量的意思是：

定义(一致性). 在回归中，从函数类 $\mathcal{F}$中选择的估计量 $\hat{f}$如果对于任意$\varepsilon > 0$

$$\lim_{n\to\infty} \sup_{\rho} \mathrm{IP}_D \left\{ I[\hat{f}] > \inf_{f\in\mathcal{F}} I[f] + \varepsilon \right\} = 0,$$

其中 ρ是数据的联合分布， $D\equiv \{(X_i, Y_i)\}^{ni}_{=1} \overset{iid}{\sim} \rho(x, y)$是从联合分布中独立同分布采样得到的数据。

在课程的后面（几节课之后），我们将讨论为什么准则（2）是有效的。我们现在将为上一讲中开发的过程提供一个贝叶斯解释。

3.2. 似然函数

任何贝叶斯公式的第一步是为数据陈述似然模型。 前面句子的更正式陈述属于所谓的似然原则，可以概括为样本中与模型参数相关的所有证据都包含在似然函数中。

到目前为止，我们的模型是

$$Y_i = f(X_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2),$$

with $f \in \mathcal{F}$ and $\mathcal{F} = \left\{ f \mid f(x) = \beta^T x \right\}$.这个模型中有两组参数：向量 β和误差的方差 σ^2或 $\theta = \{\beta, \sigma^2\}$。似然函数可以陈述为

$$\mathrm{Lik}(D; \theta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2} \right).$$

关于上述似然的一个重要思想是充分统计量的概念$T(x)$。 这意味着一旦计算出充分统计量，数据就可以被丢弃而不会丢失信息。 以已知方差 σ^2的单变量正态分布为例，给出了似然函数 $t(x_1, ...x_n) = n^{-1} \sum_i x_i$是一个充分统计量。 充分统计量可以被看作是对数据集中信息的压缩。检验统计量是否为充分统计量的标准方法是通过所谓的Neyman-Fisher分解准则：

定义(Neyman-Fisher分解)。 如果一个密度函数具有以下分解形式

$$f(x_1, ..., x_n; \theta) = g(t(x_1, ..., x_n), \theta)\, h(x_1, ..., x_n),$$然后 $t(x_1, ..., x_n)$是一个充分统计量。 上述建议我们可以将充分统计量 t与数据 $x_1, ...x_n$解耦。

有一类似然函数我们经常使用，因为它们具有良好的性质，并且在建模数据中的误差或噪声方面非常有用。
这个类别被称为指数族，上述正态似然函数是一个例子。

定义(指数族)。 密度函数 $f(x \mid \theta)$属于指数族，如果密度函数具有以下形式

$$f(x \mid \theta) = h(x)\, g(\theta)\, \exp\left(\eta(\theta)^T \cdot T(x) - A(\theta)\right),$$
其中 $T(x)$是数据的充分统计量， $\eta(\theta)$是一个函数（有时是参数的身份）， $h(x)$和 $g(\theta)$用于归一化密度。广泛的似然模型属于指数族，包括多元正态分布、二项分布、多项分布、泊松分布和指数分布。

3.2.1. 单变量正态分布

我们现在证明单变量正态分布属于指数族。

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-(x-\mu)^2/(2\sigma^2)\right)$$
$$\eta = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^T, \quad h(x) = \frac{1}{\sqrt{2\pi}}$$
$$T(x) = \begin{pmatrix} x, & x^2 \end{pmatrix}^T, \quad g(\eta) = \frac{\mu^2}{2\sigma^2} + \ln|\sigma|$$

3.2.2. 伯努利

我们现在考虑伯努利分布。 对于一个伯努利随机变量 $x \sim \mathrm{Be}(\pi)$，其中 π是随机变量 X的均值参数。 指数族分布如下

$$\begin{aligned}
\mathrm{Be}(x \mid \pi) &= \pi^x (1-\pi)^{1-x} \\
&= \exp\left[\log(\pi^x (1-\pi)^{1-x})\right] \\
&= \exp\left[x \log \pi + (1-x)\, \log(1-\pi)\right] \\
&= \exp\left[x(\log \pi - \log(1-\pi)) + \log(1-\pi))\right] \\
&= \exp\left[x \log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right]
\end{aligned}$$

将上述公式与指数族形式进行比较，我们有

$$\begin{aligned} h(x) &= 1 \\ T(x) &= x \\ \eta &= \log\left(\frac{\pi}{1-\pi}\right) \\ g(\eta) &= \log\left(\frac{1}{1-\pi}\right) \\ &= \log(1+\exp(\eta)) \end{aligned}$$

3.3. 最大后验估计

从前一节的推导中，如果我们假设标准线性回归模型，已知方差，我们可以指定似然函数为

$$\mathrm{Lik}(D;\beta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2}\right).$$

我们还可以根据James-Stein或收缩模型对效应大小参数 β进行先验设定

$$\pi(\beta) = \frac{1}{(2\pi)^{p/2}\gamma^{1/2}} \exp\left(-\frac{1}{2}\beta^T(\tau_0^2\mathbf{I}_p)^{-1}\beta\right),$$

根据贝叶斯定理，后验概率为 β是

$$\text{后验概率}(\beta \mid D) \propto \left[\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2}\right)\right] \times \frac{1}{(2\pi)^{p/2}\gamma^{1/2}} \exp\left(-\frac{1}{2}\beta^T(\tau_0^2\mathbf{I}_p)^{-1}\beta\right).$$

我们可以将后验的负对数写成

$$L = (2\sigma^2)^{-1}\sum_{i=1}^{n}\|y_i - \beta^T x_i\|^2 + \frac{1}{2\tau_0^2}\beta^T\beta,$$

可以重写为

$$\begin{aligned} L &= \frac{1}{n}\sum_{i=1}^{n}\|y_i - \beta^T x_i\|^2 + \frac{\sigma^2}{n\tau_0^2}\beta^T\beta \\ &= \frac{1}{n}\sum_{i=1}^{n}\|y_i - \beta^T x_i\|^2 + \lambda_n\beta^T\beta, \end{aligned}$$

其中正则化参数 λ_n现在是样本大小 n的函数，并且具有噪声方差与先验方差之比的解释。 我们可以最小化 L来获得所谓的最大后验概率(MAP) 估计器

$$\hat{\beta} = \arg\min_{\beta}\left[\frac{1}{n}\sum_{i=1}^{n}\|y_i - \beta^T x_i\|^2 + \lambda_n\beta^T\beta\right],$$

这实际上就是前一节中的收缩估计器，其中

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda_n n\mathbf{I})^{-1}\mathbf{X}^T Y.$$

3.4. 共轭先验

当先验和后验具有相同的形式时，我们说先验是相应似然函数的共轭先验。 对于指数族中的所有分布，我们可以推导出共轭先验。 假设先验分布为 $p(\eta \mid \tau)$，其中 τ表示超参数。 后验可以写成:$p(\eta \mid X) \propto p(X \mid \eta)\, p(\eta \mid \tau)$指数族的似然函数为:

$$p(X \mid \eta) = \left(\prod_{i=1}^{n} h(x_i)\right) \exp\left(\eta^T \sum_{i=1}^{n} T(x_i) - n g(\eta)\right)$$

假设先验分布的形式为

$$p(\eta \mid \tau) \propto \exp\left\{\eta^T \tau - \tau_0 g(\eta)\right\},$$

我们观察到超参数和参数 η之间存在内积结构。 那么后验分布可以写成:$p(\eta \mid X) \propto p(X \mid \eta)\, p(\eta \mid \tau)$

$$\begin{aligned} &\propto \exp\left(\eta^T \sum_{i=1}^{n} T(x_i) - n g(\eta)\right) \exp\left(\eta^T \tau - \tau_0 g(\eta)\right) \\ &= \exp\left\{\eta^T \left(\sum_{i=1}^{n} T(x_i) + \tau\right) - (n + \tau_0) g(\eta)\right\} \end{aligned}$$

后验分布与先验分布具有相同的指数族形式，后验超参数是将充分统计量的和添加到共轭先验的超参数中。 指数族是唯一存在共轭先验的分布族。 这是指数族的一个方便的特性，因为共轭先验简化了后验计算。 我们可以进行代数运算而不是微积分、积分。

3.4.0.1. 代数的观点。 如果一个先验分布族在采样下是封闭的，那么它被称为共轭先验分布族。这意味着对于任意的样本 $p(\ \eta) \in \mathcal{F}$，只要对于采样分布 $p(x \mid \eta)$，有 $p(\eta \mid x) \in \mathcal{F}$，那么先验分布族 $\mathcal{F}$就是共轭的。

有一个有趣的观察是在温和的条件下，共轭先验分布可以通过以下后验线性条件来刻画IE

$$\Big[\mathrm{IE}(X \mid \eta) \mid X = x\Big] = ax + b.$$

多元正态分布的有用性质

3.1. 条件和边缘

对于贝叶斯分析来说，了解如何编写多元正态分布的联合、边缘和条件分布非常有用。给定向量 $x \in \mathrm{IR}^p$，多元正态密度为

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

现在将向量分成两部分

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \text{大小为} \begin{bmatrix} q \times 1 \\ (p-q) \times 1 \end{bmatrix},$$

和

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \text{大小为} \quad \begin{bmatrix} q \times q & q \times (p-q) \\ (p-q) \times q & (p-q) \times (p-q) \end{bmatrix}.$$

现在我们陈述联合和边缘分布

$$x_1 \sim \mathrm{N}(\mu_1, \Sigma_{11}), \quad x_2 \sim \mathrm{N}(\mu_2, \Sigma_{22}), \quad x \sim \mathrm{N}(\mu, \Sigma),$$

以及条件密度

$$x_1 \mid x_2 \sim \mathrm{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

对于其他分区大小，相同的想法适用。

3.2. 共轭先验

3.2.1. 单变量正态分布

3.2.1.1. 固定方差，随机均值。 我们考虑参数 σ^2固定，因此我们对 μ的共轭先验感兴趣：

$$\pi(\mu \mid \mu_0, \sigma^2) \propto \frac{1}{\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right),$$

其中 μ_0和 σ^2是先验分布的超参数（当我们没有信息先验知识时，通常考虑 $\mu_0 = 0$ 和 σ^2较大）。

具有单变量正态似然和上述先验的后验分布将是

$$\text{Post}(\mu \mid x_1, ..., x_n) \sim \text{N}\left(\frac{\sigma_0^2}{\frac{\sigma^2}{n}+\sigma_0^2}\bar{x} + \frac{\sigma^2}{\frac{\sigma^2}{n}+\sigma_0^2}\mu_0, \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}\right).$$

3.2.1.2.固定均值，随机方差。我们将用两个参数化的尺度参数来表述这个设置：(1) 方差 σ^2，(2) 精度$\tau = \frac{1}{\sigma^2}$.

两个共轭分布是Gamma分布和逆Gamma分布（实际上它们是相同的分布，只是重新参数化）

$$\text{IG}(\alpha, \beta) : f(\sigma^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}\exp(-\beta(\sigma^2)^{-1}), \quad \text{Ga}(\alpha, \beta) : f(\tau) = \frac{\beta^\alpha}{\Gamma(\alpha)}\tau^{\alpha-1}\exp(-\beta\tau).$$

后验分布 σ^2是

$$\sigma^2 \mid x_1, ..., x_n \sim \text{IG}\Big(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum(x_i-\mu)^2\Big).$$

后验分布 τ是毫不奇怪的

$$\tau \mid x_1, ..., x_n \sim \text{Ga}\Big(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum(x_i-\mu)^2\Big).$$

3.2.1.3. 随机均值，随机方差。 我们现在将之前的先验结合起来，称之为贝叶斯分层模型：

$$\begin{aligned} x_i \mid \mu, \tau &\overset{iid}{\sim} \text{N}(\mu, (\tau)^{-1}) \\ \mu \mid \tau &\sim \text{N}(\mu_0, (\kappa_0\tau)^{-1}) \\ \tau &\sim \text{Ga}(\alpha, \beta). \end{aligned}$$

对于上述似然函数和先验分布，均值和精度的后验分布为$\mu \mid \tau, x_1, ..., x_n \sim \text{N}$

$$\Big(\frac{\mu_0\kappa_0 + n\bar{x}}{n+\kappa_0}, (\tau(n+\kappa_0))^{-1}\Big)$$

$$\tau \mid x_1, ..., x_n \sim \text{Ga}\Big(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum(x_i-\bar{x})^2 + \frac{n}{n+1}\frac{(\bar{x}-x_i)^2}{2}\Big).$$

3.2.2. 多元正态分布

给定一个向量 $x \in \text{IR}^p$，多元正态密度为

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

我们将使用精度矩阵而不是协方差，并考虑以下贝叶斯分层模型：

$$\begin{aligned} x_i \mid \mu, \Lambda &\overset{iid}{\sim} \text{N}(\mu, (\Lambda)^{-1}) \\ \mu \mid \Lambda &\sim \text{N}(\mu_0, (\kappa_0\Lambda)^{-1}) \\ \Lambda &\sim \text{Wi}(\Lambda_0, n_0), \end{aligned}$$

精度矩阵使用Wishart分布建模

$$f(\Lambda; V, n) = \frac{|\Lambda|^{(n-d-1)/2}\exp(-.5\text{tr}(\Lambda V^{-1}))}{2^{nd/2}|V|^{n/2}\Gamma_d(n/2)}.$$

对于上述似然函数和先验分布，均值和精度的后验分布为

$$\begin{aligned}
\mu \mid \Lambda, x_1, ..., x_n &\sim \mathrm{N}\Big(\frac{\mu_0\kappa_0 + n\bar{x}}{n+\kappa_0}, (\Lambda(n+\kappa_0))^{-1}\Big) \\
\Lambda \mid x_1, ..., x_n &\sim \mathrm{Wi}\left(n_0 + \frac{n}{2}, \Lambda_0 + \frac{1}{2}\left[\bar{\Sigma} + \frac{\kappa_0}{\kappa_0+n}(\bar{x}-\mu_0)(\bar{x}-\mu_0)^T\right]\right).
\end{aligned}$$

讲座 4
贝叶斯线性回归方法

贝叶斯推理背后的主要动机是一种一致的方法来建模不确定性以及推理的公理框架。
在本节中，我们将从贝叶斯公式重新表述多元线性回归。

贝叶斯推理涉及以概率分布和条件分布思考。 一个重要的概念是共轭先验。 在这门课程中，我们还将广泛使用多元正态分布及其性质。

4.1. 共轭先验

给定一个似然函数 $p(x \mid \theta)$和一个先验 $\pi(\theta)$，可以将后验写为

$$p(\theta \mid x) = \frac{p(x \mid \theta)\pi(\theta)}{\int_{\theta'} p(x \mid \theta')\pi(\theta') \, \mathrm{d}\theta'} = \frac{p(x, \theta)}{p(x)},$$

其中 $p(x)$ 是数据的边缘密度， $p(x, \theta)$ 是数据和参数 θ的联合密度。

先验和似然共轭的想法是先验和后验密度属于同一族。 现在我们举几个例子来说明这个想法。

Beta, 二项式分布：考虑具有固定的试验次数 n的二项式似然函数

$$f(x \mid p, n) = \binom{n}{x} p^x (1-p)^{n-x},$$

感兴趣的参数（成功的概率）为 $p \in [0,1]$。 对于 p的自然先验分布是具有密度的Beta分布

$$\pi(p; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \quad p \in (0,1) \text{ 且 } \quad \alpha, \beta > 0,$$

其中 $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ 是一个归一化常数。 给定先验和似然密度，后验密度模除归一化常数将采取以下形式

$$\begin{aligned} f(p \mid x) &\propto \left[\binom{n}{x} \times \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right] p^x(1-p)^{n-x} \times p^{\alpha-1}(1-p)^{\beta-1}, \\ &\propto p^{x+\alpha-1}(1-p)^{n-x+\beta-1}, \end{aligned}$$

这意味着 p的后验分布也是一个Beta分布，其中

$$p \mid x \sim \mathrm{Beta}(\alpha + x,\, \beta + n - x).$$

正态分布，正态分布：给定一个未知均值的正态分布，似然函数的密度为

$$f(x \mid \theta,\, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}(x-\theta)^2\right),$$

可以指定一个正态先验分布

$$\pi(\theta; \theta_0,\, \tau_0^2) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\theta_0)^2\right),$$

具有超参数 θ_0和 τ_0。 得到的后验分布将具有以下密度函数

$$f(\theta \mid x) \propto \exp\left(-\frac{1}{2\sigma^2}(x-\theta)^2\right) \times \exp\left(-\frac{1}{2\tau_0^2}(\theta-\theta_0)^2\right),$$

经过完成平方和重新排序后，可以写成

$$\theta \mid x \sim \mathrm{N}(\theta_1, \tau_1^2), \quad \theta_1 = \frac{\frac{\theta_0}{\tau_0^2} + \frac{x}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}, \quad \tau_1^2 = \frac{1}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}.$$

4.2. 贝叶斯线性回归

我们从似然函数开始

$$f(Y \mid \mathbf{X},\, \beta,\, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2}\right).$$

以及先验

$$\pi(\beta) \propto \exp\left(-\frac{1}{2\tau_0^2}\beta^T\beta\right).$$

后验的密度函数为

$$\text{后验概率}(\beta \mid D) \propto \left[\prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\|y_i - \beta^T x_i\|^2}{2\sigma^2}\right)\right] \times \frac{1}{(2\pi)^{p/2}\gamma^{1/2}} \exp\left(-\frac{1}{2\tau_0^2}\beta^T\beta\right).$$

通过一些操作，上述可以重写为多元
正态分布

$$\beta \mid Y, \mathbf{X},\, \sigma^2 \sim \mathrm{N}_p(\mu_1, \Sigma_1)$$

其中

$$\Sigma_1 = (\tau_0^{-2}\mathbf{I}_p + \sigma^{-2}\mathbf{X}^T\mathbf{X})^{-1}, \quad \mu_1 = \sigma^{-2}\Sigma_1\ \mathbf{X}^T Y.$$

注意上述分布与MAP估计器的相似之处。 将上述估计器的均值与MAP估计器相关联。

预测分布：给定数据 $D = \{(x_i, y_i)\}_{i=1^n}$和一个新值 x_*，我们希望估计 y_*。 可以使用后验概率进行估计，称为后验预测分布。

$$f(y_* \mid D, x_*, \sigma^2, \tau_0^2) = \int_{\mathrm{IR}^p} f(y_* \mid x_*, \beta, \sigma^2) f(\beta \mid Y, \mathbf{X}, \sigma^2, \tau_0^2) \, \mathrm{d}\beta,$$

其中通过一些操作

$$y_* \mid D, x_*, \sigma^2, \tau_0^2 \sim \mathrm{N}(\mu_*, \sigma_*^2),$$

其中

$$\mu_* = \frac{1}{\sigma^2} \Sigma_1 \mathbf{X}^T Y \, x_*, \quad \sigma_*^2 = \sigma^2 + x_*^T \Sigma_1 x_*.$$

功能分析复习*

函数空间是一个函数的空间，其中每个函数可以被看作欧几里得空间中的一个点。 功能分析大致上是对函数空间的数学理解。 在下一讲中，我们将学习一个非常有用的函数空间，称为再生核希尔伯特空间（riches），它在非线性回归中被广泛使用。

4.1. 希尔伯特空间

例子。以下是在实数子集上定义的三个函数空间的示例。 在这些示例中，我们考虑的实数子集是 $x \in [a, b]$，其中例如 $a = 0$和 $b = 10$。

(1) $C[a,\ b]$是所有实值连续函数的集合，其中 $x \in [a,\ b]$。
$y = x^3$属于 $C[a,\ b]$，而 $y = \lceil x \rceil$不属于 $C[a,\ b]$。
(2) $L_2[a,\ b]$ 是所有在 $x \in [a,\ b]$上可积的平方函数的集合。如果
$(\int_a^b |f(x)|^2\, dx)^{1/2} < \infty$，那么 $f \in L_2[a,\ b]$。
$y = x^3$在 $L_2[a, b]$中，$y = x^3 + \delta(x - c)$也在其中，其中$a < c < b$，然而第二个函数在 $x = c$处未定义。
(3) $L_1[a,\ b]$是所有绝对值在
$x \in [a,\ b]$上可积的函数集合。
$y = x^3$在 $L_1[a, b]$中，$y = x^3 + \delta(x - c)$也在其中，其中$a < c < b$，然而第二个函数在 $x = c$处未定义。

定义。 一个带范数的向量空间是一个范数被定义的空间 $\mathcal{F}$。. 一个函数$\|\cdot\|$是一个范数，当且仅当对于所有$f,\ g \in \mathcal{F}$

(1) $\|f\| \geq 0$且 $\|f\| = 0$ 当且仅当 $f = 0$。
(2) $\|f + g\| \leq \|f\| + \|g\|$
(3) $\|\alpha f\| = |\alpha|\ \|f\|$.

注意，如果除了 $\|f\| = 0$ iff $f = 0$之外，所有条件都满足，则空间具有半范数而不是范数。

定义。 内积空间是一个定义了内积的线性向量空间 $\mathcal{E}$。. 一个实值函数 $\langle \cdot, \cdot \rangle$是一个内积，当且仅当对于所有的$f, g, h \in \mathcal{E}$和 $\alpha \in \mathrm{IR}(1)$ $\langle f, g \rangle = \langle g, f \rangle$

(2) $\langle f + g,\ h \rangle = \langle f,\ h \rangle + \langle g,\ h \rangle$并且 $\langle \alpha f,\ g \rangle = \alpha \langle f,\ g \rangle$
(3) $\langle f,\ f \rangle \geq 0$并且 $\langle f,\ f \rangle = 0$当且仅当 $f = 0$.

给定一个内积空间，范数定义为 $\|f\| = \sqrt{\langle f, f \rangle}$并且向量之间的角度可以定义。

定义。 对于一个范数空间 $\mathcal{A}$一个子空间$\mathcal{B} \subset \mathcal{A}$在 $\mathcal{A}$中是稠密的当且仅当 $\mathcal{A} = \bar{\mathcal{B}}$。其中 $\bar{\mathcal{B}}$是集合 $\mathcal{B}$的闭包。

定义。 一个范数空间 $\mathcal{F}$是可分的当且仅当 $\mathcal{F}$有一个可数的稠密子集。

例子。所有有理点的集合在实数线上是稠密的，因此实数线是可分的。 注意，有理点的集合是可数的。

反例。 具有上确界范数的 $[0, 1]$上的右连续函数空间是不可分的。. 例如，阶梯函数

$$f(x) = U(x - a) \ \forall a \in [0, 1]$$

不能用可数个函数的上确界范数逼近，因为跳跃必须发生在 a处，而所有 a的集合是不可数的。

定义。 在一个赋范空间 $\mathcal{F}$中，序列 $\{x_n\}$被称为柯西序列，如果 $\lim_{n\to\infty} \sup_{m \geq n} \|x_n - x_m\| = 0$。

定义。 一个赋范空间 $\mathcal{F}$被称为完备的，如果其中的每个柯西序列都收敛。

定义。*Hilbert*空间， $\mathcal{H}$是一个完备的内积空间，可分离，通常是无限维的。*Hilbert*空间具有可数的基础。

例子。 以下是*Hilbert*空间的例子。

(1) IR^n是*Hilbert*空间的典型例子。 空间中的每个点 $x \in \mathrm{IR}^n$都可以表示为一个向量 $x = \{x_1, ..., x_n\}$，该空间中的度量为 $\|x\| = \sqrt{\sum}$ 该空间具有非常自然的基础，由 n个基函数组成 $e_1 = \{1, 0, ..., 0\}$, $e^2 = \{0, 1, ..., 0\}$,...,$e_n = \{0, 0, ..., 1\}$。 向量x和基向量e_i之间的内积简单地是 x在第$_i$个坐标上的投影 $x_i = \langle x, e\ i \rangle$。

注意，这不是一个无限维的希尔伯特空间。

(2) L_2也是一个希尔伯特空间。 这个希尔伯特空间是无限维的。

4.2. 函数和算子

定义。 在希尔伯特空间 $\mathcal{H}$上的线性泛函是一个从H到IR的线性变换 $T : V \to \mathrm{IR}$。

线性泛函接受一个希尔伯特空间中的元素，并输出一个实数，积分是线性泛函的一个例子。

定理 (里兹表示定理)。 设 V是一个有限维内积空间， $T : V \to \mathrm{IR}$是一个线性泛函。那么存在一个向量$w \in V$使得对于所有的 $v \in V$，有$Tv = \langle v, w \rangle$。

积分变换是算子的一个例子（在课程的其余部分中，所有算子的例子都是积分变换）。 算子 T将一个向量空间映射到另一个。

定义。 一个积分变换 T将一个函数映射为另一个函数，如下所示

$$g(u) = (T\,f)(u) := \int_{t_1}^{t_2} K(t,\,u) f(t)\,\mathrm{d}t.$$

第5讲

再生核希尔伯特空间

再生核希尔伯特空间（rkhs）是具有一些非常好的性质的假设空间。 这些空间的主要特性是再生性质，它将希尔伯特空间中的范数与线性代数联系起来。 在高斯过程的背景下，这类函数还有一个很好的解释。 因此，它们在计算、统计和功能方面都非常重要。

5.1. 再生核希尔伯特空间（rkhs）

我们将使用两种表述来描述rkhs。 第一种表述较为具体和构造性。 第二种表述较为一般和抽象。 这两种表述的关键思想是存在一个核函数 $K : X \times X \ \rightarrow$IR，这个核函数与之相关联的希尔伯特空间 $\mathcal{H}_K$对于优化和推断具有奇妙的性质。

我们将在下一讲中详细研究的算法如下

$$f := \arg\min_{f \in \mathcal{H}_k} \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}_K},$$

其中 $\mathcal{H}_K$ 是一个rkhs， $\|f\|_{\mathcal{H}_K}$ 是由再生核 K定义的特定范数。 rkhs的美妙之处在于上述无限维函数空间中的优化问题可以被重新表述为一个仅涉及向量和矩阵的二次规划问题。

在本讲的剩余部分，我们将将希尔伯特空间限制在一个紧致域 X内。

5.1.1. 构造性表述

在本小节中，rkhs的发展在大多数支持向量机（SVM）和核机器的表述中都有所体现。它不太通用，因为它依赖于再生核是Mercer核。然而，它对于大多数人来说需要较少的函数分析知识，并且更加直观。

我们首先定义核函数或再生核函数。

定义。再生核函数（rk），K（$\cdot$，$\cdot$）是一个对称的实值函数，其两个变量s，$\mathbf{t}\in X$。

K（s, t）： $X\times X\rightarrow$IR。

此外，K（s, t）必须是正定的，即对于所有实数a_1，...，a_n和t_1，...，$t_n \in X$。

$$\sum_{i,j=1}^{n} a_i \cdot a_j K(t_i, t_j) \geq 0。$$

如果上述不等式严格成立，则K（s, t）是严格正定的。

在这个表述中，我们考虑连续再生核函数K：$X \times X \to \mathrm{IR}$。我们通过以下积分变换定义一个积分算子$LK$：$L_2[X] \to C[X]$（5.1）。

$$L_K f := \int_X K(s, t) f(t) dt = g(t).$$

如果 K是正定的，那么 L_K也是正定的（反之亦然），因此（5.1）的特征值是非负的。

我们将（5.1）的特征值和特征向量表示为 $\{\lambda_1, ..., \lambda_k\}$和 $\{\phi_1, ..., \phi_k\}$ 分别为

$$\int_X K(s, t) \phi_k(t) dt = \lambda_k \phi_k(t) \ \ \forall k.$$

现在我们陈述Mercer定理。

定理。 给定由对称正定核 K定义的积分方程的特征函数和特征值

$$\int_X K(s, t) \phi(s) ds = \lambda \, \phi(t).$$

核函数的展开式

$$K(s, t) = \sum_j \lambda_j \phi_j(s) \phi_j(t),$$

其中收敛性是在 $L_2[X]$范数中定义的。

我们可以将rkhs定义为由核函数定义的积分算子的特征函数张成的函数空间$\mathcal{H}_K = \{ f \,|\, f(s) =$

$$\sum_k c_k \phi_k(s) \text{ 和 } \ \|f\|_{\mathcal{H}_K} < \infty\},$$

其中rkhs范数 $\| \cdot \|_{\mathcal{H}_K}$ 定义如下

$$\|f(s)\|^2_{\mathcal{H}_K} = \left\langle \sum_j c_j \phi_j(s), \sum_j c_j \phi_j(s) \right\rangle^2_{\mathcal{H}_K} := \sum_j \frac{c_j^2}{\lambda_j}.$$

类似地，内积的定义如下

$$\langle f, g \rangle = \left\langle \sum_j c_j \phi_j(s), \sum_j d_j \phi_j(s) \right\rangle_{\mathcal{H}_K} := \sum_j \frac{d_j c_j}{\lambda_j}.$$

作业问题的一部分将是证明表示器属性

$$\langle f(\cdot, K(\cdot, x)) \rangle_{\mathcal{H}_k} = f(x),$$

使用Mercer定理和上述rkhs范数的定义。

5.1.2. 核函数和特征空间

rkhs的概念在SVM和核机器文献中被称为核技巧。

定义域中的点 $x \in X \subset \mathrm{IR}^d$ 通过再生核的特征值和特征函数映射到更高维度的空间（该空间的维度等于积分算子定义的非零特征值的数量）$x \to \Phi(x) := \{$

$$\sqrt{\lambda_1}\phi_1(x), ..., \sqrt{\lambda_k}\phi_k(x)\}.$$

两个映射到特征空间的点的标准 L_2内积可以通过一个核函数来计算，根据Mercer定理$K(s,t) = \langle \Phi(s), \Phi(t) \rangle_L$

$_2$.

5.1.3. 核函数的例子

任何（半）正定函数都可以用作核函数。例如

(1) 线性核函数: $k(u, v) = \langle u, v \rangle$
(2) 多项式核函数: $k(u, v) = (\langle u, v \rangle + b)^p$
(3) 高斯核函数: $k(u, v) = \exp(-\kappa \|u - v\|^2)$
(4) 双指数核函数: $k(u, v) = \exp(-\kappa \|u - v\|)$

5.2. 抽象公式

命题.线性评估函数 L_t*evaluates each function in a Hilbert space* $f \in \mathcal{H}$ *at a point* t. 它将 $f \in \mathcal{H}$关联到一个数字 $f(t) \in \mathrm{IR}$, $L_t[f] = f(t)$.

(1) $L_t[f + g] = f(t) + g(t)$
(2) $L_t[af] = af(t)$.

例子. Del*t*a函数 $\delta(x - t)$是 $C[a, b]$的线性评估函数

$$f(t) = \int_a^b f(x)\delta(x - t)dx.$$

命题. 如果存在一个 M，使得对于希尔伯特空间H中的所有函数 f，有$|L_t[f]| = |f(t)| \leq M\|f\|$成立，则线性评估函数是有界的。其中 $\|f\|$是希尔伯特空间的范数。

例子. 对于希尔伯特空间 $C[a, b]$和上确界范数，存在一个有界的线性评估函数，因为对于 $C[a, b]$中的所有函数，有 $|f(x)| \leq M$成立。. 这是由于定义域的连续性和紧致性。 评估函数简单地是$L_t[f] : t \to f(t)$，其中 M= 1。

反例。 对于Hib**ert**空间 $L_2[a, b]$，不存在有界线性评估函数。. 下面的函数在 $L_2[a, b]$中，$y = x^3 + \delta(x - c)$，其中 $\mathbf{a} < c < b$。

在点 $x = c$处，不存在 M 使得 $|f(c)| \leq M$，因为该函数被评估为 “∞”。 这是一个在空间中甚至没有被逐点定义的函数的例子。

定义。 如果$H\mathbf{i}lb\mathbf{e}r\mathbf{t}$空间有一个有界线性评估函数 L_t，则它是一个再生核$H\mathbf{i}lb\mathbf{e}r\mathbf{t}$空间（$rkhs$）， $\mathcal{H}_K$。

rkhs的以下性质非常重要，是Riesz表示定理的结果。

命题。 如果 $\mathcal{H}_K$是一个$\mathbf{r}kh\mathbf{s}$，那么存在一个元素在空间$_{Kt}$中，具有以下性质：对于所有 $f\in \mathcal{H}_K$，$L_t[f] = \langle\ K_t, f\rangle = f(t)$。

内积在$rkhs$范数中，元素 K_t被称为评估的代表
的 t。

备注。 上述性质有点令人惊讶，它表明如果一个希尔伯特空间有一个有界线性评估函数，那么这个空间中存在一个元素，通过内积评估空间中的所有函数。

在空间 $L_2[a,\ b]$中，我们说delta函数评估 $L_2[a,\ b]$中的所有函数。

$$L_t[f] = \int_a^b f(x)\delta(x-t)dx.$$

然而，delta函数不在 $L_2[a,\ b]$中。

一个rkhs与其再生核之间存在着深刻的关系。 这可以通过以下定理来描述。

定理。对于每个再生核希尔伯特空间（$\mathbf{r}kh\mathbf{s}$），都存在一个唯一的再生核，反之亦然，给定一个正定函数 $K\ on\ X\times X$，我们可以构造一个以 K为再生核的实值函数的唯一$\mathbf{r}kh\mathbf{s}$。

证明。
如果 $\mathcal{H}_K$是一个rkhs，那么根据Reisz再生核定理，rkhs中存在一个元素是表示器的评估。 我们定义再生核K（s，t）：$=\langle K_s，\ K_t\rangle$

其中 K_s和 K_t是在 s和 t处的表示器。 根据希尔伯特空间的性质和表示器的性质，以下成立

$$\begin{aligned}\left\|\sum_j a_j K_{t_j}\right\|^2 &\geq 0\\ \left\|\sum_j a_j K_{t_j}\right\|^2 &= \sum_{i,j} a_i a_j \langle K_{t_i},\ K_{t_j}\rangle\\ \sum_{i,j} a_i a_j K(t_i,\ t_j) &= \sum_{i,j} a_i a_j \langle K_{t_i},\ K_{t_j}\rangle.\end{aligned}$$

因此 $K(s,\ t)$ 是正定的。

我们现在证明反过来。 给定一个rk $K(\cdot,\cdot)$，我们构造 $\mathcal{H}_K$。 对于每个 $t\in X$，我们定义实值函数

$$K_t(\cdot) = K(t,\cdot)。$$

我们可以证明rkhs只是由集合 K_t生成的函数空间的完备性。

$$\mathcal{H} = \{f \,|\, f = \sum_i^i a_i K_{t_i} \quad \text{其中} \quad a_i \in \mathrm{IR}, t_i \in X, \quad i \in \mathrm{IN}\}$$

具有以下内积

$$\left\langle \sum_i a_i K_{t_i}, \sum_i a_i K_{t_i} \right\rangle = \sum_{i,j} a_i a_j \langle K_{t_i}, K_{t_j} \rangle = \sum_{i,j} a_i a_j K(t_i, t_j)。$$

由于 $K(\cdot, \cdot)$ 是正定的，上述内积是良定义的。 对于任意 $f \in \mathcal{H}_K$ ，我们可以验证

$$\langle K_t, f \rangle = f(t)$$

因为对于上述线性空间中的任何函数，范数收敛意味着逐点收敛

$|f_n(t) - f(t)| = |\langle f_n - f, K_t \rangle| =\leq \|f_n - f\| \|K_t\|$，最后一步是由于柯西-施瓦茨不等式。因此，该空间中的每个柯西序列都收敛，并且它是完备的。 □

讲座 6

非线性回归

我们将在下一讲中详细研究的算法如下

$$f := \arg\min_{f\in\mathcal{H}_k} \frac{1}{n}\sum_{i=1}^{n}(f(x_i)-y_i)^2 + \lambda\|f\|^2_{\mathcal{H}_K}.$$

我们将看到上述最小化器作为一种特定形式，非常适合优化，这是由于表示定理。

6.1. 表示定理的结果

以下是三种标准的正则化方法：

（1）蒂赫诺夫正则化：通过添加惩罚项间接约束假设空间。

$$\min_{f\in\mathcal{H}}\left[n^{-1}\sum_{i=1}^{n}V(f,\,z_i)+\lambda\Omega(f)\right].$$

(2) 伊万诺夫正则化：直接约束假设空间

$$\min_{f\in\mathcal{H}}\left[n^{-1}\sum_{i=1}^{n}V(f,\,z_i)\right] \quad \text{满足} \quad \Omega(\qquad f)\le\tau.$$

(3) 菲利普斯正则化：直接约束假设空间

$$\min_{f\in\mathcal{H}}\Omega(f) \text{ 满足} \qquad \left[n^{-1}\sum_{i=1}^{n}V(f,\,z_i)\right]\le\kappa.$$

考虑到rkhs范数将作为正则化函数

$$\Omega(f)=\|f\|^2_{\mathcal{H}_K}.$$

这定义了以下优化问题：

$$(P1) \qquad \min_{f\in\mathcal{H}} \left[n^{-1}\sum_{i=1}^{n} V(f,\, z_i) + \lambda \|f\|^2_{\mathcal{H}_K} \right],$$

$$(P2) \qquad \min_{f\in\mathcal{H}} \left[n^{-1}\sum_{i=1}^{n} V(f,\, z_i) \right] \quad \text{满足} \quad \|f\|^2_{\mathcal{H}_K} \le \tau,$$

$$(P3) \qquad \min_{f\in\mathcal{H}} \|f\|^2_{\mathcal{H}_K} \text{满足} \quad \left[n^{-1}\sum_{i=1}^{n} V(f,\, z_i) \right] \le \kappa.$$

上述所有优化问题都涉及包含无限多个函数的函数空间。 使用第5.1.1节中的公式，我们可以将任何函数写成rhks中的形式

$$\mathcal{H}_K = \left\{ f \,|f(x) = \sum_k c_k \phi_k(x) \right\},$$

所以优化过程是在系数 c_k上进行的。 展开式中非零系数的数量定义了rkhs的维度，这可以是无限的，例如高斯核函数。

所有上述优化问题的一个令人惊奇的方面是，最小化问题的形式满足

$$f(x) = \sum_{i=1}^{n} c_i K(x,\, x_i)。$$

所以优化过程是在实数变量上进行的。 这在以下的“Represe*n*ter定理”中得到了形式化。

定理。 给定一组点 $\{(x_i, y_i)\}^{ni}_{=1}$，形式为

$$f(x) = \sum_{i=1}^{n} c_i K(x,\, x_i),$$

是以下优化过程的最小化器。

$$c\left((f(x_1), y_1), ..., (f(x_n), y_n)\right) + \lambda g(\|f\|_{\mathcal{H}_K}),$$

其中 $\|f\|_{\mathcal{H}_K}$ 是一个*rkhs*范数， $g(\cdot)$是单调递增的，而 c是任意的代价函数。

过程（P1）是上述定理中所述优化过程的特殊情况。

证明。 为了简化表示，证明中的所有范数和内积都是rkhs范数和内积。

假设函数 f具有以下形式

$$f = \sum_{i=1}^{n} b_i \phi_i(x_i) + v,$$

其中

$$\langle \phi_i(x_i),\, v \rangle = 0 \quad \forall i = 1, .., n.$$

正交条件简单地确保 v不在 $\{\phi_i(x_i)\}^{ni}_{=1}$的张成空间中。

所以对于任意点 x_j $(j = 1, ..., n)$

$$f(x_j) = \left\langle \sum_{i=1}^{n} b_i \phi(x_i) + v,\, \phi(x_j) \right\rangle = \sum_{i=1}^{n} b_i \langle \phi(x_i),\, \phi(x_j) \rangle,$$

所以 v对代价函数

$$c\left((f(x_1), y_1), ..., (f(x_n), y_n)\right)$$没有影

现在我们来看一下rkhs范数

$$g(\|f\|) = g\left(\left\|\sum_{i=1}^{n} b_i \phi_i(x_i) + v\right\|\right) = g\left(\sqrt{\left\|\sum_{i=1}^{n} b_i \phi_i(x_i)\right\|^2 + \|v\|^2}\right) \geq g\left(\sqrt{\left\|\sum_{i=1}^{n} b_i \phi_i(x_i)\right\|^2}\right).$$

因此，额外的因子 v增加了rkhs范数，并对成本函数产生影响，因此必须为零，函数的形式为$f =$

$$\sum_{i=1}^{n} b_i \phi_i(x_i),$$

通过再生性质

$$f(x) = \sum_{i=1}^{n} a_i K(x,\, x_i). \ \square$$

作业：证明另外两种正则化形式的再现定理。

6.2. 核岭回归

核岭回归（KRR）算法已经被发明和重新发明了很多次，并被称为各种名称，如正则化网络，最小二乘支持向量机（LSSVM），正则化最小二乘分类（RLSC）。

我们从Tikhonov正则化开始

$$\min_{f \in \mathcal{H}_K} \left[n^{-1} \sum_{i=1}^{n} V(f,\, z_i) + \lambda \Omega(f) \right]$$

然后将正则化函数设置为RKHS范数

$$\Omega(f) = \|f\|^2_{\mathcal{H}_K}$$

并使用平方损失函数

$$n^{-1} \sum_{i=1}^{n} V(f,\, z_i) = n^{-1} \sum_{i=1}^{n} (f(x_i) - y_i)^2.$$

由此产生的优化问题是

$$\min_{f \in \mathcal{H}_K} \left[n^{-1} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|^2_{\mathcal{H}_K} \right], \tag{6.1}$$

我们通过Representer定理知道其最小化器的形式为

$$f(x) = \sum_{i=1}^{n} c_i K(x,\, x_i)。$$

这意味着我们只需要解决关于 c_i的优化问题。 这将把优化无限维函数的问题转化为优化N个实数的问题。

使用Representer定理，我们推导出实际解决核岭回归的优化问题。

我们首先定义一些符号。 我们将使用符号 K来表示核函数 K或N $\times$N矩阵 K，其中$K_{ij} \equiv K(x_i, x_j)$。

使用这个定义，函数 $f(x)$ 在训练点 x_j处的值可以用矩阵表示为

$$\begin{aligned} f(x_j) &= \sum_{i=1}^{n} K(x_i, x_j)c_i \\ &= [Kc]_j, \end{aligned}$$

其中 $[Kc]_j$,是通过将核矩阵 K与向量 c相乘得到的向量的第j个元素。 在这个符号表示法中，我们可以将方程(6.1)重写为

$$\min_{f \in \mathcal{H}_K} \frac{1}{n} (Kc - y)^2 + \lambda \|f\|_K^2,$$

其中 y是 y值的向量。 同时，根据表示定理，RKHS范数可以使用线性代数进行计算$||f||_K^2 = c^T Kc,$

其中 c^T是向量 c的转置。 将上述范数代入方程(6.1)得到一个关于向量 c的优化问题

$$\arg\min_{c \in \mathbb{R}^n} \left[g(c) := \frac{1}{\ell}(Kc - y)^2 + \lambda c^T Kc. \right]$$

这是一个凸可微函数 c的，所以我们可以通过对 c求导并将导数设为0来简化最小化过程。

$$\frac{\partial g(c)}{\partial c} = \frac{2}{\ell} K(Kc - y) + 2\lambda Kc = 0.$$

我们证明了上述方程的解是以下线性系统$c = (K + \lambda \ell I)^{-1}\mathrm{y}$，其中 I是单位矩阵

：

微分	$0 = \frac{2}{\ell} K(Kc - y) + 2\lambda Kc$
乘法	$K(Kc) + \lambda \ell Kc = Ky$
“左乘以 K^{-1}”	$(K + \lambda \ell I)c = y$
求逆	$c = (K + \lambda \ell I)^{-1} y.$

矩阵 $K + \lambda \ell I$是正定的，如果 λ不太小，它将具有良好的条件。

线性系统的几个性质是：

(1) 如果$\lambda>0$，矩阵 $(K+\lambda\ell I)$ 是可逆的。 当 $\lambda\to 0$时，正则化最小二乘解趋向于标准高斯最小二乘解，最小化经验损失。 当 $\lambda\to\infty$时，解趋向于 $f(\mathbf{x})=0$。

(2) 在实践中，我们实际上不会求逆 $(K+\lambda\ell I)$，而是使用求解线性系统的算法。
(3) 为了使用这种方法，我们需要计算并存储整个核矩阵 K。 这使得在处理非常大的训练集时变得不切实际。

最后，没有什么能阻止我们使用上述算法进行分类。 通过这样做，我们实质上将我们的分类问题视为一个回归问题，其中y的值为1或-1。

6.2.1. 求解 c

共轭梯度（CG）算法是解正定线性系统的一种流行算法。 对于本课程的目的，我们需要知道CG是一种迭代算法。 CG中的主要操作是将向量 v乘以矩阵 A。 请注意，矩阵 A不一定需要明确提供，我们只需要找到一种形成乘积 Av的方法。

对于普通的正半定系统，CG将与直接方法竞争。 如果有一种快速乘以 A的方法，CG可能会快得多。

例子。 假设我们的核函数 K是线性的：

$$K(x,\,y)=\langle x,\,y\rangle。$$

那么我们的解 x 可以写成

$$\begin{aligned} f(x) &= \sum c_i\langle x_i,\,x\rangle \\ &= \left\langle \left(\sum c_i x_i\right),\,x\right\rangle \\ &:= \langle w,x\rangle, \end{aligned}$$

我们可以在时间 d而不是时间 nd内将我们的函数应用于新的例子。这是$Tikhonov$正则化与线性核的一般特性，与平方损失的使用无关。

我们可以使用CG算法来解决带有线性核（$K(\mathbf{x_1},\mathbf{x_2})=\mathbf{x_1}\cdot\mathbf{x_2}$）的正则化最小二乘回归问题，从而节省大量时间。 对于任意的核函数，我们必须明确地形成一个乘积 Kv— 我们将一个向量乘以 K。 对于线性核，我们注意到 $K=AA^T$，其中 A是一个以数据点为行向量的矩阵。利用这一点：

$$\begin{aligned} (K+\lambda nI)v &= (AA^T+\lambda nI)v \\ &= A(A^Tv)+\lambda nIv. \end{aligned}$$

假设我们有 n个点在d个维度中。 明确地形成核矩阵 K需要 n^2 d时间，并且将向量乘以 K需要 n^2时间。

如果我们使用线性表示，我们不需要支付任何费用来形成核矩阵，并且将向量乘以 K需要$2dn$时间。

如果 $d\ll n$，我们可以节省大约一个因子的 n每次迭代2_d 内存的节省更加重要，因为我们无法完全存储核矩阵

对于大型训练集，如果需要重新计算核矩阵的条目，每次迭代将花费 n^2d的时间。

还要注意，如果训练数据稀疏（它们由大量维度组成，但每个点的大多数维度为零），则将向量乘以 K的成本可以写为$2\bar{d}n$，其中$\bar{d}$是每个数据点的平均非零条目数。

这在与文本相关的应用中经常是这样，其中维度
将对应于“字典”中的单词。 可能有成千上万个
单词，但在任何给定的文档中只会出现几百个。

6.3. 三种形式的等价性

正则化的三种形式具有一定的等价性。 等价性是给定一组点 $\{(x_i, y_i)\}^{ni}_{=1}$可以设置参数$\lambda, \tau,$和 κ使得相同的函数 $f(x)$ 最小化 (P1), (P2), 和 (P3)。 鉴于这种等价性和 (P1) 的表示定理，很明显 (P2) 和 (P3) 也有表示定理。

命题。 给定一个凸损失函数，以下优化
过程是等价的

$$(P1) \qquad \min_{f\in\mathcal{H}_K} \left[n^{-1}\sum_{i=1}^{n} V(f,\, z_i) + \lambda \|f\|^2_{\mathcal{H}_K} \right],$$

$$(P2) \qquad \min_{f\in\mathcal{H}_K} \left[n^{-1}\sum_{i=1}^{n} V(f,\, z_i) \right] \quad \text{满足} \qquad \|f\|^2_{\mathcal{H}_K} \le \tau,$$

$$(P3) \qquad \min_{f\in\mathcal{H}_K} \|f\|^2_{\mathcal{H}_K} \text{ 满足} \qquad \left[n^{-1}\sum_{i=1}^{n} V(f,\, z_i) \right] \le \kappa.$$

等价是指如果 $f_0(x)$是其中一个问题的解，则存在参数τ, κ, λ使得 $f_0(x)$是其他问题的解。

证明。

令 f_0为（P2）的解，对于固定的 τ，并假设优化下的约束是紧的（$\|f_0\|^2_{\mathcal{H}} = \tau$）。让 $\left[n^{-1}\sum_{i=1}^{n} V(f_0, z_i)\right] = b$。通过检查，当 $\kappa = b$时，（P3）的解将是 f_0。

令 f_0为（P3）的解，对于固定的 κ，并假设优化下的约束是紧的（$[n^{-1}\sum^{n}_{i=1} V(f_0, z_i)] = \kappa$）。让 $\|f_0\|^2_{\mathcal{H}_K} = b$。通过检查，当 $\tau = b$时，（P2）的解将是 f_0。

对于(P2)和(P3)，上述论证可以调整为约束不紧但解不一定唯一的情况。

设 f_0为固定 τ下(P2)的解。 使用拉格朗日乘子，我们可以将(P2)重新写成

$$\min_{f\in\mathcal{H}_K,\alpha} \left[n^{-1}\sum_{i=1}^{n} V(f,\, z_i) \right] + \alpha \left(\|f\|^2_{\mathcal{H}_K} - \tau \right), \tag{6.2}$$

其中 $\alpha \ge 0$，最优 $\alpha = \alpha_0$。 根据Karush-Kuhn-Tucker (KKT)条件（互补松弛性）在最优性处

$$\alpha_0 \left(\|f_0\|^2_{\mathcal{H}_K} - \tau \right) = 0.$$

如果 $\alpha_0 = 0$，则 $\|f\|_{\mathcal{H}_K}^2 < \tau$并且我们可以将方程(6.2)重新写成

$$\min_{f\in\mathcal{H}_K} \left[n^{-1}\sum_{i=1}^{n} V(f, z_i) \right],$$

这对应于 $\lambda = 0$的(P1)，且最小值为 f_0。如果 $\alpha_0 > 0$，则$\|f\|_{\mathcal{H}_K}^2 = \tau$并且我们可以将方程(6.2)重新写成以下等价的优化过程

$$(P2) \qquad \min_{f\in\mathcal{H}_K} \left[n^{-1}\sum_{i=1}^{n} V(f, z_i) \right] + \alpha_0 \left(\|f\|_{\mathcal{H}_K}^2 - \tau \right),$$

$$(P2) \qquad \min_{f\in\mathcal{H}_K} \left[n^{-1}\sum_{i=1}^{n} V(f, z_i) \right] + \alpha_0 \|f\|_{\mathcal{H}_K}^2,$$

这对应于（P1），其中 $\lambda = \alpha_0$，最小值为 f_0。

设 f_0为固定 κ下（P3）的解。使用拉格朗日乘数法，我们可以将（P3）重写为

$$\min_{f\in\mathcal{H}_K,\alpha} \|f\|_{\mathcal{H}_K}^2 + \alpha \left(\left[n^{-1}\sum_{i=1}^{n} V(f, z_i) \right] - \kappa \right), \tag{6.3}$$

其中 $\alpha \geq 0$，最优 $\alpha = \alpha_0$。根据最优性的KKT条件

$$\alpha_0 \left(\left[n^{-1}\sum_{i=1}^{n} V(f_0, z_i) \right] - \kappa \right) = 0.$$

如果 $\alpha_0 = 0$，则$\left[n^{-1}\sum_{i=1}^{n} V(f_0, z_i) \right] < \kappa$，我们可以将方程（6.3）重写为最小化

$$\min_{f\in\mathcal{H}_K} \|f\|_{\mathcal{H}_K}^2,$$

这对应于（P1），其中 $\lambda = \infty$，最小值为 f_0。如果 $\alpha_0 > 0$，则$\left[n^{-1}\sum_{i=1}^{n} V(f_0, z_i) \right] = \kappa$并且我们可以将方程(6.3)重写为以下等价的优化过程

$$(P3) \qquad \min_{f\in\mathcal{H}_K} \|f\|_{\mathcal{H}_K}^2 + \alpha_0 \left(\left[n^{-1}\sum_{i=1}^{n} V(f, z_i) \right] - \kappa \right),$$

$$(P3) \qquad \min_{f\in\mathcal{H}_K} \|f\|_{\mathcal{H}_K}^2 + \alpha_0 \left[n^{-1}\sum_{i=1}^{n} V(f, z_i) \right],$$

$$(P3) \qquad \min_{f\in\mathcal{H}_K} \left[n^{-1}\sum_{i=1}^{n} V(f, z_i) \right] + \frac{1}{\alpha_0}\|f\|_{\mathcal{H}_K}^2,$$

对应于(P1)的 $\lambda = 1/\alpha_0$和最小值为 f_0. □

凸优化综述*

凸优化中的概念，如Karush-Kuhn-Tucker (KKT)条件，在本讲座的前几节中使用过。 在本节中，我们对这些条件进行了简要介绍和推导。

定义。 集合$\mathcal{X} \in \mathrm{IR}^n$是凸的，如果

$$\forall x_1, x_2 \in \mathcal{X}, \quad \forall \lambda \in [0,1], \lambda x_1 + (1-\lambda) x_2 \in \mathcal{X}.$$

个集合是凸的，如果给定集合中的任意两点，连接它们的线段完全位于集合内部。

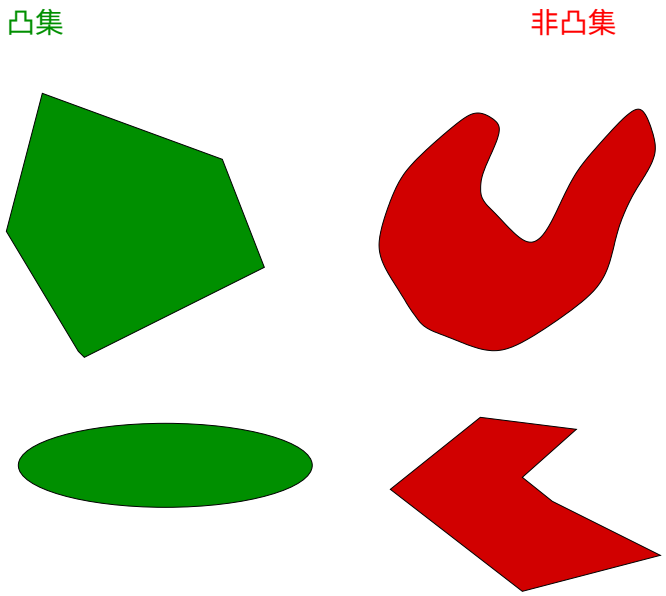


图1. 凸集和非凸集的示例在IR^2中。

定义。 一个函数 f：$\mathrm{IR}^n \to \mathrm{IR}$如果满足以下条件

，则为凸函数：(1) 对于函数 f的定义域中的任意 x_1和 x_2，

对于任意 $\lambda \in [0,1]$，有$f(\lambda x_1 + (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda$

(2) 连接两个点 $f(x_1)$和 $f(x_2)$的线段完全位于或位于函数 f的上方。

(3) 位于或位于函数 f上方的点的集合是凸的。

如果我们用“在或以上”替换“以上”，或者用“$\leq$”替换 $<$，那么一个函数就是严格凸的。

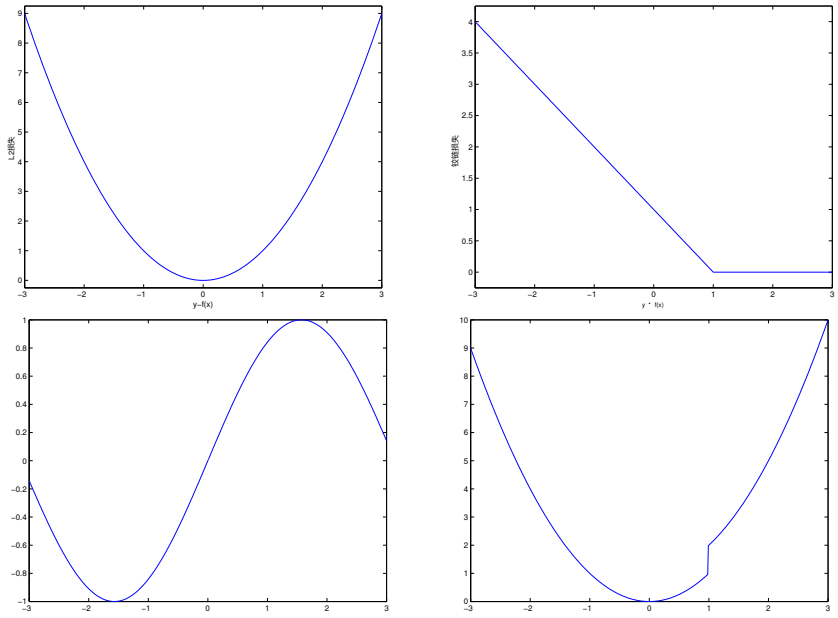


图2. 前两个图是凸函数。. 第一个函数是严格凸的。 底部的图是非凸函数。

定义。 如果存在$\varepsilon > 0$，使得对于所有 $\| \mathbf{x} - \mathbf{x}^* \| \leq \varepsilon$，都有 $f(\mathbf{x}^*) < f(\mathbf{x})$，那么点 $\mathbf{x}^*$被称为函数 f的局部最小值。

定义。 如果对于所有可行的 $\mathbf{x}$，都有 $f(\mathbf{x}^*) < f(\mathbf{x})$，那么点 $\mathbf{x}^*$被称为函数 f的全局最小值。

无约束的凸函数（定义域为整个实数集IR^n）很容易最小化。 凸函数几乎处处可微。方向导数总是存在。 如果我们在局部移动无法改善我们的解决方案，那么我们已经达到最优解。 如果我们找不到一个方向来改善我们的解决方案，那么我们已经达到最优解。

凸函数在凸集上（凸域）也很容易最小化。
如果集合和函数都是凸的，如果我们找不到一个方向可以使函数减小，那么我们就完成了。 局部最优解是全局最优解。

例子。 线性规划问题总是一个凸问题

$$\begin{aligned} \min_{c} \quad & \langle c, x \rangle \\ \text{约束条件：} \quad & Ax = b \\ & Cx \leq d。 \end{aligned}$$

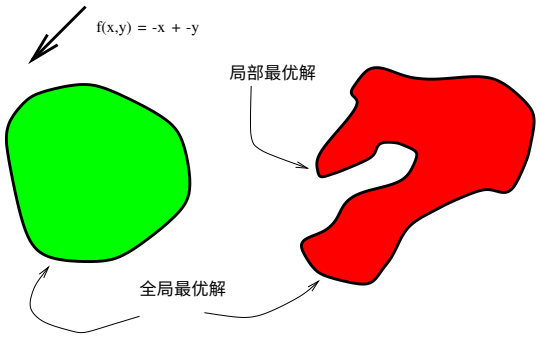


图3。 优化一个凸函数在凸和非凸集合上。. 在左边的例子中，集合是凸的，函数是凸的，所以局部最小值对应于全局最小值。 在右边的例子中，集合是非凸的，函数是凸的，可以找到不是全局最小值的局部最小值。

例子。二次规划是一个凸问题，当且仅当矩阵 Q是半正定的

$$\begin{aligned} \text{最小化} \quad & x'Qx + \langle c,\, x\rangle \\ \text{约束条件为:} \quad & Ax = b \\ & Cx \leq d. \end{aligned}$$

定义。 下面的约束优化问题 P将被称为原始问题

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{约束条件为:} \quad & g_i(\mathbf{x}) \geq 0 \quad i= 1, \ldots, m \\ & h_i(\mathbf{x}) = 0 \quad i = 1, \ldots, n \\ & x \in \mathcal{X}. \end{aligned}$$

在这里， f是我们的目标函数， g_i是不等式约束， h_i是等式约束， $\mathcal{X}$是某个集合。

定义。 我们定义一个拉格朗日对偶问题 D:

$$\begin{aligned} \max \quad & \Theta(\mathbf{u}, \mathbf{v}) \\ \text{subject to :} \quad & \mathbf{u} \geq 0 \end{aligned}$$

其中 $\Theta(\mathbf{u}, \mathbf{v}) := \inf \left\{ f(\mathbf{x}) - \sum_{i=1}^{n} u_i g_i(\mathbf{x}) - \sum_{j=1}^{n} v_i h_i(\mathbf{x}) : x \in \mathcal{X} \right\}$.

定理(弱对偶性). 假设 x 是问题 P的可行解。. 假设 $\mathbf{u}, \mathbf{v}$是问题 D的可行解。那么对于所有 $\mathbf{u} \geq 0$

$$f(\mathbf{x}) \geq \Theta(\mathbf{u}, \mathbf{v}).$$

证明。

$$\begin{aligned}\Theta(\mathbf{u},\mathbf{v}) &= \inf\left\{f(\mathbf{y})-\sum_{i=1}^{m}u_i g_i(\mathbf{y})-\sum_{j=1}^{n}v_i h_i(\mathbf{y}) : y\in\mathcal{X}\right\}\\ &\leq f(\mathbf{x})-\sum_{i=1}^{m}u_i g_i(\mathbf{x})-\sum_{i=1}^{n}v_i h_i(\mathbf{x})\\ &\leq f(\mathbf{x}).\end{aligned}$$

弱对偶性表明，对于 P的每个可行解，其费用至少与 D的每个可行解一样昂贵。 这是对偶性的一个非常普遍的性质，我们没有依赖于任何凸性假设来证明它。

定义。当原始问题和对偶问题的最优解等价时，强对偶性成立 $Opt(P) = Opt(D)$。

如果强对偶性不成立，我们可能会出现对偶间隙的情况。强对偶性非常有用，因为通常意味着我们可以从计算上更方便的对偶或原始问题中解决，并且通常可以从一个问题的解得到另一个问题的解。

命题。如果目标函数 f是凸的，并且可行域是凸的，在轻微的技术条件下，我们有强对偶性。

现在我们来看看所谓的拉格朗日函数的鞍点。 我们将拉格朗日函数定义为对偶问题。

$$L(\mathbf{x},\mathbf{u},\mathbf{v}) = f(\mathbf{x})-\sum_{i=1}^{m}\mathbf{u}_i\mathbf{g}_i(\mathbf{x})-\sum_{j=1}^{n}\mathbf{v}_j\mathbf{h}_j(\mathbf{x})。$$

对于 P和 D的可行解集合$(\mathbf{x}^*,\mathbf{u}^*,\mathbf{v}^*)$被称为拉格朗日函数的鞍点，如果

如果 $\mathbf{u}$和 $\mathbf{v}$固定在 $\mathbf{u}^*$和 $\mathbf{v}^*$，并且 $\mathbf{u}^*$和 $\mathbf{v}^*$最大化L，那么$\mathbf{x}$在$\mathbf{x}^*$处最小化 L。

定义。 点 $(\mathbf{x}^*,\mathbf{u}^*,\mathbf{v}^*)$满足$Karush\,Kuh\mathbf{n}\ Tucke r\ (KKT)$条件或者是$KKT$点，如果它们对于 P和 D是可行的，并且$\nabla f(\mathbf{x}^*)-\nabla\mathbf{g}(\mathbf{x}^*)'\mathbf{u}^*-\nabla\mathbf{h}(\mathbf{x}^*)'\mathbf{v}\ =0\mathbf{u}^*\mathbf{g}(\mathbf{x}^*)=0$。

在一个凸、可微的问题中，满足KKT条件的点等价于拉格朗日函数的鞍点，只需满足一些次要的技术条件。

讲座 7
支持向量机

支持向量机已经在许多应用中被使用，并且是最流行的机器学习算法之一。我们将从两个角度推导出支持向量机算法：Tikhonov正则化和更常见的几何角度。

7.1. 从Tikhonov正则化到支持向量机

我们从Tikhonov正则化开始

$$\min_{f\in\mathcal{H}}\left[n^{-1}\sum_{i=1}^{n}V(f,\,z_i)+\lambda\Omega(f)\right]$$

然后将正则化函数设置为RKHS范数

$$\Omega(f)=\|f\|_{\mathcal{H}_K}^2$$

并使用合页损失函数

$$n^{-1}\sum_{i=1}^{n}V(f,\,z_i):=n^{-1}\sum_{i=1}^{n}(1-y_if(x_i))_+,$$

其中 $(\quad k)_+:=\max(k,0)$.

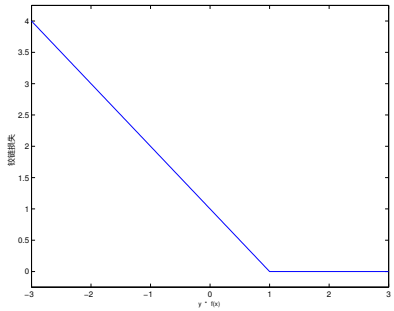


图1. 合页损失函数.

由此产生的优化问题是

$$\min_{f\in\mathcal{H}_K}\left[n^{-1}\sum_{i=1}^{n}(1-y_if(x_i))_+ + \lambda\|f\|^2_{\mathcal{H}_K}\right], \tag{7.1}$$

在 $(1-y_if(x_i))=0$ 处不可微，因此我们引入松弛变量并写出以下约束优化问题：
min

$$\begin{aligned} &\min_{f\in\mathcal{H}_K} && n^{-1}\sum_{i=1}^{n}\xi_i+\lambda||f||_K^2 \\ &\text{subject to :} && y_if(x_i)\geq 1-\xi_i \qquad i=1,\ldots,n \\ &&& \xi_i\geq 0 \qquad\qquad i=1,\ldots,n. \end{aligned}$$

通过表示定理，我们可以将上述受限优化问题重新写成一个受限二次规划问题min

$$\begin{aligned} &\min_{c\in\mathbb{R}^n} && n^{-1}\sum_{i=1}^{n}\xi_i+\lambda c^TKc \\ &\text{subject to :} && y_i\sum_{j=1}^{n}c_jK(x_i,x_j)\geq 1-\xi_i \quad i=1,\ldots,n \\ &&& \xi_i\geq 0 \qquad\qquad i=1,\ldots,n. \end{aligned}$$

SVM 包含一个未正则化的偏置项 b，因此表示定理导致一个函数$f(x)=$

$$\sum_{i=1}^{n}c_iK(x,x_i)+b.$$

将这个形式代入上述受限二次问题中得到了“原始”SVMmin

$$\begin{aligned} &\min_{\mathbf{c}\in\mathbb{R}^n,\xi\in\mathbb{R}^n} && n^{-1}\sum_{i=1}^{n}\xi_i+\lambda\mathbf{c}^TK\mathbf{c} \\ &\text{约束条件：} && y_i\left(\sum_{j=1}^{n}c_jK(x_i,x_j)+b\right)\geq 1-\xi_i \quad i=1,\ldots,n \\ &&& \xi_i\geq 0 \qquad\qquad i=1,\ldots,n. \end{aligned}$$

现在我们使用拉格朗日乘子技术推导出 Wolfe 对偶二次规划问题：

$$\begin{aligned} L(\mathbf{c},\xi,b,\alpha,\zeta) &= n^{-1}\sum_{i=1}^{n}\xi_i+\lambda\mathbf{c}^TK\mathbf{c} \\ &\quad -\sum_{i=1}^{n}\alpha_i\left(y_i\left\{\sum_{j=1}^{n}c_jK(x_i,x_j)+b\right\}-1+\xi_i\right) \\ &\quad -\sum_{i=1}^{n}\zeta_i\xi_i. \end{aligned}$$

我们希望最小化 L关于 $\mathbf{c}$, b, 和 ξ的值，并且最大化 L关于 α和 ζ的值，同时满足原始问题和非负约束条件对于 α和 ζ. 我们首先通过偏导数消除 b和 ξ:∂L

$$\begin{aligned} \frac{}{\partial b}=0 &\implies \sum_{i=1}^{n}\alpha_iy_i=0 \\ \frac{\partial L}{\partial\xi_i}=0 &\implies \frac{1}{n}-\alpha_i-\zeta_i=0\Longrightarrow 0\leq\alpha_i\leq\frac{1}{n}. \end{aligned}$$

上述两个条件将成为必须在最优解时满足的约束条件。 这导致了一个简化的拉格朗日函数:

$$L^R(\mathbf{c}, \alpha) = \lambda \mathbf{c}^T K \mathbf{c} - \sum_{i=1}^{n} \alpha_i \left(y_i \sum_{j=1}^{n} c_j K(x_i, x_j) - 1 \right).$$

我们现在消除 $\mathbf{c}$

$$\frac{\partial L^R}{\partial \mathbf{c}} = 0 \implies 2\lambda K\mathbf{c} - KY\alpha = 0 \implies c_i = \frac{\alpha_i y_i}{2\lambda},$$

对角元素是 y i;$Y\alpha$是一个向量其第$_i$个元素是 $\alpha_i y_i$. 将我们对 $\mathbf{c}$的表达式代入，我们得到以下“对偶”问题：最大化

$$\begin{aligned} \max_{\alpha \in \mathrm{IR}^{\mathbf{n}}} \quad & \textstyle\sum_{i=1}^{\mathbf{n}} \alpha_i - \frac{1}{4\lambda} \alpha^T Q \alpha \\ \text{约束条件：} \quad & \textstyle\sum_{i=1}^{\mathbf{n}} y_i \alpha_i = 0 \\ & 0 \le \alpha_i \le \frac{1}{n} \qquad i = 1, \ldots, n, \end{aligned}$$

其中 Q 是由以下矩阵定义的

$$Q = \mathbf{y} K \mathbf{y^T} \Longleftrightarrow Q_{ij} = y_i y_j K(x_i, x_j).$$

在大多数支持向量机文献中，正则化参数不是通过正则化参数 λ来控制的，而是通过参数 C来控制的，这个参数是通过以下关系定义的

$$C = \frac{1}{2\lambda n}.$$

使用这个定义（在将目标函数乘以常数 $\frac{1}{2n}$ ，之后），基本的正则化问题变成了

$$\min_{f \in \mathcal{H}_K} \; C \sum_{i=1}^{n} V(y_i, f(\mathbf{x}_i)) + \frac{1}{2} ||f||_K^2.$$

和 λ一样，参数 C也控制着分类准确性和函数范数之间的权衡。 原始问题和对偶问题分别为：$\min$

$$\begin{aligned} \min_{\mathbf{c} \in \mathrm{IR}^n, \xi \in \mathrm{IR}^n} \quad & C \textstyle\sum_{i=1}^{\mathbf{n}} \xi_i + \frac{1}{2} \mathbf{c}^T K \mathbf{c} \\ \text{约束条件：} \quad & y_i \left(\textstyle\sum_{j=1}^{\mathbf{n}} c_j K(x_i, x_j) + b \right) \ge 1 - \xi_i \quad i = 1, \ldots, n \\ & \xi_i \ge 0 \qquad\qquad\qquad\qquad i = 1, \ldots, n \end{aligned}$$

$$\begin{aligned} \max_{\alpha \in \mathrm{IR}^{\mathbf{n}}} \quad & \textstyle\sum_{i=1}^{\mathbf{n}} \alpha_i - \frac{1}{2} \alpha^T Q \alpha \\ \text{约束条件：} \quad & \textstyle\sum_{i=1}^{\mathbf{n}} y_i \alpha_i = 0 \\ & 0 \le \alpha_i \le C \qquad i = 1, \ldots, n. \end{aligned}$$

7.2. 从几何角度看支持向量机

发展支持向量机数学的“传统”方法是从分离超平面和间隔的概念开始。 通常在线性空间中发展理论，从感知器的概念开始，感知器是一个线性超平面，用于分离正例和负例。 将间隔定义为超平面到最近样本的距离，基本观察是

直观上，我们期望具有较大间隔的超平面比具有较小间隔的超平面更好地泛化。

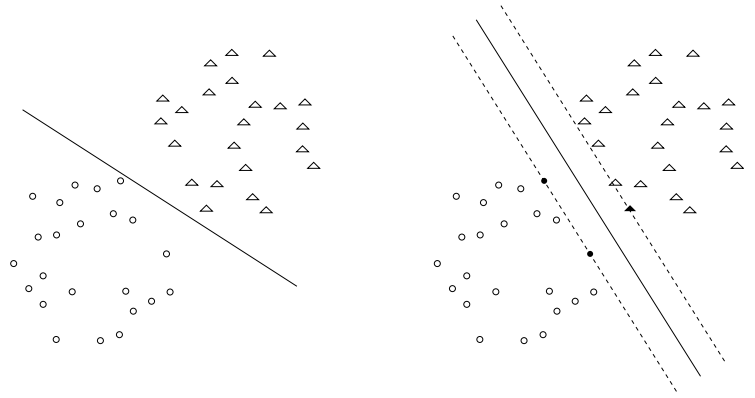


图**2.** 两个超平面(a)和(b)完全分离数据。. 然而，超平面(b)具有更大的间隔，直观上预期在新的观察中更准确。

我们用 $\mathbf{w}$表示我们的超平面，并且我们将通过函数(7.2)对一个新的点 $\mathbf{x}$进行分类。

$$f(x) = \text{sign } [\langle \mathbf{w}, \mathbf{x} \rangle] .$$

给定一个分离的超平面 $\mathbf{w}$，我们让 $\mathbf{x}$成为离 $\mathbf{w}$最近的数据点，并且我们让$\mathbf{x}^{\mathbf{w}}$成为离 $\mathbf{w}$最近的 x上的唯一点。 显然，找到一个最大间隔 $\mathbf{w}$等价于最大化 $||\mathbf{x} - \mathbf{x}^{\mathbf{w}}||$ 。 所以对于一些 k（假设$k > 0$ 为了方便起见），

$$\begin{aligned} \langle \mathbf{w}, \mathbf{x} \rangle &= k \\ \langle \mathbf{w}, \mathbf{x}^{\mathbf{w}} \rangle &= 0 \\ \langle \mathbf{w}, (\mathbf{x} - \mathbf{x}^{\mathbf{w}}) \rangle &= k。 \end{aligned}$$

注意到向量 $\mathbf{x} - \mathbf{x}^{\mathbf{w}}$与法向量 w平行，

$$\begin{aligned} \langle \mathbf{w}, (\mathbf{x} - \mathbf{x}^{\mathbf{w}}) \rangle &= \left\langle \mathbf{w}, \left(\frac{||\mathbf{x} - \mathbf{x}^{\mathbf{w}}||}{||\mathbf{w}||} \mathbf{w} \right) \right\rangle \\ &= ||\mathbf{w}||^2 \frac{||\mathbf{x} - \mathbf{x}^{\mathbf{w}}||}{||\mathbf{w}||} \\ &= ||\mathbf{w}||\ ||\mathbf{x} - \mathbf{x}^{\mathbf{w}}|| \\ k &= ||\mathbf{w}||\ ||(\mathbf{x} - \mathbf{x}^{\mathbf{w}})|| \\ \frac{k}{||\mathbf{w}||} &= ||\mathbf{x} - \mathbf{x}^{\mathbf{w}}||. \end{aligned}$$

k是一个“无关参数”，为了不失一般性，我们将 k固定为1，并且可以看到，最大化 $||\mathbf{x} - \mathbf{x}^{\mathbf{w}}||$等价于最大化 $\frac{1}{||w||}$，进而最小化 $||w||$ or $||w||^2$。 现在我们可以将间隔定义为超平面 $\langle \mathbf{w}, \mathbf{x} \rangle = 0$ 和 $\langle \mathbf{w}, \mathbf{x} \rangle = 1$ 之间的距离。

所以如果数据是线性可分的情况，并且超平面通过原点，那么最大间隔超平面就是那个

$$\begin{array}{ll} \min\limits_{\mathbf{w}\in \mathrm{IR}^n} & ||\mathbf{w}||^2 \\ \text{满足条件：} & y_i\langle \mathbf{w}, \mathbf{x}_i\rangle \geq 1 \quad i = 1, \ldots, n. \end{array}$$

Vapnik引入的SVM包括一个未正则化的偏置项 b，通过以下形式的函数进行分类：

$$f(x) = \mathrm{sign}[\langle \mathbf{w}, \mathbf{x}\rangle + b].$$

此外，我们还需要处理线性不可分的数据集，因此我们引入松弛变量 ξ_i，就像以前一样。 我们仍然可以将间隔定义为超平面 $\langle \mathbf{w}, \mathbf{x}\rangle = 0$和 $\langle \mathbf{w}, \mathbf{x}\rangle = 1$之间的距离，但几何直观不再那么清晰或有说服力。

加上偏差项和松弛变量后，原始SVM问题变为

$$\begin{array}{lll} \min\limits_{\mathbf{w}\in \mathrm{IR}^n, b\in \mathrm{IR}} & C\sum_{i=1}^{n} \xi_i + \frac{1}{2}||\mathbf{w}||^2 & \\ \text{约束条件：} & y_i\left(\langle \mathbf{w}, \mathbf{x}\rangle + b\right) \geq 1 - \xi_i & i = 1, \ldots, n \\ & \xi_i \geq 0 & i = 1, \ldots, n. \end{array}$$

使用拉格朗日乘子，我们可以从前一节中推导出相同的对偶形式。

从历史上看，大多数发展都是从几何形式开始的，推导出一个与我们上面推导出的对偶问题相同的对偶程序，然后才观察到这个对偶程序只需要点积，并且这些点积可以用核函数来替代。 在线性可分的情况下，我们还可以将分离超平面推导为一个与正负类中最近两点连接向量平行的向量，通过这个向量的垂直平分线。 这就是Vapnik在20世纪70年代提出的“肖像法”，最近被Keerthi重新发现（包括非可分扩展）

7.3. 最优性条件

原始问题和对偶问题都是可行的凸二次规划。 因此，它们都有最优解，并且原始问题和对偶问题的最优解具有相同的目标值。

我们使用（现在重新参数化的）拉格朗日乘子法从原始问题推导出对偶问题：

$$\begin{aligned} L(\mathbf{c}, \xi, b, \alpha, \zeta) \quad = \quad & C\sum_{i=1}^{n} \xi_i + \mathbf{c}^T K\mathbf{c} \\ & - \sum_{i=1}^{n} \alpha_i \left(y_i \left\{ \sum_{j=1}^{n} c_j K(x_i, x_j) + b \right\} - 1 + \xi_i \right) \\ & - \sum_{i=1}^{n} \zeta_i \xi_i. \end{aligned}$$

现在我们考虑与原始约束相关的对偶变量：

$$\begin{aligned}\alpha_i &\implies y_i\left\{\sum_{j=1}^{n} c_j K(x_i, x_j) + b\right\} - 1 + \xi_i \\ \zeta_i &\implies \xi_i \geq 0.\end{aligned}$$

互补松弛条件告诉我们，在最优性时，要么原始不等式满足等式，要么对偶变量为零。 换句话说，如果 $\mathbf{c}$, ξ, b, α和 ζ是原始问题和对偶问题的最优解，则

$$\begin{aligned}\alpha_i\left(y_i\left\{\sum_{j=1}^{n} c_j K(x_i, x_j) + b\right\} - 1 + \xi_i\right) &= 0 \\ \zeta_i \xi_i &= 0\end{aligned}$$

所有最优解必须满足：

$$\begin{aligned}\sum_{j=1}^{n} c_j K(x_i, x_j) - \sum_{j=1}^{n} y_i \alpha_j K(x_i, x_j) = 0 \qquad & i = 1, \ldots, n \\ \sum_{i=1}^{n} \alpha_i y_i = 0 \qquad & \\ C - \alpha_i - \zeta_i = 0 \qquad & i = 1, \ldots, n \\ y_i\left(\sum_{j=1}^{n} y_j \alpha_j K(x_i, x_j) + b\right) - 1 + \xi_i \geq 0 \qquad & i = 1, \ldots, n \\ \alpha_i\left[y_i\left(\sum_{j=1}^{n} y_j \alpha_j K(x_i, x_j) + b\right) - 1 + \xi_i\right] = 0 \qquad & i = 1, \ldots, n \\ \zeta_i \xi_i = 0 \qquad & i = 1, \ldots, n \\ \xi_i, \alpha_i, \zeta_i \geq 0 \qquad & i = 1, \ldots, n\text{以上}\end{aligned}$$

最优性条件既是必要条件也是充分条件。 如果我们有 $\mathbf{c}$,ξ, b, α和 ζ满足上述条件，我们知道它们代表了原始问题和对偶问题的最优解。 这些最优性条件也被称为Karush-Kuhn-Tucker（KKT）条件。

假设我们有最优的 α_i。 还假设（在实践中“总是”发生）存在一个满足$0 < \alpha\ i < C$的$_i$。 那么$\alpha_i < C \implies \zeta_i > 0 \implies \xi_i = 0$

$$\begin{aligned}&\implies y_i\left(\sum_{j=1}^{n} y_j \alpha_j K(x_i, x_j) + b\right) - 1 = 0 \\ &\implies b = y_i - \sum_{j=1}^{n} y_j \alpha_j K(x_i, x_j)\end{aligned}$$

因此，如果我们知道最优的 α，我们可以确定 b。

将我们的分类函数 $f(x)$定义为

$$f(x) = \sum_{i=1}^{\ell} y_i \alpha_i K(x, x_i) + b,$$

我们可以推导出"简化"的最优条件。 例如，考虑一个 i，使得

$$\begin{aligned} y_i f(x_i) < 1 \quad &\Longrightarrow \quad \xi_i > 0 \\ &\Longrightarrow \quad \zeta_i = 0 \\ &\Longrightarrow \quad \alpha_i = C. \end{aligned}$$

反过来，假设 $\alpha_i = C$:

$$\begin{aligned} \alpha_i = C \quad &\Longrightarrow \quad y_i f(x_i) - 1 + \xi_i = 0 \\ &\Longrightarrow \quad y_i f(x_i) \leq 1. \end{aligned}$$

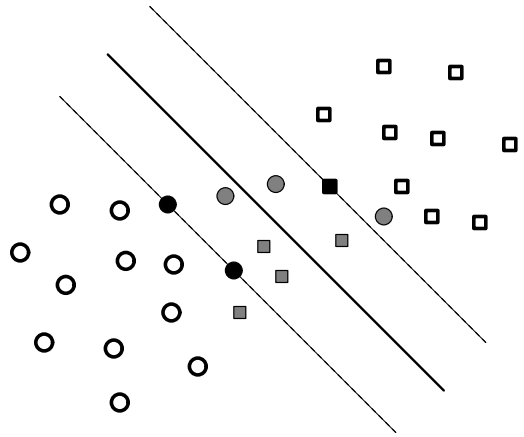


图3. 减少的最优条件的几何解释。
开放的方块和圆圈对应于 $\alpha_i = 0$. 深色圆圈和方块对应于 $y_i f(x_i) = 1$ 和 $\alpha_i \leq C$ ，这些是边界上的样本。 灰色圆圈和方块对应于 $y_i f(x_i) < 1$ 和 $\alpha_i = C$.

7.4. 解决SVM优化问题

我们的计划是解决对偶问题，找到 α的值，并使用它来找到 b和我们的函数 f。 对偶问题比原始问题更容易解决。 它具有简单的箱约束和单个不等式约束，更好的是，我们将看到该问题可以分解为一系列较小的问题。

我们可以使用标准软件解决QPs问题。 有许多可用的代码。 主要问题是 Q矩阵是密集的，是 $n \times n$，所以我们无法写下来。 标准的QP软件需要 Q矩阵，因此不适用于大问题。

为了解决这个内存问题，我们将数据集分成一个工作集W和剩余的点 R。 我们可以将对偶问题重写为：最大化

$$\max_{\alpha_W \in \mathrm{IR}^{|W|},\, \alpha_R \in \mathrm{IR}^{|R|}} \quad \sum_{i \in W}^{n} \alpha_i + \sum_{i \in R} \alpha_i - \frac{1}{2}[\alpha_{\mathbf{W}}\ \alpha_{\mathbf{R}}] \begin{bmatrix} Q_{WW} & Q_{WR} \\ Q_{RW} & Q_{RR} \end{bmatrix} \begin{bmatrix} \alpha_{\mathbf{W}} \\ \alpha_{\mathbf{R}} \end{bmatrix}$$

约束条件：

$$\sum_{i \in W} y_i \alpha_i + \sum_{i \in R} y_i \alpha_i = 0$$
$$0 \le \alpha_i \le C,\ \forall i.$$

假设我们有一个可行解 α. 通过将
$\alpha_{\mathbf{W}}$视为变量，将 $\alpha_{\mathbf{R}}$视为常数，我们可以得到一个更好的解。 我们可以解决简化的对偶问题：

$$\max_{\alpha_W \in \mathrm{IR}^{|W|}} \quad (\mathbf{1} - Q_{WR}\alpha_{\mathbf{R}})\alpha_W - \frac{1}{2}\alpha_{\mathbf{W}} Q_{WW}\alpha_{\mathbf{W}}$$

受限于:

$$\sum_{i \in W} y_i \alpha_i = -\sum_{i \in R} y_i \alpha_i$$
$$0 \le \alpha_i \le C, \quad \forall i \in W。$$

简化问题是固定大小的，可以使用标准QP代码解决。 收敛证明很困难，但这种方法在实践中似乎总是收敛到最优解。

在分解中选择工作集是一个重要问题。有许多不同的方法。 基本思想是检查不在工作集中的点，找到违反简化最优性条件的点，并将它们添加到工作集中。删除工作集中远离违反最优性条件的点。

讲座 8

正则化逻辑回归

SVM的一个缺点是该方法不会明确输出标签的概率或似然，而是输出一个实数值，其大小应与概率$P(y = \pm 1|x) \propto y\, f(x)$成单调关系。

可以通过使用基于逻辑回归或二元回归的损失函数来解决这个问题。 逻辑回归的主要思想是通过函数 $f(x)$来建模对数似然比。

$$f(x) = \log\left(\frac{P(y = 1|x)}{P(y = -1|x)}\right).$$

由于 $P(y{=}\,1|x)$是一个伯努利随机变量，我们可以将上述方程重写为

$$\begin{aligned} f(x) &= \log\left(\frac{P(y = 1|x)}{P(y = -1|x)}\right) \\ &= \log\left(\frac{P(y = 1|x)}{1 - P(y = 1|x)}\right) \end{aligned}$$

这意味着

$$\begin{aligned} P(y = 1|x) &= \frac{1}{1 + \exp(f(x))} \\ P(y = -1|x) &= \frac{1}{1 + \exp(-f(x))} \\ P(y = \pm 1|x) &= \frac{1}{1 + \exp(y\, f(x))}. \end{aligned}$$

给定一个数据集 $D = \{(x_i, y_i)\}^{ni}{}_{=1}$和一个函数类 $f{\in}\,\mathcal{H}$最大似-然估计 (MLE) 是使观测到数据集 D的似然函数最大化的函数

$$f^*_{\mathrm{MLE}} := \arg\max_{f\in\mathcal{H}} [P(D|f)] = \arg\max_{f\in\mathcal{H}} \left[\prod_{i=1}^{n} \frac{1}{1 + \exp(y_i\, f(x_i))}\right].$$

就像经验风险最小化的情况一样，MLE估计可能会过拟合数据
因为没有平滑或正则化项。一种经典的施加方式

在这个背景下，通过对函数 $f\in\mathcal{H}$进行先验设定来实现平滑性

$$P(f) \propto e^{-\|f\|^2_{\mathcal{H}_K}}.$$

给定先验和似然，我们可以使用贝叶斯规则计算后验分布 $P(f|D)$

$$P(f|D) = \frac{P(D|f)\,P(f)}{P(D)}.$$

如果我们将先验和似然函数代入贝叶斯规则，我们可以计算出最大后验概率（MAP）估计量

$$\begin{aligned} f^*_{\text{MAP}} &:= \arg\max_{f\in\mathcal{H}} \left[\frac{P(D|f)P(f)}{P(D)}\right] \\ &= \arg\max_{f\in\mathcal{H}} \left[\frac{\prod_{i=1}^n \frac{1}{1+\exp(y_i\,f(x_i))}\; e^{-\|f\|^2_{\mathcal{H}_K}}}{P(D)}\right] \\ &= \arg\max_{f\in\mathcal{H}} \left[\sum_{i=1}^n \log\left(\frac{1}{1+\exp(y_i\,f(x_i))}\right) - \|f\|^2_{\mathcal{H}_K}\right]. \end{aligned}$$

通过一些简单的代数运算，上述MAP估计量可以以Tikhonov正则化的形式重写

$$f^*_{\text{MAP}} = \arg\min_{f\in\mathcal{H}_K} \left[n^{-1}\sum_{i=1}^n \log(1+\exp(-y_i\,f(x_i))) + \lambda\|f\|^2_{\mathcal{H}_K}\right],$$

其中 λ是正则化参数。 根据表示定理，上述方程有以下形式的解

$$f^*(x) = \sum_{i=1}^n c_i K(x,\, x_i)。$$

根据上述表示定理，我们可以通过以下优化问题来解决变量 c_i

$$\min_{\mathbf{c}\in\mathrm{IR}^n} \left[n^{-1}\sum_{i=1}^n \log(1+\exp(-y_i\,(\mathbf{c}^T\mathbf{K})_i)) + \lambda\mathbf{c}^T\mathbf{K}\mathbf{c}\right],$$

其中$(\mathbf{c}^T\mathbf{K})_i$是向量 $\mathbf{c}^T\mathbf{K}$的第i个元素。 这个优化问题是凸的且可微的，因此解决 $\mathbf{c}$ 的经典方法是使用牛顿-拉夫逊方法。

8.1. 牛顿-拉夫逊

牛顿-拉夫逊方法最初用于求解多项式的根，将其应用于优化问题相对简单。 我们首先描述一维情况下的牛顿-拉夫逊方法，即优化问题是关于一个变量的。 然后我们描述多元形式，并将其应用于逻辑回归的优化问题。

- 牛顿法求根：牛顿法主要用于求解多项式的根。 该方法由牛顿在1669年左右提出

拉普森在1690年对该方法进行了改进，因此称为牛顿-拉普森方法。 给定一个多项式 $f(x)$，在点$x = x_0 + \varepsilon$附近的泰勒级数展开式为

$$f(x_0 + \varepsilon) = f(x_0) + f'(x_0)\varepsilon + \frac{1}{2}f''(x_0)\varepsilon^2 + ...$$

将展开式截断至一阶项后，得到

$$f(x_0 + \varepsilon) \approx f(x_0) + f'(x_0)\varepsilon。$$

从上面的表达式中，我们可以估计需要的偏移量 ε来接近根$(x : f(x) = 0)$，从初始猜测 x_0开始。 这是通过设置 $f(x_0 + \varepsilon) = 0$并解出 ε来完成的。

$$\begin{aligned} 0 &= f(x_0 + \varepsilon) \\ 0 &\approx f(x_0) + f'(x_0)\varepsilon \\ -f(x_0) &\approx f'(x_0)\varepsilon \\ \epsilon_0 &\approx -\frac{f(x_0)}{f'(x_0)}. \end{aligned}$$

这是根位置的一阶或线性调整。 这可以通过设置 $x_1 = x_0 + \varepsilon_0$，计算一个新的 ε_1，然后迭代直到收敛来转化为迭代过程：

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$

- 牛顿-拉夫逊方法作为标量优化方法：给定一个凸最小化问题$\min$

$$\min_{x \in [a,b]} g(x),$$

其中 $g(x)$是一个凸函数。 函数 $g(x)$的极值点将出现在一个值of $x_m$such that $g'(x_m) = 0$，由于函数是凸的，这个极值点将是一个最小值。 如果 $g(x)$是一个多项式，那么 $g'(x)$也是一个多项式我们可以将牛顿法应用于 $g'(x)$进行根查找。 如果 $g(x)$不是一个多项式，那么我们将根查找方法应用于多项式approximation of $g(x)$. 我们现在描述涉及的步骤。

(1) Taylor展开 $g(x)$：对 $g(x)$进行截断的Taylor展开结果是对 $g(x)$的二阶多项式近似

$$g(x) \approx g(x_0) + g'(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^2 g''(x_0).$$

(2) 将导数设为零：对Taylor展开进行求导并将其设为零

$$\frac{dg}{dx} = f(x) = g'(x_0) + g''(x_0)(x - x_0) = 0.$$

这使我们面临一个寻找根的问题，即寻找 $f(x)$的根
我们使用牛顿法来寻找根。

(3) 更新规则：更新规则简化为

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{g'(x_n)}{g''(x_n)}.$$

上述过程中的一个关键点是 $g(x)$ 的凸性。 为了确保该过程收敛，优化域内的二阶导数 $g''(x)$ 必须为正，即区间 $[a, b]$。 凸性保证了 $g(x)$ 的这一点。

- 牛顿-拉夫逊方法作为向量的优化方法：给定一个凸最小化问题$\min$

$$\min_{\mathbf{x}\in\mathcal{X}} g(\mathbf{x}),$$

其中$\mathcal{X}\subseteq \mathrm{IR}^n$是凸集， $g(\mathbf{x})$ 是凸函数。 我们按照标量情况的逻辑进行，只是使用向量微积分。

(1) Taylor展开 $g(\mathbf{x})$：对 $g(\mathbf{x})$进行截断的Taylor展开结果是对 $g(\mathbf{x})$的二阶多项式近似

$$g(\mathbf{x}) \approx g(\mathbf{x}_0) + (\mathbf{x}-\mathbf{x}_0)^T \cdot \nabla g(\mathbf{x}_0) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T \cdot \mathbf{H}(x_0) \cdot (\mathbf{x}-\mathbf{x}_0),$$

其中 $\mathbf{x}$是一个长度为 n的列向量， $\nabla g(\mathbf{x}_0)$是在 $\mathbf{x}_0$处评估的 g的梯度，也是一个长度为 n的列向量， $\mathbf{H}(x_0)$是在 $\mathbf{x}_0$处评估的Hessian矩阵

$$\mathbf{H}_{i,j}(x_0) = \frac{\partial^2 g(\mathbf{x})}{\partial \mathbf{x}^i \partial \mathbf{x}^j}\Big|_{\mathbf{x}_0}, \quad i, j = 1, ..., n.$$

(2) 将导数设为零：对Taylor展开进行求导并将其设为零

$$\nabla g(\mathbf{x}) = \nabla g(\mathbf{x}_0) + \frac{1}{2}\mathbf{H}(\mathbf{x}_0) \cdot (\mathbf{x}-\mathbf{x}_0) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^T \cdot \mathbf{H}(\mathbf{x}_0) = 0,$$

Hessian矩阵是对称的且二次可微的（由于凸性），因此我们可以将上述问题简化为$\nabla g(\mathbf{x}) = \nabla g(\mathbf{x}_0) + \mathbf{H}(\mathbf{x}_0) \cdot (\mathbf{x}-\mathbf{x}_0) = 0.$

这意味着在极小值 $\mathbf{x}^*$处，梯度为零

$$0 = \mathbf{H}(\mathbf{x}_0) \cdot (\mathbf{x}^* - \mathbf{x}_0) + \nabla g(\mathbf{x}_0).$$

(3) 更新规则：解上述线性方程组得到 $\mathbf{x}^*$的值，得到以下更新规则

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \mathbf{H}^{-1}(\mathbf{x}_n) \cdot \nabla g(\mathbf{x}_n)$$

，其中 $-\mathbf{H}^{-1}(\mathbf{x}_n) \cdot \nabla g(\mathbf{x}_n)$被称为牛顿方向。
为了使上述过程收敛到极小值，牛顿方向
必须是下降方向

$$\nabla g^T(\mathbf{x}_n) \cdot (\mathbf{x}_{n+1} - \mathbf{x}_n) < 0.$$

如果海森矩阵是正定的，那么牛顿方向将是下降方向，这是正二阶导数的矩阵类比。 在定义域 $\mathcal{X}$上， $g(\mathbf{x})$的凸性确保了海森矩阵是正定的。
如果函数 $g(\mathbf{x})$是二次的，该过程将在一次迭代中收敛。

- 正则化逻辑回归的牛顿-拉夫逊方法：正则化逻辑回归的优化问题是f^*_{MAP} $= \arg\min$

$$f\in\mathcal{H}_K \left[n^{-1}\sum_{i=1}^{n} \log(1+\exp(-y_i\, f(x_i))) + \lambda \|f\|^2_{\mathcal{H}_K} \right],$$

根据表示定理

$$f^*(x) = \sum_{i=1}^{n} c_i K(x,\, x_i) + b,$$

$\|f\|_{\mathcal{H}_K}$是一个不惩罚常数的半范数，就像SVM的情况一样。
优化问题可以重写为

$$\min_{\mathbf{c}\in\mathrm{IR}^n, b\in\mathrm{IR}} \left[L[\mathbf{c},\, b] = n^{-1}\sum_{i=1}^{n} \log(1+\exp(-y_i\,((\mathbf{c}^T\mathbf{K})_i + b))) + \lambda \mathbf{c}^T\mathbf{K}\mathbf{c} \right],$$

其中$(\mathbf{c}^T\mathbf{K})_i$是向量 $\mathbf{c}^T\mathbf{K}$的第i个元素。

讲座 9

高斯过程回归

高斯过程回归的思想是在函数空间 $\mathcal{H}$上放置一个分布。 例如，考虑一个rkhs $\mathcal{H}_K$，我们希望进行贝叶斯推断。 假设一个带有标准噪声假设的回归模型

$$Y_i = f(X_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2), \quad f \in \mathcal{H}_K.$$

如果我们知道如何在函数空间中放置先验，理论上我们可以进行贝叶斯推断。

9.1. 高斯过程

高斯过程是对函数$f(x)$的概率分布的规范化，其中$f\in \mathcal{H}$且 $x\in \mathcal{X}$由均值函数 μ和协方差函数 $K(\cdot,\cdot)$ 参数化。 这个想法可以非正式地陈述为

$$p(f) \propto \exp\Big(-\frac{1}{2}\|f\|^2_{\mathcal{H}_K}\Big), \text{ 对于所有的函数} f\in \mathcal{H},\ p(f)\geq 0。 \int_{f\in\mathcal{H}} p(f)\,\mathrm{d}f = 1,$$

这里我们使用非正式这个术语，因为$\mathrm{d}f$没有被很好地定义，不清楚$p(f)$的归一化常数是什么，函数空间 $\mathcal{H}$与 $\mathcal{H}_K$的关系也不清楚。 我们将从另一个角度来发展高斯过程，而不是让所有的点都变得清晰明了。 有许多定义和思考高斯过程的方式。 一个标准的表述是，高斯过程是多元高斯分布的无限版本，有两个参数：均值函数 μ对应于均值向量和正定协方差或核函数 K对应于正定协方差矩阵。

定义无限维对象的一种常见方法是通过定义其有限维投影。 这是我们将采用的高斯过程的方法。 将 $\mathcal{X}$中的有限点集 $x_1, ..., x_n$视为一个有限集合。 对于函数 $f \in \mathcal{H}$的高斯过程， $\mathbf{f} = \{f(x_1), ..., f(x_n)\}^T$的概率密度是多元正态分布，其中 $\boldsymbol{\mu} = \{\mu(x_1), ..., \mu(x_n)\}$，协方差 $\mathbf{\Sigma}_{ij} = K(x_i, x_j)$。

$$\mathrm{f} \sim \mathrm{N}\Big(\boldsymbol{\mu}, \mathbf{\Sigma}\Big),$$

其中 $\mu(x) = \mathrm{IE} f(x)$， $K(x_i, x_j) = \mathrm{IE}\Big[(f(x_i) - \mu(x_i))(f(x_j) - \mu(x_j))\Big]$ 和 $f \sim \mathcal{G}P(\mu(\cdot), K(\cdot,\cdot))$.

定义。如果对于任意的 $\{x_1, ..., x_n\} \in \mathcal{X}$和$n \in \mathrm{IN}$， $\mathbf{f} = \{f(x_1), ..., f(x_n)\}^T$的分布是

$$\mathbf{f} = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim N\left(\begin{bmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{bmatrix}, \begin{bmatrix} K(x_1, x_1) & \cdots & K(x_1, x_n) \\ \vdots & \ddots & \vdots \\ K(x_1, x_n) & \cdots & K(x_n, x_n) \end{bmatrix} \right).$$

9.2. 高斯过程回归

考虑数据 $D = \{(x_i, y_i)\}_{i=1}^n$ drawn from the model

$$Y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2),$$

我们将在函数空间上放置一个先验，使用高斯过程

$$f \sim \mathcal{G}P(\mu(\cdot), K(\cdot,\cdot)).$$

我们还给出了一些新的变量或测试数据 $T = \{x_i^*\}^{mi}{}_{=1}$each of which would have a corresponding y_i^*.

现在我们提供一些符号

$$\mathbf{X} = \begin{bmatrix} -x_1- \\ \vdots \\ -x_n- \end{bmatrix}, \quad \mathbf{X}^* = \begin{bmatrix} -x_1^*- \\ \vdots \\ -乘^*乘- \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{Y}^* = \begin{bmatrix} y_1^* \\ \vdots \\ y_m^* \end{bmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \boldsymbol{\varepsilon}^* = \begin{bmatrix} \varepsilon_1^* \\ \vdots \\ \varepsilon_m^* \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} f(\mathsf{X}_1) \\ \vdots \\ f(x_n) \end{bmatrix}, \quad \mathbf{f}^* = \begin{bmatrix} f(x_1^*) \\ \vdots \\ f(x_m^*) \end{bmatrix}.$$

我们的最终目标是指定 $\mathbf{Y}^*$上的预测分布，我们知道它将是多元正态分布

$$\mathbf{Y}^* \mid \mathbf{X}^*, \mathbf{X} \sim \mathrm{N}(\mu^*, \Sigma^*).$$

现在首先观察

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{Y}^* \end{bmatrix} \Big| \mathbf{X}^*, \mathbf{X} = \begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon}^* \end{bmatrix} \sim \mathrm{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I} & K(\mathbf{X},\mathbf{X}^*) \\ K(\mathbf{X}^*,\mathbf{X}) & K(\mathbf{X}^*,\mathbf{X}^*) + \sigma^2\mathbf{I} \end{bmatrix} \right),$$

其中 $K(\mathbf{X},\mathbf{X})$ 是一个 $n \times n$矩阵，其中 $\mathbf{K}_{ij} = K(x_i, x_j)$，而 $K(\mathbf{X}^*,\mathbf{X}^*)$ 是一个 $m \times m$ 矩阵，其中 $\mathbf{K}_{ij}^* = K(x_{i,}^*{}_{x_j^*})$。

为了得到关于 $\mathbf{Y}^*$的预测分布，我们写出条件 $\mathbf{Y}$ $^* \mid \mathbf{X}^*, \mathbf{X}$。给定上述多元正态分布，我们简单地对所有其他变量进行条件处理，以获得后验预测密度的均值和协方差：

$$\begin{aligned} \mu^* &= K(\mathbf{X}^*,\mathbf{X})(K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I})^{-1}\,\mathbf{Y} \\ \Sigma^* &= K(\mathbf{X}^*,\mathbf{X}^*) + \sigma^2\mathbf{I} - K(\mathbf{X}^*,\mathbf{X})(K(\mathbf{X},\mathbf{X}) + \sigma^2\mathbf{I})^{-1}K(\mathbf{X},\mathbf{X}^*) \end{aligned}$$

。高斯过程回归的美妙之处在于我们可以使用核函数对函数进行先验设定，并在有限数量的点上评估函数值的方差，这些都只基于多元正态分布的性质

分布。 这是一个非常强大的非线性预测工具。 核函数、RKHS和高斯过程之间存在着强烈的关系。 也有一些微妙的差异。 主要的差异来自所谓的$\mathrm{Kalianpur0}-1$定律

定理$(\mathrm{Kallianpur}\ 1970)$.如果 $Z\sim \mathcal{G}P(\mu, K)$是一个具有协方差核 K和均值 $\mu\in \mathcal{H}_K$的高斯过程，并且 $\mathcal{H}_K$是无限维的，那么$\mathbf{P}(Z\in \mathcal{H}_K) = 0$.

上述定理的要点是，如果我们指定一个核 K并确保高斯过程的均值在与核$_K$对应的$\mathrm{rkhs}\ \mathcal{H}\ K$中，从这个高斯过程中抽取的样本将不在$\mathrm{rkhs}$中。 可以正式证明的是，如果我们取任意的随机函数，称之为 g，那么对于所有的 g，以下命题成立

$$\int_{\mathcal{X}} g(u)K(x,\, u)\ \mathrm{d}u \in \mathcal{H}_K.$$

讲座 10
稀疏回归

我们之前已经看到，在 p □ n的情况下，以下岭回归模型允许我们进行稳定的推断

$$\arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|^2, \quad \lambda > 0.$$

通常我们不仅想要一个好的预测模型，还想知道哪些变量与预测相关。 同时推断一个好的回归模型和选择变量的问题被称为同时回归和变量选择。 在本讲座中，我们将介绍一些标准的同时回归和变量选择方法。

我们首先陈述标准模型

$$Y_i = (\beta^*)^T x_i + \varepsilon_i,$$

然而，我们现在假设大多数坐标的回归系数为零 $(i= 1, .., p)$。 真实模型的非零坐标子集$\mathcal{A}_* = \{j : |\beta_*^{(j)}| \neq 0\}$和非零系数的数量表示为 $|\mathcal{A}_*|$。

我们的目标是给定数据 $D = \{(x_i, y_i)_{i=1}^n\}$ 来推断 β ，使得

(1) 选择一致性：非零子集 β的表示为 $\mathcal{A} = \{j : |\beta^{(j)}| \neq 0\}$。 我们希望两个子集 $\mathcal{A}_*$和 $\mathcal{A}$在任何有限的情况下都接近，并且在finite n和$\to\infty$时相同。

(2) 估计一致性：选择集中的系数收敛得有多好：

$$\lim_{n\to\infty} \beta_{\mathcal{A}_*} = \beta^*_{\mathcal{A}_*}$$

我们将用以下最小化问题来进行同时回归和变量选择的方法

$$\arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|_q^q, \quad \lambda > 0,$$

其中 $\|\beta\|^{q\,q}$是通过 q-范数进行惩罚。 我们已经看到最小化2-范数会导致岭回归。现在我们将探讨另外两个范数：1-范数和0-范数。

我们从零范数开始。

$$\beta := \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|_0^0, \quad \lambda > 0,$$

这等价于以下最小化问题

$$\beta := \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\sum_{j=1}^{p} I(\beta_j = 0), \quad \lambda > 0,$$

这意味着使用尽可能少的变量来最小化平方误差， λ作为变量数量和误差之间的权衡。 上述最小化问题是NP-hard问题，因为它可以约化为三个集合的精确覆盖问题。 这意味着我们无法以任何高效的方式实现上述优化问题，即使当 p= 10时，也需要在一个巨大的空间中进行搜索。

10.1. LASSO: 最小绝对选择和收缩算子

套索程序背后的思想是最小化

$$\beta := \arg\min_{\beta\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|_1, \quad \lambda > 0,$$

出于我们将讨论的原因，最小化上述惩罚损失函数会导致变量选择和回归，导致回归系数恰好为零。 有人认为最小化$_1$-范数正则化问题是0-范数最小化问题的很好近似。 我们将探讨为什么这个最小化问题近似于0-范数以及最小化$_1$-范数的方法。

10.1.1. 多面体的几何

回想一下，$\arg\min_{\beta\in\mathbb{R}^p}\frac{1}{n}$之间存在等价关

$$\begin{aligned} &\text{系} && -\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda\|\beta\|_1, \\ &\arg\min_{\beta\in\mathbb{R}^p}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 \text{ 满足} && \|\beta\|_1 \le \tau. \end{aligned}$$

我们将对比以下两个最小化问题

$$\begin{aligned} &\arg\min_{\beta\in\mathbb{R}^p}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 \text{ 满足} && \|\beta\|_1 \le \tau. \\ &\arg\min_{\beta\in\mathbb{R}^p}\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 \text{ 满足} && \|\beta\|_2^2 \le \tau. \end{aligned}$$

上面问题的解受限于原点周围的1范数球
而下面问题的解受限于原点周围的2范数球.
考虑真实的 β向量为 β_*，平方损失的几何形状是椭圆，作为等损失的轮廓。 最小化器是与 p-范数球边界相交的最小损失值。在下面的图中，我们展示了两个变量的情况。

具有稀疏解的最小化器将触及/相交于轴上的误差椭圆的等高线，即稀疏面的 p维多面体。 例如，当约束是原点周围的2-范数球时，相交点集中在轴上的可能性非常低。 特别是在高维空间中，1-范数球的几何形状与椭圆相交于少数点。 例如，0-范数是一个位于轴上的星形或尖峰，因此它总是稀疏的。

尽管我们考虑了约束优化的1-范数问题，但相同的结果适用于套索回归。

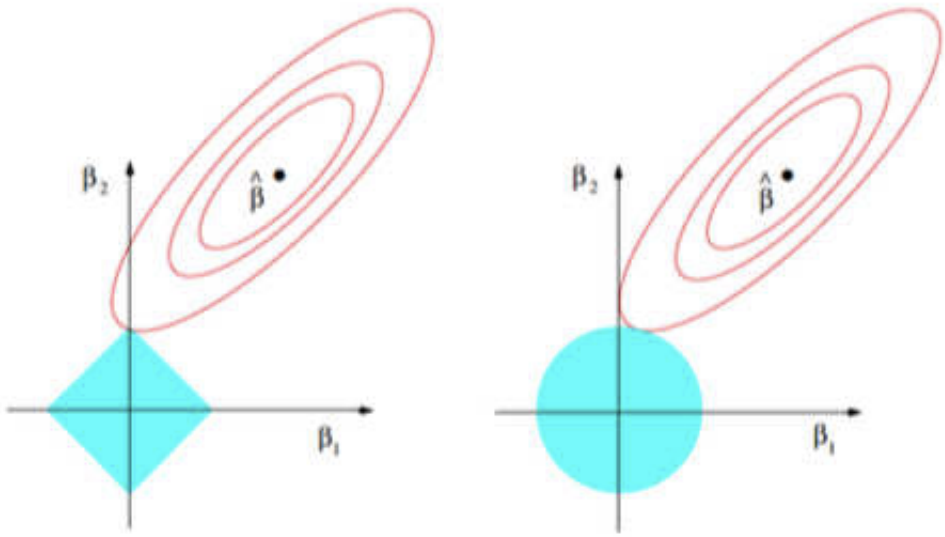


图1。 两个变量的1-范数最小化在左边，2-范数最小化在右边。

10.1.2. 正则化路径

回顾优化问题

$$\beta := \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \|\beta\|_1, \quad \lambda > 0,$$

众所周知，对于 $\lambda = 0$，解 $\beta_{\text{LASSO}} = \beta_{\text{OLS}}$，对于 $\lambda = \infty$，解为 $\beta_{\text{LASSO}} = 0$。 Lasso回归系数 β_λ是一个 p维向量，对于较大的 λ值，其中很多值被设为零。 正则化路径的思想是观察 β随着 λ的变化的情况，可以将 λ看作是 x轴， β看作是 y轴上的点。

数学事实是 β的图形将是分段连续的，并且在某一点趋近于零。

正则化路径的思想是帮助选择模型中保留多少个变量。 在岭回归模型中，很难将正则化参数解释为系数，因为系数不会趋近于零，变化缓慢。 在Lasso模型中，这种情况有所缓解。

在下面的图中，我们考虑了两个回归分析，一个使用岭回归，另一个使用套索回归，数据集是一个与前列腺癌相关的问题。 响应变量是PSA，一种前列腺癌的生物标志物，协变量是几个其他生物标志物。

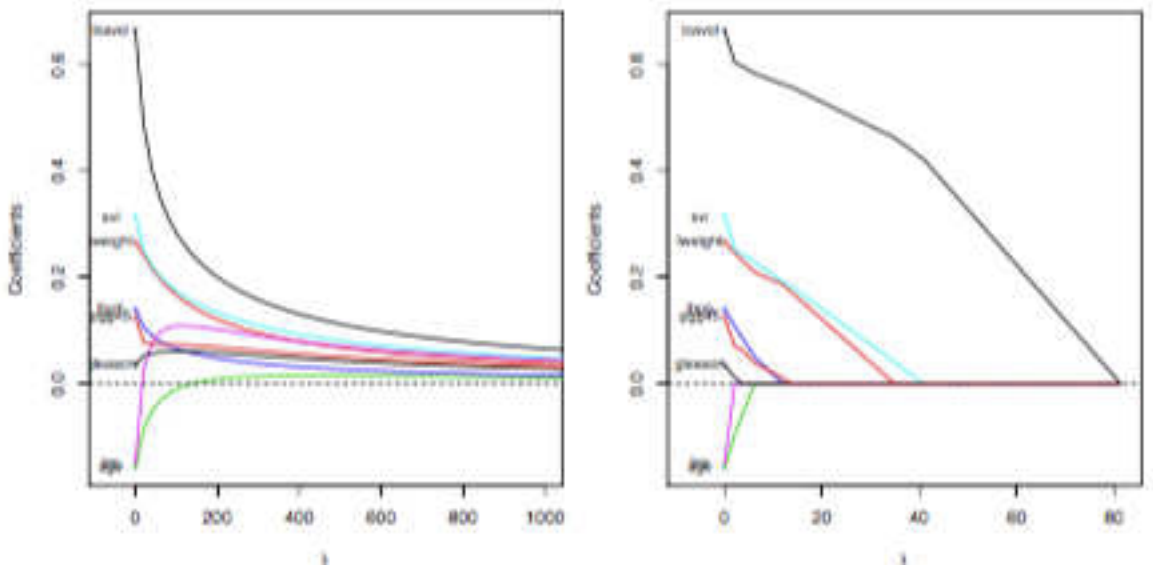


图**2**。 左图是岭回归的正则化路径，x轴是λ参数，y轴是绘制的系数。. 右图是套索回归的相同图。 响应变量是PSA，一种前列腺癌的生物标志物，协变量是几个生物标志物。

讲座 11
增强学习假设和Adaboost

投票算法或者最终分类或回归函数是“简单”或“较弱”分类器的加权组合的算法在各种应用中被广泛使用。

我们将更深入地研究两个投票算法的例子：自助采样聚合（BAGGING）和提升（boosting）算法。

11.1. 提升

提升算法，特别是AdaBoost（自适应提升），对各种实际算法产生了重要影响，并且也成为各个领域理论研究的焦点。 提升这个正式术语和第一个提升算法起源于理论计算机科学中的计算复杂性领域。特别是，提升的学习形式来自于可能近似正确（PAC）学习的概念。

11.2. PAC学习

可能近似正确（PAC）学习的概念是由Leslie Valiant于1984年提出的，旨在描述可学习性。让 $\mathcal{X}$是一个集合，该集合包含了学习问题中所有感兴趣对象的编码。学习算法的目标是从已知概念类别 $\mathcal{C}$中推断出 $\mathcal{X}$的某个未知子集，称为概念。 与之前的统计学习问题的形式化不同，这个问题的表示问题是由于计算问题而引起的。

- 概念类别是对 $\mathcal{X}$的表示类别，表示为一个二元组 $(\sigma, \mathcal{C})$，其中$\mathcal{C} \subseteq \{0,1\}^*$且 $\sigma : \mathcal{C} \to 2^{\mathcal{X}}$。 对于 $c \in \mathcal{C}$， $\sigma(c)$ 是对 $\mathcal{X}$的一个概念， $\sigma(\mathcal{C})$ 是由 $(\sigma, \mathcal{C})$ 表示的概念类别。 对于 $c \in \mathcal{C}$，正例是 $\mathrm{pos}(c) = \sigma(c)$，负例是$\mathrm{neg}(c) = \mathcal{X} - \sigma(c)$。 记号 $c(x) = 1$ 等价于 $x \in \mathrm{pos}(c)$且 $c(x) = 0$ 等价于 $x \in \mathrm{neg}(c)$。 我们假设域中的点$x \in \mathcal{X}$和表示 $c \in \mathcal{C}$可以通过编码分别以长度 $|x|$和 $|c|$高效地表示。

- 参数化表示 我们将研究由索引参数化的表示类，导致域 $\mathcal{X} = \cup\, n_{\geq 1}\mathcal{X}_n$和表示类 $\mathcal{C} = \cup_{n\geq 1}\mathcal{C}_n$。 索引 $_n$ 作为 $\mathcal{C}$中概念复杂性的度量。 例如， $\mathcal{X}$可以是集合 $\{0,1\}$ $_n$和$\mathcal{C}$是所有布尔公式的集合，其中变量个数为 $_n$。

- 表示类 $\mathcal{C}$的高效评估 如果 $\mathcal{C}$是 $\mathcal{X}$上的表示类，则如果存在一个（概率）多项式时间评估算法 $\mathcal{A}$，给定表示 $c\in \mathcal{C}$和域点 $x\in \mathcal{X}$，输出 $c(\ x)$。

- 样本 从域 $\mathcal{X}$中的标记示例是一对$< x, b >$，其中 $x\in \mathcal{X}$且 $b\in \{0,1\}$。 样本 $S= (< x_1, b_1>, ..., < x_m, b_m >)$ 是一系列有限的标记示例。 表示 $c\in \mathcal{C}$的标记示例具有形式$< x, c(\ x) >$。 表示 h和示例$< x, b >$一致，如果 $h(\ x) = b$。 表示 h和样本 S一致，如果 h与 S中的每个示例一致。

- 对于一个表示类 $\mathcal{C}$的学习算法，它将从单个表示 $c\in \mathcal{C}$中接收示例，我们称之为目标表示。 目标表示的示例是以概率方式生成的： D_c是一个固定但任意的正(c)和负(c)分布。 这是目标分布。 学习算法将获得一个称为oracle的访问权限 EX，它以单位时间返回根据目标分布 D_c绘制的目标表示示例。

- 误差度量给定目标表示 $c\in \mathcal{C}$和目标分布 D，表示 $h\in \mathcal{H}$的误差为

$$e_c(h) = D(h(x)\ = c(x))。$$

在上述公式中， $\mathcal{C}$是目标类。 在上述公式中， $\mathcal{H}$是假设类。 算法 $\mathcal{A}$是 $\mathcal{C}$的学习算法，输出$h_{\mathcal{A}}\in \mathcal{H}$是 $\mathcal{A}$的假设。现在我们可以定义可学习性：

定义（强学习）。 让 $\mathcal{C}$和 $\mathcal{H}$是在 $\mathcal{X}$上可多项式评估的表示类。. 如果存在一个（概率）算法 $\mathcal{A}$可以访问 EX，以，δ为输入，并且满足以下条件，则 $\mathcal{C}$可以通过 $\mathcal{H}$进行多项式学习：对于任何目标表示 $c\in \mathcal{C}$，任何目标分布 D和任何输入值 $0< \varepsilon$，$\delta < 1$，算法 $\mathcal{A}$在多项式时间内终止，并输出一个表示$_{\mathcal{A}}\in \mathcal{H}$，使得 ^{1} $_{\bar{\delta}}$的概率大于ε满足c（A）$< \varepsilon$。

参数 ε是准确度参数，参数 δ是置信度参数。 这两个参数描述了名为“可能（δ）近似（ε）正确（ε_c（A））”的概念。 上述定义有时被称为分布无关学习，因为该属性适用于目标表示和目标分布。

在PAC学习中，大量的研究集中在哪些表示类 $\mathcal{C}$是多项式可学习的上。

到目前为止，我们将学习定义为近似地接近目标概念。 另一种学习模型称为弱学习，考虑的情况是学习算法需要比随机稍微好一点。

定义(弱学习)。设 $\mathcal{C}$和 $\mathcal{H}$是在 $\mathcal{X}$上多项式可评估的表示类。 如果存在一个（概率）算法 $\mathcal{A}$可以访问 EX，并且接受输入ε, δ，满足以下条件：对于任何目标表示 $c\in\mathcal{C}$，任何目标分布 D，任何输入值 $0<\varepsilon,\delta<1$，算法 $\mathcal{A}$在多项式时间内终止，输出一个表示$_{\mathcal{A}}\in\mathcal{H}$，使得概率大于 $^{1}\ _{\overline{\delta}}$满足$c(\ h)<\ _{2}{}^{1}-$

$-\quad\frac{1}{p(|c|)}$, 其中 p 是一个多项式。

定义(样本复杂度)。 给定一个学习算法 $\mathcal{A}$对于一个表示类 $\mathcal{C}$。. 学习算法$_{\mathcal{A}}$在输入ε, δ 的最坏情况下对目标表示 $c\in\mathcal{C}$ 和目标分布 D进行的调用次数 $\mathcal{A}(\varepsilon,\delta)$是算法 $\mathcal{A}$ 的样本复杂度。

现在我们列举一些布尔类，我们将它们的可学习性陈述为可学习性的正面或负面例子。

- 类 M_n由布尔变量 $x_i, ..., x_n$的单项式组成。
- 对于一个常数 k，类$kCNF_n$(合取范式)由形式为 $C_1\wedge\cdots\wedge C_l$的布尔公式组成，其中每个 C_i是布尔变量 $x_i, ..., x_n$的最多 k个单项式的析取。
- 对于一个常数 k，类$kDNF_n$（析取范式）由形式为 $T_1\vee\cdots\vee T_l$的布尔公式组成，其中每个 T_i是一个布尔变量 x_i， $..., x_n$上至多 k个单项式的合取。
- 布尔阈值函数 I（$\sum^{n}{}_{i=1}\, w_i x_i> t$）其中 $w_i\in\{0,1\}$且 I是指示函数。

定义（经验风险最小化，ERM）。一致算法 $\mathcal{A}$是指输出与样本 S和可能的假设范围一致的假设 h的算法 $\mathcal{A}$。

上述算法是在目标概念的假设空间中具有零误差的ERM。

定理。 如果假设类是有限的，则一致算法 $\mathcal{A}$可以学习假设类 $\mathcal{C}$。

定理。 布尔阈值函数是不可学习的。

定理。 $\{f\vee g: f\in kCNF, g\in kDNF\}$是可学习的。

定理。 $\{f\wedge g: f\in kDNF, g\in kCNF\}$是可学习的。

定理。 设 $\mathcal{C}$是一个具有有限VC维度$VC(\mathcal{C})=d<\infty$的概念类。
那么 $\mathcal{C}$可以通过一致算法 $\mathcal{A}$来学习。

11.3. 假设增强问题

在20世纪80年代末，一个重要的理论和实践问题是强可学习性和弱可学习性是否等价。 这就是假设增强问题：

猜想。 如果一个概念类 $\mathcal{C}$是弱可学习的，那么它也是强可学习的。

这个猜想在1990年被Robert Schapire证明为真。

定理。如果一个概念类 $\mathcal{C}$是弱可学习的，那么它也是强可学习的。

上述定理的证明基于一个特定的算法。 下面的算法输入一个弱学习器、一个错误参数 ε、一个置信度参数 δ、一个预言机 EX，并输出一个强学习器。 在算法的每次迭代中，错误率为 ε的弱学习器被提升，使其错误率降低到$3\varepsilon^2 - 2\varepsilon^3$。

算法1 : Learn(ε, δ, EX)

输入 ：错误参数 ε、置信度参数 δ、示例预言机 EX

返回：概率$\geq 1-\delta$的 ε接近目标概念 c的 h。

如果 $\varepsilon \geq 1/2 - 1/p(n,s)$ 则返回 $\mathbf{WeakLearn}(\delta, EX)$；
$\alpha \leftarrow g^{-1}(\varepsilon)$：其中 $g(x) = 3x^2 - 2x^3$;

期望$_1$ ←期望;
好$_1$ ← 学习(α, $\delta/5$, 期望$_1$);
$\tau_1 \leftarrow \varepsilon/3$;
让 $\hat{a}_1$ 成为一个估计值 $a_1 = \Pr_{x\sim D}[h_1(x) = c(x)]$
选择一个足够大的样本
$|a_1 - \hat{a}_1| \leq \tau_1$的概率$\geq 1 - \delta/5$
如果 $\hat{a}_1 \leq \varepsilon - \tau_1$ 那么 返回 h_1;

定义期望$_2$()
{抛硬币
如果是正面 那么 返回第一个 $x : h_1(x) = c(x)$
否则 返回第一个 $x : h_1(x) = c(x)$}
好$_2$ ← 学习(α, $\delta/5$, EX_2);
$\tau_2 \leftarrow (1-2\alpha)\varepsilon/9$;
让$\hat{e}$是 e的估计值$= \Pr_{x\sim D}[h_2(x) = c(x)]$
选择一个足够大的样本使得
$|e - \hat{e}| \leq \tau_2$的概率 $\geq 1 - \delta/5$
如果 $\hat{e} \leq \varepsilon - \tau_2$ 那么 返回 h_2;

定义 EX_3()
{返回第一个 $x : h_1(x) = h_2(x)$ };
$h_3 \leftarrow$ 学习(α, $\delta/5$, EX_3);

定义 $h(x)$
{ $b_1 \leftarrow h_1(x)$, $b_2 \leftarrow h_2(x)$
如果 $b_1 = b_2$ 那么 返回 b_1
否则 返回 $h_3(x)$}
返回 h

上述算法可以总结如下：

(1) 在前 N个训练点上学习一个初始分类器

(2) 在一个新的样本中学习 N个点，其中一半被 h_1误分类。

(3) 学习 N个点，其中 h_1和 h_2不一致。
(4) 提升的分类器 $h=$ 多数投票(h_1, h_2, h_3)。

基本结果是，如果个别分类器 h_1, h_2和 h_3的错误率为 ε ，则提升的分类器的错误率为$2\varepsilon^2 - 3\varepsilon^3$。

为了证明定理，需要证明算法在以下意义上是正确的。

定理。对于 $0< \varepsilon < 1/2$和 $0< \delta <\leq 1$，通过调用$Learn(\varepsilon, \delta, EX)$返回的假设与目标概念的接近程度至少为 $1-\delta$的概率。

我们首先定义一些量

$$\begin{aligned} p_i &= \Pr_{x\sim D}[h_i(x) = c(x)] \\ q &= \Pr_{x\sim D}[h_1(x) = h_2(x)] \\ w &= \Pr_{x\sim D}[h_2(x) = h_1(x) = c(x)] \\ v &= \Pr_{x\sim D}[h_1(x) = h_2(x) = c(x)] \\ y &= \Pr_{x\sim D}[h_1(x) = h_2(x) = c(x)] \\ z &= \Pr_{x\sim D}[h_1(x) = h_2(x) = c(x)]. \end{aligned}$$

给定上述量

$$w + v = \Pr_{x\sim D}[h_1(x) = c(x)] = 1 - a_1 \tag{11.1}$$

$$y + z = \Pr_{x\sim D}[h_1(x) = c(x)] = a_1. \tag{11.2}$$

我们可以明确地表示 EX_i返回一个实例 x的机会，以上述变量为基础:

$$\begin{aligned} D_1(x) &= D(x) \\ D_2(x) &= \frac{D(x)}{2}\left(\frac{p_1(x)}{a_1} + \frac{1-p_(x)}{1-a_1}\right) \\ D_3(x) &= \frac{D(x)q(x)}{w+y}. \end{aligned} \tag{11.3}$$

从方程(11.3)我们有

$$\begin{aligned} 1 - a_2 &= \sum_{x\in\mathcal{X}_n} D_2(x)(1-p_2(x)) \\ &= \frac{1}{2\text{一个}_1}\sum_{x\in\mathcal{X}_n} D(x)p_1(x)(1-p_2(x)) + \frac{1}{2(1-a_1)}\sum_{x\in\mathcal{X}_n} D(x)(1-p_1(x))(1-p_2(x)) \\ &= \frac{y}{2\text{一个}_1} + \frac{z}{2(1-a_1)}. \end{aligned}$$

将上述方程与方程(11.1)和(11.2)结合起来，我们可以解出 w 和 z的值，以y, a_1, a_2为变量。

$$\begin{aligned} w &= (2\text{一个}_2 - 1)(1-a_1) + \frac{y(1-}{a_1}a^1) \\ z &= a_1 - y. \end{aligned}$$

我们现在控制着数量

$$\begin{aligned}
\Pr_{x\sim D}[h(x)=c(x)] &= \Pr_{x\sim D}[(h_1(x)=h_2(x))\vee(h_1(x)=h_2(x)\wedge h_3(x)=c(x))] \\
&= z+\sum_{x\in\mathcal{X}_n} D(x)q(x)p_3(x) \\
&= z+\sum_{x\in\mathcal{X}_n}(w+y)D_3(x)p_3(x) \\
&= z+a_3(w+y) \\
&\leq z+\alpha(w+y) \\
&= \alpha(2a_2-1)(1-a_1)+a_1+\frac{y(\alpha-a_1)}{a_1} \\
&\leq \alpha(2a_2-1)(1-a_1)+\alpha \\
&\leq \alpha(2\alpha-1)(1-\alpha)+\alpha = 3\alpha^2-2\alpha^3=\varepsilon,
\end{aligned}$$

这些不等式是由于 $a_i \leq \alpha < 1/2$ 和 $y \leq a_1$ 的事实得出的。 □还需要证明该算法在多项式时间内运行。. 以下引理说明了这一点。 该引理的证明超出了讲座笔记的范围。

引理。在良好的运行中，**Lea**$rn(\varepsilon,\delta/2,EX)$的预期执行时间是多项式的，与 $m, 1/\delta, 1/\epsilon$成比例。

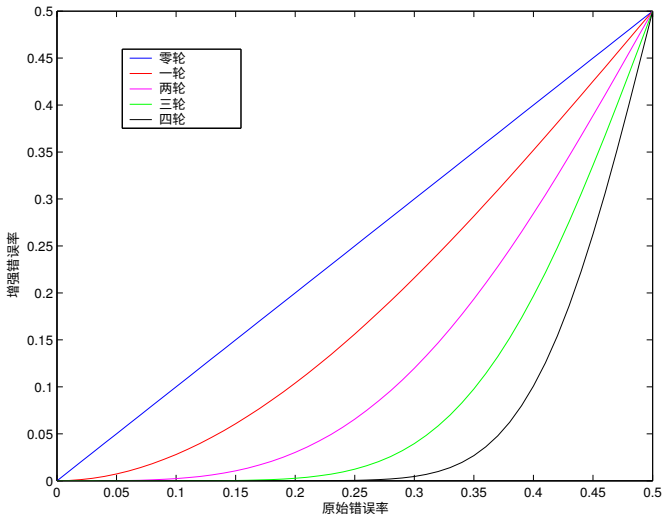


图1. 绘制了增强错误率作为初始错误的函数
对于不同数量的增强轮数。

11.4. 自适应增强 (AdaBoost)

我们将称上述增强的公式为多数投票算法。
Schapire提出的多数投票增强公式涉及通过过滤进行增强，因为一个弱学习器作为另一个的过滤器。 Yoav Fruend还基于过滤器开发了另一种多数投票增强公式。

后来，这两种算法都进行了调整，以便使用采样权重代替过滤器。 然而，所有这些算法都存在一个问题，即弱学习器的强度必须事先知道。

Freund和Schapire开发了以下自适应增强算法AdaBoost来解决这些问题。

算法2： AdaBoost

输入：样本 $S= (x_i, y_i)^{Ni}{}_{=1}$, 弱学习器, 迭代次数 T

返回：$h(x) =$ 符号 $\left[\sum_{i=1}^{T} \alpha_i h_i(x)\right]$

对于 $i=1$ 到 N 执行 $w_i^0 = 1/N$;

对于 $t=1$到 T执行

- $h_t \leftarrow$使用权重 w^t调用WeakLearn函数;
- $\varepsilon_t = \sum_{j=1}^{N} w_j^t I_{\{y_j = h_t(x_j)\}}$;
- $\alpha_t = \log((1-\varepsilon_t)/\varepsilon_t)$;
- 对于 $j=1$ 到 N 执行 $w_j{}^{t+1} = w_j{}^t \exp\ (-\alpha_t y_j h_t(x_j))$;
- $Z_t = \sum_{j=1}^{N} w_j^{t+1}$;

对于 $j=1$ 到 N 执行 $w_j{}^{t+1} = w_j{}^{t+1}/Z_t$;

对于上述算法，我们可以证明训练误差会随着增强迭代而减小。 上述算法的优点是我们不需要在所有轮次上都使用相同的 γ。 我们只需要在每个增强轮次中存在一个$\gamma_t > 0$。

定理。假设$AdaBoost$调用 $WeakLearn$生成具有误差 $\varepsilon_1, ..., \varepsilon_T$的假设。 假设每个 $\varepsilon_i \leq 1/2$，并且令 $\gamma_i = 1/2 - \varepsilon_i$，则以下上界对于假设 h成立

$$\frac{|j : h(x_j) = y_j|}{N} \leq \prod_{i=1}^{T} \sqrt{1 - 4\gamma_i^2} \leq \exp\left(-2\sum_{i=1}^{T} \gamma_i^2\right).$$

证明。

如果 $y_i = h(x_i)$，那么 $y_i h(x_i) \leq 0$，并且 $e^{-y_i h(x_i)} \geq 1$。因此

$$\begin{aligned} \frac{|j : h(x_j) = y_j|}{N} &\leq \frac{1}{N}\sum_{i=1}^{N} e^{-y_i h(x_i)}, \\ &= \sum_{i=1}^{N} w_i^{T+1} \prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} Z_t. \end{aligned}$$

此外，由于 $\alpha_t = \log((1-\varepsilon_t)/\varepsilon_t)$和$1 + \quad x \leq e^x$

$$Z_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)} = \sqrt{1 - 4\gamma_t^2} \leq e^{-2\gamma_t}. \ \square$$

11.5. Adaboost的统计解释

在本节中，我们将将boosting重新解释为一种贪婪算法来拟合一个附加模型。 我们首先将我们的弱学习器定义为一个参数化的函数类$h_\theta(x) = h(x;\theta)$，其中 $\theta \in \theta$。 如果我们将每个弱学习器视为基函数那么增强的假设 $h(x)$可以被视为弱学习器的线性组合

$$h(x) = \sum_{i=1}^{T} \alpha_i h_{\theta_i}(x),$$

其中 $h_{\theta_i}(x)$是由 θ_i参数化的第i个弱学习器。 一种设置参数 θ_i和权重 α_i的方法被称为前向逐步建模。 在这种方法中，我们按顺序添加新的基函数或弱学习器，而不调整当前解的参数和系数。 以下算法实现了前向逐步加法建模。

算法3：前向逐步加法建模

输入：样本 $S{=}\,(x_i, y_i)^{Ni}{}_{=1}$, 弱学习器, 迭代次数 T, 损失函数 L

返回：$h(x) = \left[\sum_{i=1}^{T} \alpha_i h_{\theta_i}(x)\right]$

$h_0(x) = 0$;
对于 $i{=}1$到 T执行
  $(\alpha_t, \theta_t) = \arg\min_{\alpha\in \mathrm{IR}^+, \theta\in\Theta} \sum_{i=1}^{N} L(y_i, h_{t-1}(x_i) + \alpha h_\theta(x))$;
  $h_t(x) = h_{t-1}(x) + \alpha_t h_{\theta_i}(x)$;

我们现在将展示上述具有指数损失的算法

$$L(y, f(x)) = e^{-yf(x)}$$

等价于AdaBoost。
在每次迭代中执行以下最小化

$$\begin{aligned}
(\alpha_t, \theta_t) &= \arg\min_{\alpha\in \mathrm{IR}^+, \theta\in\Theta} \sum_{i=1}^{N} \exp[-y_i(h_{t-1}(x_i) + \alpha h_\theta(x))], \\
(\alpha_t, \theta_t) &= \arg\min_{\alpha\in \mathrm{IR}^+, \theta\in\Theta} \sum_{i=1}^{N} \exp[-y_i h_{t-1}(x_i)] \exp[-y_i \alpha h_\theta(x)], \\
(\alpha_t, \theta_t) &= \arg\min_{\alpha\in \mathrm{IR}^+, \theta\in\Theta} \sum_{i=1}^{N} w_i^t \exp[-y_i \alpha h_\theta(x)], \qquad (11.4)
\end{aligned}$$

其中 $w_i{}^t = \exp[-y_i h_{t-1}(x_i)]$ 不影响优化函数。对于任意 $\alpha > 0$，方程（11.4）中的目标函数可以重写为

$$\begin{aligned}
\theta_t &= \arg\min_{\theta\in\Theta}\left[e^{-\alpha}\sum_{y_i=h_\theta(x_i)} w_i^t + e^{\alpha}\sum_{y_i=h_\theta(x_i)} w_i^t\right], \\
\theta_t &= \arg\min_{\theta\in\Theta}\left[(e^{-\alpha}-e^{\alpha})\sum_{i=1}^N w_i^t I_{\{y_i=h_\theta(x_i)\}} + e^{\alpha}\sum_{i=1}^N w_i^t\right], \\
\theta_t &= \arg\min_{\theta\in\Theta}\sum_{i=1}^N w_i^t I_{\{y_i=h_\theta(x_i)\}}.
\end{aligned}$$

因此，最小化方程（11.4）的弱学习器将最小化加权错误率，如果我们将其代入方程（11.4），我们可以解出 α_t。

$$\alpha_t = \frac{1}{2}\log\frac{1-\varepsilon_t}{\varepsilon_t},$$

其中

$$\varepsilon_t = \sum_{i=1}^N w_i^t I_{\{y_i=h_t(x_i)\}}。$$

最后要展示的是线性模型的更新

$$h_t(x) = h_{t-1}(x) + \alpha_t h_t(x),$$

等价于AdaBoost中使用的重新加权。由于指数损失函数和每次迭代的加法更新，上述求和可以重写为

$$w_i^{t+1} = w_i{}^t e^{-\alpha y_i h_t(x_i)}。$$

因此，AdaBoost可以被解释为通过前向逐步加法建模来最小化指数损失准则的算法。

现在我们给出一些动机，解释为什么指数损失是分类问题中一个合理的损失函数。 第一个论点是，就像SVM分类中的铰链损失一样，指数损失也作为错分损失的一个上界（见图2）。

使用指数损失的另一个简单动机是期望损失相对于某个函数类$\mathcal{H}$的最小化器。

$$f^*(x) = \arg\min_{f\in\mathcal{H}} \mathrm{IE}_{Y|x}\left[e^{-Yf(x)}\right] = \frac{1}{2}\log\frac{\Pr(Y=1|x)}{\Pr(Y=-1|x)}$$

，估计了log-odds比的一半

$$\Pr(Y=1|x) = \frac{1}{1+e^{-2f^*(x)}}.$$

11.6. Adaboost的边界解释

我们通过最大化间隔，在可分离的情况下，开发了支持向量机的几何形式。 我们将AdaBoost制定为最大化间隔的问题。

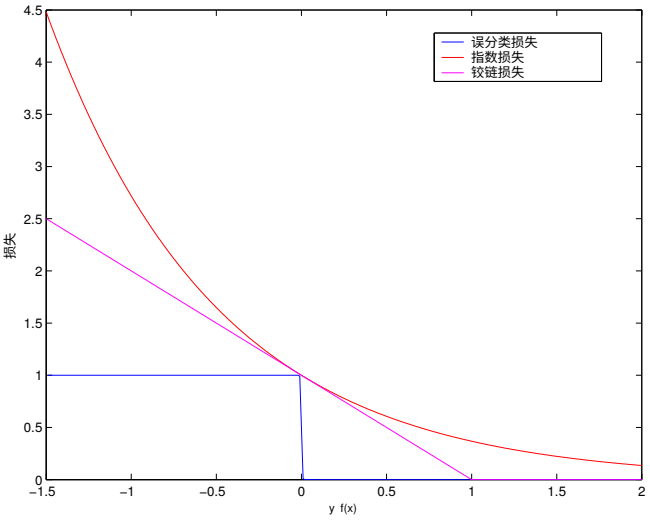


图**2.** 分类损失函数的比较。

回想一下，在具有IR^d中的数据点的线性可分离SVM中，给定数据集 S，以下优化问题描述了最大间隔分类器

$$\hat{w} = \arg\max_{w\in\mathrm{IR}^d} \min_{x_i\in S} \frac{y_i\langle w, x_i\rangle}{||w||_{L_2}}.$$

在AdaBoost的情况下，我们可以构建一个具有T个弱分类器维度的坐标空间，$u \in \mathrm{IR}^T$，其中 $\{u_1, ..., u_T\}$对应于弱分类器 $\{u_1 = f_1(x), ..., u_T = f_T(x)\}$的输出。我们可以证明AdaBoost是解决以下最小最大问题的迭代方法

$$\hat{w} = \arg\max_{w\in\mathrm{IR}^T} \min_{u_i\in S} \frac{y_i\langle w, u_i\rangle}{||w||_{L_1}},$$

其中 $u_i = \{f_1(x_i), ..., f_T(x_i)\}$，最终分类器的形式为

$$h_T(x) = \sum_{i=1}^{T} \hat{w}_i^T f_i(x).$$

这是根据前向加性分步建模解释得出的，因为在可分性下，每次将弱分类器添加到线性展开中，将会得到一个增强的假设 h_t，作为 t的函数，将在 $y_i h_t(x_i) \quad \forall i$上是非递减的，同时受到 w^t的 L_1范数的约束， $||w||_{L_1} = 1$，这是因为每次迭代的权重必须满足分布要求。

两个不同的权重范数在两个不同的优化问题中产生了一个有趣的几何结构。主要思想是我们希望将 w的范数与在IR^d（在SVM情况下）或IR^T（在提升情况下）中的点的范数属性相关联。 根据H¨older不等式的对偶范数 $||x||_{L_q}$ 和 $||w||_{L_p}$ 其中 $\frac{1}{p} + \frac{1}{q} = 1$，$p, q \in [1, \infty]$，以下成立

$$|\langle x, w\rangle| \le ||x||_{L_q} ||w||_{L_p}.$$

上述不等式意味着最小化 w的 L_2范数等价于最大化超平面与数据之间的 L_2距离。 类似地，最小化 w的 L_1范数等价于最大化超平面与数据之间的 L_∞范数。

一维浓度不等式

11.1. 大数定律

在本讲座中，我们将研究固定函数的浓度不等式或大数定律。 设$(\Omega, \mathcal{L}, \mu)$是一个概率空间。 设 $x_1, ..., x_n$是Ω上的实随机变量。 如果一个随机变量序列 y_n几乎必然收敛于一个随机变量 Y，则 $\mathrm{IP}(y_n \to Y) = 1$。 如果一个随机变量序列 y_n以概率收敛于一个随机变量 Y，则对于每个$\epsilon > 0$，$\lim_{n\to\infty} \mathrm{IP}(|y_n - Y| > \epsilon) = 0$。 如果一个序列 $x^1, ..., x_n$满足强大数定律，则存在常数 c，使得 $\hat{\mu}^n$ 收敛于 c几乎必然。 如果一个序列 $x_1, ..., x\ n$满足弱大数定律，则存在常数 c，使得 $\hat{\mu}^n$ 收敛于 c以概率。 一般情况下，常数 c将是随机变量 $\mathrm{IE}\ x$的期望。

给定随机变量x的函数 $f(\ x)$如果其经验平均值与期望值$\mathrm{IE}\, f(x)$之间的偏差趋近于零，那么它集中。 也就是说$f(x)$满足大数定律。

11.2. 多项式不等式

定理（Jensen）。如果 ϕ是一个凸函数，那么 $\phi(\mathrm{IE}x) \leq \mathrm{IE}\phi(x)$。定

理（$\mathrm{B}i\mathbf{e}\mathrm{n}a\mathrm{y}\mathbf{m}\acute{}\mathbf{e}\text{-}\mathrm{C}\mathbf{he}\mathrm{bys}\mathbf{he}v$）。 对于任意随机变量 x，$\epsilon > 0$

$$\mathrm{IP}(|x| \geq \epsilon) \leq \frac{\mathrm{IE}x^2}{\epsilon^2}.$$

证明。

对于任意随机变量x，$\epsilon > 0$

$$\mathrm{IP}(|x| \geq \epsilon) \leq \frac{\mathrm{IE}e^{\lambda x}}{e^{\lambda\epsilon}}$$

和

$$\mathrm{IP}(|x| \geq \epsilon) \leq \inf_{\lambda<0} e^{-\lambda\epsilon}\mathrm{IE}e^{\lambda x}.$$

证明。

$$\mathrm{IP}(x > \epsilon) = \mathrm{IP}(e^{\lambda x} > e^{\lambda \epsilon}) \leq \frac{\mathrm{IE}e^{\lambda x}}{e^{\lambda \epsilon}}. \ \square$$

11.3. 指数不等式

对于独立随机变量的和或平均值，上述界限可以从多项式改进为指数级别的 ϵ.

定理（Bennet）. 设 $x_1, ..., x_n$为独立随机变量，满足 $\mathrm{IE}x = 0$, $\mathrm{IE}x^2 = \sigma^2$, 且 $|x_i| \leq M$. 对于$\epsilon > 0$

$$\mathrm{IP}\left(|\sum_{i=1}^{n} x_i| > \epsilon \right) \leq 2e^{\frac{-n\sigma^2}{M^2}\phi\left(\frac{\epsilon M}{n\sigma^2}\right)},$$

其中

$$\phi(z) = (1+z)\log(1+z) - z.$$

证明。 我们将证明上述定理的一个边界

$$\mathrm{IP}\left(\sum_{i=1}^{n} x_i > \epsilon \right).$$

$$\mathrm{IP}\left(\sum_{i=1}^{n} x_i > \epsilon \right) \leq e^{-\lambda\epsilon}\mathrm{IE}e^{\lambda \sum x_i} = e^{-\lambda\epsilon}\Pi_{i=1}^{n}\mathrm{IE}e^{\lambda x_i}$$
$$= e^{-\lambda\epsilon}(\mathrm{IE}e^{\lambda x})^n.$$

$$\mathrm{IE}e^{\lambda x} = \mathrm{IE}\sum_{k=0}^{\infty}\frac{(\lambda x)^k}{k!} = \sum_{k=0}^{\infty}\lambda^k\frac{\mathrm{IE}x^k}{k!}$$
$$= 1 + \sum_{k=2}^{\infty}\frac{\lambda^k}{k!}\mathrm{IE}x^2x^{k-2} \leq 1 + \sum_{k=2}^{\infty}\frac{\lambda^k}{k!}M^{k-2}\sigma^2$$
$$= 1 + \frac{\sigma^2}{M^2}\sum_{k=2}^{\infty}\frac{\lambda^k M^k}{k!} = 1 + \frac{\sigma^2}{M^2}(\mathrm{\varepsilon}^{\lambda M} - 1 - \lambda M)$$
$$\leq \mathrm{\varepsilon}^{\frac{\sigma^2}{M^2}(\mathrm{\varepsilon}^{\lambda M} - \lambda M - 1)}.$$

最后一行成立，因为$1 + x \leq \mathrm{\varepsilon}^x$。

因此，

$$\mathrm{IP}\left(\sum_{i=1}^{n} x_i > \epsilon \right) \leq \mathrm{\varepsilon}^{-\lambda\epsilon\ \frac{\sigma^2}{M^2}(\mathrm{\varepsilon}^{\lambda M} - \lambda M - 1)}. \tag{11.5}$$

现在我们通过对 λ求导来进行优化

$$\begin{aligned} 0 &= -\epsilon + \frac{n\sigma^2}{M^2}(M\,\mathrm{\varepsilon}^{\lambda M} - M), \\ \mathrm{\varepsilon}^{\lambda M} &= \frac{\epsilon M}{n\sigma^2} + 1, \\ \lambda &= \frac{1}{M}\log\left(1 + \frac{\epsilon M}{n\sigma^2}\right). \end{aligned}$$

通过将 λ代入方程（11.5）证明了定理。 □Bennet不等式的问题在于很难得到一个简单的表达式，作为超过 ϵ的概率的函数。

定理（伯恩斯坦）。设 $x_1, ..., x_n$为独立随机变量，满足 $\mathrm{IE}x = 0$, $\mathrm{IE}x^2 = \sigma^2$, 且 $|x_i| \le M$。对于$\epsilon > 0$

$$\mathrm{IP}\left(|\sum_{i=1}^{n} x_i| > \epsilon \right) \le 2e^{-\frac{\epsilon^2}{2n\sigma^2 + \frac{2}{3}\epsilon M}}.$$

证明。
参考Bennet不等式的证明，并注意

$$\phi(z) \ge \frac{z^2}{2 + \frac{2}{3}z}. \ \square$$

备注。 利用伯恩斯坦不等式，可以计算出 ϵ作为超过 ϵ的概率的一个简单表达式

$$\sum_{i=1}^{n} x_i \le \frac{2}{3}uM + \sqrt{2n\sigma^2 u}.$$

大纲。

$$\mathrm{IP}\left(\sum_{i=1}^{n} x_i > \epsilon \right) \le 2e^{-\frac{\epsilon^2}{2n\sigma^2 + \frac{2}{3}\epsilon M}} = e^{-u},$$

其中

$$u = \frac{\epsilon^2}{2n\sigma^2 + \frac{2}{3}\epsilon M}.$$

我们现在解决 ϵ

$$\epsilon^2 - \frac{2}{3}\epsilon M - 2n\sigma^2 \epsilon = 0$$

和

$$\epsilon = \frac{1}{3}uM + \sqrt{\frac{u^2 M^2}{9} + 2n\sigma^2 u}.$$

因为 $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$

$$\epsilon = \frac{2}{3}uM + \sqrt{2n\sigma^2 u}.$$

所以以很大的概率

$$\sum_{i=1}^{n} x_i \le \frac{2}{3}uM + \sqrt{2n\sigma^2 u}. \ \triangle$$

如果我们想要限制

$$|n^{-1}\sum_{i=1}^{n} f(x_i) - \mathrm{IE}f(x)|$$

我们考虑

$$|f(x_i) - \mathrm{IE}f(x)| \le 2M.$$

因此

$$\sum_{i=1}^{n} (f(x_i) - \mathrm{IE}f(x)) \le \frac{4}{3}uM + \sqrt{2n\sigma^2 u}$$

和

$$n^{-1}\sum_{i=1}^{n} f(x_i) - \mathrm{IE}f(x) \leq \frac{4}{3}\frac{uM}{n} + \sqrt{\frac{2\sigma^2 u}{n}}.$$

同样地,

$$\mathrm{IE}f(x) - n^{-1}\sum_{i=1}^{n} f(x_i) \geq \frac{4}{3}\frac{uM}{n} + \sqrt{\frac{2\sigma^2 u}{n}}.$$

在上述界限中

$$\sqrt{\frac{2\sigma^2 u}{n}} \geq \frac{4uM}{n}$$

这意味着 $u \leq \frac{n\sigma^2}{8M^2}$ 因此

$$|n^{-1}\sum_{i=1}^{n} f(x_i) - \mathrm{IE}f(x)| \lesssim \sqrt{\frac{2\sigma^2 u}{n}} \text{对于 } u \lesssim n\sigma^2,$$

这对应于高斯随机变量的尾概率，并由中心极限定理（CLT）条件预测，即$\lim_{n\to\infty} n\sigma^2 \to \infty$.
如果$\lim_{n\to\infty} n\sigma^2 = C$，其中 C是一个固定常数，那么

$$|n^{-1}\sum_{i=1}^{n} f(x_i) - \mathrm{IE}f(x)| \lesssim \frac{C}{n}$$

这对应于泊松随机变量的尾概率.

现在我们来看一个更简单的指数不等式，我们不需要方差的信息。

定理（Hoeffding）。设 $x_1, ..., x_n$是独立随机变量，满足 $\mathrm{IE}x = 0$ 和 $|x_i| \leq M_i$. 对于$\epsilon > 0$

$$\mathrm{IP}\left(|\sum_{i=1}^{n} x_i| > \epsilon\right) \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n} M_i^2}}.$$

证明。

$$\mathrm{IP}\left(\sum_{i=1}^{n} x_i > \epsilon\right) \leq e^{-\lambda\epsilon}\mathrm{IE}e^{\lambda\sum_{i=1}^{n} x_i} = e^{-\lambda\epsilon}\Pi_{i=1}^{n}\mathrm{IE}e^{\lambda x_i}.$$

可以证明（作业问题）

$$\mathrm{IE}(e^{\lambda x_i}) \leq e^{\frac{\lambda^2 M_i^2}{8}}.$$

通过对 λ进行优化，可以证明以下界限

$$e^{-\lambda\epsilon}\Pi_{i=1}^{n} e^{\frac{\lambda^2 M_i^2}{8}}. \ \square$$

将Hoeffding不等式应用于

$$n^{-1}\sum_{i=1}^{n} f(x_i) - \mathrm{IE}f(x)$$

我们可以以概率$1 - e^{-u}$的方式陈述

$$n^{-1}\sum_{i=1}^{n} f(x_i) - \mathrm{IE}f(x) \leq \sqrt{\frac{2M\,u}{n}},$$

这是一个亚高斯分布，类似于中心极限定理，但是没有方差信息，我们永远无法达到当随机变量具有泊松尾分布时所达到的速率。

我们将在后面关于科尔莫哥洛夫链和达德利熵积分的讲座中使用Hoeffding不等式的以下版本。

定理 (Hoeffding).设 $x_1, ..., x_n$是独立随机变量，满足 $\mathrm{IP}(x_i = M_i) = 1/2$和 $\mathrm{IP}(x_i = -M_i) = 1/2$. 对于$\epsilon > 0$

$$\mathrm{IP}\left(|\sum_{i=1}^{n} x_i| > \epsilon \right) \leq 2e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n} M_i^2}}.$$

证明。

$$\mathrm{IP}\left(\sum_{i=1}^{n} x_i > \epsilon \right) \leq e^{-\lambda\epsilon}\mathrm{IE}e^{\lambda\sum_{i=1}^{n} x_i} = e^{-\lambda\epsilon}\Pi_{i=1}^{n}\mathrm{IE}e^{\lambda x_i}.$$

$$\begin{aligned} \mathrm{IE}(e^{\lambda x_i}) &= \frac{1}{2}e^{\lambda M_i} + \frac{1}{2}e^{-\lambda M_i}, \\ \frac{1}{2}e^{\lambda M_i} + \frac{1}{2}e^{-\lambda M_i} &= \sum_{k=0}^{\infty} \frac{(M_i\lambda)^{2k}}{(2k)!} \leq e^{\frac{\lambda^2 M_i^2}{2}}. \end{aligned}$$

对 λ进行以下优化

$$e^{-\lambda\epsilon}\Pi_{i=1}^{n} e^{\frac{\lambda^2 M_i^2}{2}}. \ \square$$

11.4. 鞅不等式

在前一节中，我们提出了一些独立随机变量和的浓度不等式。 现在我们来看看更复杂的独立随机变量函数，并引入一个特殊的鞅不等式来证明浓度。

设 $(\Omega, \mathcal{L}, \mu)$ 是一个概率空间。 设 $x_1, ..., x_n$是Ω上的实随机变量。 设函数 $Z(x_1, ..., x_n) : \Omega^n \to \mathrm{IR}$ 是从随机变量到实数的映射。

如果 $Z(x_1, ..., x_n)$ 和 IE_x 之间的偏差很小，则函数 Z浓度。 $^{1,..,x_n}Z(x_1, .., x_n)$随着 n趋向于无穷大而趋于零。

定理(McDiarmid)。设 $x_1, ..., x_n$为独立随机变量，令 $Z(x_1, ..., x_n) : \Omega^n \to \mathrm{IR}$ 使得

对于任意的$x_1, ..., x_n, x'_1, ..., x'_n$ $|Z(x_1, .., x_n) - Z(x_1, ..., x_{i-1}, x'_i, x_{i+1}, x_n)| \leq c_i,$

则

$$\mathrm{IP}\left(Z - \mathrm{IE}Z > \epsilon\right) \leq e^{-\frac{\epsilon^2}{2\sum_{i=1}^{n} c_i^2}}.$$

证明。

$$\mathrm{IP}(Z - \mathrm{IE}Z > \epsilon) = \mathrm{IP}(e^{\lambda(Z-\mathrm{IE}Z)} > e^{\lambda\epsilon}) \leq e^{-\lambda\epsilon}\mathrm{IE}e^{\lambda(Z-\mathrm{IE}Z)}.$$

我们将使用以下非常有用的分解

$$\begin{aligned} Z(x_1, ..., x_n) - \mathrm{IE}_{x'_1,..,x'_n} Z(x'_1, .., x'_n) &= [Z(x_1, ..., x_n) - E_{x'_1} Z(x'_1, x_2, ..., x_n)] \\ &+ [E_{x'_1} Z(x'_1, x_2, ..., x_n) - E_{x'_1,x'_2} Z(x'_1, x'_2, x_3, ..., x_n)] \\ &+ ... \\ &+ [E_{x'_1,...,x'_{n-1}} Z(x'_1, x'_2, ...x'_{n-1}, x_n) - E_{x'_1,...,x'_n} Z(x'_1, ..., x'_n)]. \end{aligned}$$

我们将随机变量

$$z_i(x_i, ..., x_n) := \mathrm{IE}_{x'_1,...,x'_{i-1}} Z(x'_1, ..., x'_{i-1}, x_i, ..., x_n) - \mathrm{IE}_{x'_1,...,x'_i} Z(x'_1, ..., x'_i, x_{i+1}, ..., x_n),$$

and

$$Z(x_1, ..., x_n) - \mathrm{IE}_{x'_1,..,x'_n} Z(x'_1, .., x'_n) = z_1 + ... + z_n.$$

以下不等式成立（见下面的引理证明）

$$\mathrm{IE}_{x_i} e^{\lambda z_i} \leq e^{\lambda^2 c_i^2/2} \ \ \forall \lambda \in \mathrm{IR}.$$

$$\begin{aligned} \mathrm{IE} e^{\lambda(Z - \mathrm{IE}Z)} &= \mathrm{IE} e^{\lambda(z_1 + ..., + z_n)} \\ \mathrm{IEIE}_{x_1} e^{\lambda(z_1 + ..., + z_n)} &= \mathrm{IE} e^{\lambda(z_2 + ..., + z_n)} \mathrm{IE}_{x_1} e^{\lambda z_1} \\ &\leq \mathrm{IE} e^{\lambda(z_2 + ..., + z_n)} e^{\lambda c_1^2/2}, \end{aligned}$$

通过归纳

$$\mathrm{IE} e^{\lambda(Z - \mathrm{IE}Z)} \leq e^{\lambda^2 \sum_{i=1}^n c_i^2/2}.$$

为了推导出界，我们对 λ进行优化

$$e^{-\lambda\epsilon + \lambda^2 \sum_{i=1}^n c_i^2/2}. \ \square$$

引理。对于所有的 $\lambda \in \mathrm{IR}$

$$\mathrm{IE}_{x_i} e^{\lambda z_i} \leq e^{\lambda^2 c_i^2/2}.$$

证明。

对于任意的 $t \in [-1,1]$，函数 $e^{\lambda t}$关于 λ是凸的。

$$\begin{aligned} e^{\lambda t} &= e^{\lambda(\frac{1+t}{2}) - \lambda(\frac{1-t}{2})} \\ &\leq \frac{1+t}{2} e^{\lambda} + \frac{1-t}{2} e^{-\lambda} \\ &= \frac{e^{\lambda} + e^{-\lambda}}{2} + t \frac{e^{\lambda} - e^{-\lambda}}{2} \\ &\leq e^{\lambda^2/2} + t \, \mathrm{sh}(\lambda). \end{aligned}$$

设 $t = \frac{z_i}{c_i}$ 并注意到 $\frac{z_i}{c_i} \in [-1,1]$ 所以，

$$e^{\lambda z_i} = e^{\lambda c_i \frac{z_i}{c_i}} \leq e^{\lambda^2 c_i^2/2} + \frac{z_i}{c_i} \mathrm{sh}(\lambda c_i),$$

和

$$\mathrm{IE}_{x_i} e^{\lambda z_i} \leq e^{\lambda^2 c_i^2/2}. \ \square$$

例子. 我们可以使用麦克迪尔米德不等式来证明经验最小值的集中性。. 给定一个数据集 $\{v_1 = (x_1, y_1), ..., v_n = (x_n, y_n)\}$经验最小值是

$$Z(v_1, ..., v_n) = \underset{f \in \mathcal{H}_K}{\text{最小值}} \ n^{-1} \sum_{i=1}^n V(f(x_i), y_i).$$

如果损失函数有界，可以证明对于所有 $(v_1, ..., v_n, v_i')$

$$|Z(v_1, ..., v_n) - Z(v_1, ..., v_{i-1}, v_i', ...v_n)| \leq \frac{k}{n}.$$

因此，以概率 $1 - e^{-u}$

$$|Z - \mathrm{I\!E}Z| \leq \sqrt{\frac{2ku}{n}}.$$

因此经验最小值集中。

讲座 12
Vapnik-ˇCervonenkis理论

12.1. 大数定律

在之前的讲座中，我们考虑了单个或固定函数的大数定律。我们将其称为一维集中不等式。现在我们来看一下大数定律的均匀性，即在一类函数上均匀成立的大数定律。

这些均匀极限定理的要点是，如果大数定律对假设空间中的所有函数都成立，则对于经验最小化者也成立。

这一章被称为Vapnik-Cervonenkis理论的原因是他们提供了研究这些类的基本工具。

12.2. 一个函数的泛化界

在研究均匀结果之前，我们证明当假设空间 $\mathcal{H}$只包含一个函数 f_1时的泛化结果。

在这种情况下，经验风险最小化者是 f_1

$$f_1 = f_S := \arg\min_{f\in\mathcal{H}} \left[n^{-1} \sum_{i=1}^{n} V(f, z_i) \right].$$

定理。给定 $0 \le V(f_1, z) \le M$对于所有 z和 $S = \{z_i\}^{ni}{}_{=1}$独立同分布抽取，则以概率至少 $1 - e^{-t}$ $(t > 0)$

$$\mathrm{IE}_z V(f_1, z) \le n^{-1} \sum_{i=1}^{n} V(f_1, z_i) + \sqrt{\frac{M^2 t}{n}}.$$

证明。

根据Hoeffding不等式

$$\mathrm{IP}\left(\mathrm{IE}_z V(f_1, z) - n^{-1} \sum_{i=1}^{n} V(f_1, z_i) > \varepsilon \right) \le e^{-n\varepsilon^2/M^2}$$

所以

$$\mathrm{IP}\left(\mathrm{IE}_z V(f_1, z) - n^{-1} \sum_{i=1}^{n} V(f_1, z_i) \le \varepsilon \right) > 1 - e^{-n\varepsilon^2/M^2}.$$

设 $t = n\varepsilon^2/M^2$，结论得证。 □

12.3. 有限个函数的泛化界

现在我们来看一下在假设空间 $\mathcal{H}$上的ERM情况，其中函数的个数是有限的， $k = |\mathcal{H}|$。 在这种情况下，经验最小化器将是 k个函数之一。

定理。给定 $0 \leq V(f_j, z) \leq M$对于所有 $f_j \in \mathcal{H}, z$和 $S = \{z_i\}^{ni}{}_{=1}$独立同分布抽取的情况下，对于经验最小化器，至少以概率 $1 - e^{-t}$ $(t > 0)$成立，

$$\mathrm{IE}_z V(f_S, z) < n^{-1} \sum_{i=1}^{n} V(f_S, z_i) + \sqrt{\frac{M^2(\log K + t)}{n}}.$$

证明。

以下事件的蕴含关系成立

$$\left\{ \max_{f_j \in \mathcal{H}} \ \mathrm{IE}_z V(f_j, z) - n^{-1} \sum_{i=1}^{n} V(f_j, z_i) < \varepsilon \right\} \Rightarrow \left\{ \mathrm{IE}_z V(f_S, z) - n^{-1} \sum_{i=1}^{n} V(f_S, z_i) < \varepsilon \right\}.$$

$$\begin{aligned} & \mathrm{IP}\left(\max_{f_j \in \mathcal{H}} \ \mathrm{IE}_z V(f_j, z) - n^{-1} \sum_{i=1}^{n} V(f_j, z_i) \geq \varepsilon \right) \\ = \ & \mathrm{IP}\left(\bigcup_{f \in \mathcal{H}} \left\{ \mathrm{IE}_z V(f, z) - n^{-1} \sum_{i=1}^{n} V(f, z_i) \geq \varepsilon \right\} \right) \\ \leq \ & \sum_{f_j \in \mathcal{H}} \mathrm{IP}\left(\mathrm{IE}_z V(f_j, z) - n^{-1} \sum_{i=1}^{n} V(f_j, z_i) \geq \varepsilon \right) \\ \leq \ & k e^{-n\varepsilon^2/M^2}, \end{aligned}$$

最后一步来自我们的单函数结果。 设 $e^{-t} = k e^{-n\varepsilon^2/M^2}$，结论就得出了。 □

12.4. 紧致假设空间的泛化界

我们现在证明了具有无限个函数的假设空间的泛化的充分条件，并给出了一些例子。

我们首先假设我们的假设空间是连续函数空间的子集，$\mathcal{H} \subset \mathcal{C}(\mathcal{X})$。

定义。 度量空间是紧致的当且仅当它是完全有界和完备的。

定义。设 R为度量空间， ϵ为任意正数。 如果对于集合 $M \subset R$中的每个元素 x，存在至少一个点 $a \in A$使得 $\rho(x, a) < \epsilon$，则集合 $A \subset R$被称为 M的 ϵ-网。 这里 $\rho(\cdot, \cdot)$是一种范数。

定义。 给定度量空间 R和子集 $M \subset R$，假设对于每个$\epsilon > 0$， M都有一个有限的ϵ-网。. 则 M被称为完全有界。

命题。 紧致空间对于所有$\epsilon > 0$都有一个有限的 ϵ-网。

在本节的剩余部分中，我们将使用上确界范数，$\rho(a, b) = \mathrm{s}$

$$\mathrm{up}。 \qquad \sup_{x \in \mathcal{X}} |a(x) - b(x)|.$$

定义. 给定一个假设空间 $\mathcal{H}$和超范数，覆盖数 $\mathcal{N}(\mathcal{H}, \epsilon)$是最小的数 $\ell \in \mathrm{IN}$，使得对于每个 $f \in \mathcal{H}$都存在函数 $\{g_i\}_{i=1}^{\ell}$，使得

$$\sup_{x\in\mathcal{X}} |f(x) - g_i(x)| \leq \epsilon \text{对于某个 } i.$$

现在我们给出这种情况的一个泛化界。 在界中我们假设 $V(f, z) = (f(x) - y)^2$但是结果可以很容易地扩展到任何Lipschitz损失

$$|V(f_1, z) - V(f_2, z)| \leq C||f_1(x) - f_2(x)||_\infty \ \forall z.$$

定理。 设 $\mathcal{H}$是 $\mathcal{C}(\mathcal{X})$的一个紧致子集。. 给定 $0 \leq |f(x) - y| \leq M$对于所有 $f \in \mathcal{H}$, z和 $S = \{z_i\}^{ni}{}_{=1}$独立同分布地抽取，则对于经验最小化函数，至少以概率 $1 - e^{-t}$ $(t > 0)$ 有

$$\mathrm{IE}_{x,y}(f_S(x) - y)^2 < n^{-1}\sum_{i=1}^{n}(f_S(x_i) - y_i)^2 + \frac{\sqrt{M^2(\log \mathcal{N}(\mathcal{H}, \varepsilon/8M) + t)}}{n}.$$

我们首先证明两个有用的引理。定义

$$D(f, S) := \mathrm{IE}_{x,y}(f(x) - y)^2 - n^{-1}\sum_{i=1}^{n}(f(x_i) - y_i)^2.$$

引理。 如果 $|f_j(x) - y| \leq M$对于 $j = 1, 2$，那么

$$|D(f_1, S) - D(f_2, S)| \leq 4M||f_1 - f_2||_\infty.$$

证明。 请注意

$$(f_1(x) - y)^2 - (f_2(x) - y)^2 = (f_1(x) - f_2(x))(f_1(x) + f_2(x) - 2y)$$

所以

$$\begin{aligned} |\mathrm{IE}_{x,y}(f_1(x) - y)^2 - \mathrm{IE}_{x,y}(f_2(x) - y)^2| &= \left|\int (f_1(x) - f_2(x))(f_1(x) + f_2(x) - 2y)d\mu(x, y)\right| \\ &\leq ||f_1 - f_2||_\infty \int |f_1(x) - y + f_2(x) - y|du(x, y) \\ &\leq 2M||f_1 - f_2||_\infty, \end{aligned}$$

and

$$\begin{aligned} |n^{-1}\sum_{i=1}^{n}[(f_1(x_i) - y_i)^2 - (f_2(x_i) - y_i)^2]| &= n^{-1}\left|\sum_{i=1}^{n}(f_1(x_i) - f_2(x_i))(f_1(x_i) + f_2(x_i) - 2y_i)\right| \\ &\leq ||f_1 - f_2||_\infty \frac{1}{n}\sum_{i=1}^{n}|f_1(x_i) - y_i + f_2(x_i) - y_i| \\ &\leq 2M||f_1 - f_2||_\infty. \end{aligned}$$

结果由上述不等式得出。 □

引理。 设 $\mathcal{H} = B_1 \bigcup ... \bigcup B_\ell$ 且 $\varepsilon > 0$。则

$$\mathrm{IP}\left(\sup_{f\in\mathcal{H}} D(f, S)\right) \leq \sum_{j=1}^{\ell}\mathrm{IP}\left(\sup_{f\in B_j} D(f, S)\right).$$

证明。

结果由以下等价性和并集界得出

$$\sup_{f\in\mathcal{H}} D(f,\,S)\geq\varepsilon \Longleftrightarrow \exists j\leq \ell \text{ s.t. } \sup_{f\in B_j} D(f,\,S)\geq\varepsilon.\ \square$$

我们现在证明定理12.4。

设 $\ell=\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{4M}\right)$ 且函数 $\{g_j\}_{j=1}^{\ell}$ 具有以下性质：以 f_j 为中心，半径为 $\frac{\varepsilon}{4M}$ 覆盖 $\mathcal{H}$. 根据第一个引理，对于所有 $f\in B_j$

$$|D(f,\,S)-D(f_j,\,S)|\leq 4M||f-f_j||_\infty\leq 4M\frac{\varepsilon}{4M}=\varepsilon,$$

这意味着对于所有 $f\in B_j$

$$\sup_{f\in B_j}|D(f,\,S)|\geq 2\varepsilon\Rightarrow|D(f_j,\,S)|\geq\varepsilon.$$

所以

$$\mathrm{IP}\left(\sup_{f\in B_j}|D(f,\,S)|\geq 2\varepsilon\right)\leq \mathrm{IP}\left(|D(f_j,\,S)|\geq\varepsilon\right)\leq 2e^{-\varepsilon^2 n/M^2}.$$

这与第二个引理结合起来意味着

$$\mathrm{IP}\left(\sup_{f\in\mathcal{H}} D(f,\,S)\right)\leq\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{8M}\right)e^{-\varepsilon^2 n/M^2}.$$

由于以下事件的蕴含成立

$$\left\{\sup_{f\in\mathcal{H}}\ \mathrm{IE}_z V(f_j,\,z)-n^{-1}\sum_{i=1}^{n}V(f_j,\,z_i)<\varepsilon\right\}\Rightarrow\left\{\mathrm{IE}_z V(f_S,\,z)-n^{-1}\sum_{i=1}^{n}V(f_S,\,z_i)<\varepsilon\right\}$$

通过设置 $e^{-t}=\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{8M}\right)e^{-n\varepsilon^2/M^2}$. $\square$ 来获得结果

上述定理的一个结果是ERM的一致收敛和一致性的以下充分条件。

推论。 对于所有$\varepsilon>0$，ERM是一致的

$$\lim_{n\to\infty}\frac{\log\mathcal{N}\left(\mathcal{H},\,\varepsilon\right)}{n}=0.$$

证明。

这直接来自于陈述

$$\mathrm{IP}\left(\sup_{f\in\mathcal{H}} D(f,\,S)\right)\leq\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{8M}\right)e^{-\varepsilon^2 n/M^2}.\ \square$$

现在我们计算几种假设空间的覆盖数。

我们还需要定义包装数。

定义。给定一个假设空间 $\mathcal{H}$和$sup\mathbf{no}rm$，如果 ℓ函数 $\{g_i\}_{i=1}^{\ell}$是 ϵ分离的，那么

$$\sup_{x\in\mathcal{X}}|g_j(x)-g_i(x)|>\epsilon\ \ \forall i=j.$$

装箱数 $\mathcal{P}(\mathcal{H},\,\epsilon)$是 ϵ-分离集合的最大基数。

装箱数和覆盖数之间的以下关系非常有用。

引理。 给定度量空间 (A, ρ)。. 那么对于所有的$\epsilon > 0$和每个 $W \subset A$，
覆盖数和装箱数满足

$$\mathcal{P}(W, 2\epsilon, \rho) \leq \mathcal{N}(W, \epsilon, \rho) \leq \mathcal{P}(W, \epsilon, \rho)。$$

证明。

对于第二个不等式，假设 P是最大基数的 ϵ-装箱集合，
$\mathcal{P}(W, \epsilon, d)$。 那么对于任意的 $w \in W$，必须存在一个 $u \in P$使得 $\rho(u, w) < \epsilon$，否则
w不是 P的元素，且 $P \cup w$是一个 ϵ-装箱集合。 这与假设相矛盾
即 P是最大的 ϵ-装箱集合。 因此，任何最大的 ϵ-装箱集合都是一个 ϵ-覆盖集合。

对于第一个不等式，假设 C是 W的一个 ϵ-覆盖，并且 P是W的一个最大基数为P(W, □, d)的2ϵ-packing。 我们将证明 $|\mathcal{P}| \leq |C|$。
假设 $|C| > |P|$。 那么对于两个点 $w_1, w_2 \in P$和一个点 $u \in C$，以下成立

$$\rho(w_1, u) \leq \epsilon \text{和 } \rho(w_2, u) \leq \epsilon \Longrightarrow \rho(w_1, w_2) \leq 2\epsilon。$$

这与 P中的点是2ϵ-分离的事实相矛盾。□一般来说，我们将计算假设空间的packing数，并使用上述引理得到覆盖数。

以下命题将会很有用。

命题。 给定 $x \in \mathrm{IR}^d$，将空间限制在单位球 $B = \{x : ||x|| \leq M\}$上，并使用标准欧几里得度量 $\rho(x, y) = ||x - y||$，那么对于 $\epsilon \leq M$

$$\mathcal{P}(B, \epsilon, \rho) \leq \left(\frac{3M}{\epsilon}\right)^d.$$

证明。

点 ℓ $w_1, .., w_\ell$构成了一个最优 ϵ-包装，所以

$$\begin{aligned}
\mathrm{Vol}\left(M + \frac{\epsilon}{2}\right) &= C_d \left(M + \frac{\epsilon}{2}\right)^d \\
\mathrm{Vol}\left(\frac{\epsilon}{2}\right) &= C_d \left(\frac{\epsilon}{2}\right)^d \\
\ell[C_d \left(\frac{\epsilon}{2}\right)^d] &= C_d \left(M + \frac{\epsilon}{2}\right)^d \\
\ell &\leq \left(\frac{2M + \epsilon}{\epsilon}\right)^d \\
&\leq \left(\frac{3M}{\epsilon}\right)^d \text{ 对于所有 } \epsilon \leq M. \quad \square
\end{aligned}$$

例子。 有限维有界$RKHS$的覆盖数。

对于一个有限维有界$RKHS$

$$\mathcal{H}_K = \left\{ f : f(x) = \sum_{p=1}^{m} c_p \phi_p(x) \right\},$$

$with\ \|f\|_K^2 \leq M.$

通过再生性质和柯西-施瓦茨不等式，可以通过$RKHS$范数来限制$supnorm$：

$$\begin{aligned} ||f(\mathbf{x})||_\infty &= ||\langle K(\mathbf{x},\cdot), f(\cdot)\rangle_K||_\infty \\ &\leq ||K(\mathbf{x},\cdot)||_K||f||_K \\ &= \sqrt{\langle K(x,\cdot), K(x,\cdot)\rangle}||f||_K \\ &= \sqrt{K(\mathbf{x},\mathbf{x})}||f||_K \\ &\leq \kappa||f||_K \end{aligned}$$

这意味着如果我们可以用$RKHS$范数覆盖，我们也可以用$supnorm$覆盖。

我们封面中的每个函数， $\{g_i\}_{i=1}^{\ell}$ 都可以写成

$$g_i(x) = \sum_{p=1}^{m} d_{ip}\phi_p(x)$$

所以如果我们找到 ℓ向量 d_i，对于所有 c： $\sum_{p=1}^{m} \frac{c_p^2}{\lambda_p} \leq M$存在一个 d_i使得

$$\sum_{p=1}^{m} \frac{(c_p - d_{ip})^2}{\lambda_p} < \epsilon^2,$$

我们有一个尺度为 ϵ的封面。 上述问题简化为使用欧几里得度量覆盖半径为 M的球的问题。 使用命题*12.4*，我们可以用$\mathrm{R}KHS$范数和$supnor^m$来限制装箱数。

$$\begin{aligned} \mathcal{P}(\mathcal{H}, \epsilon, ||\cdot||_{\mathcal{H}_k}) &\leq \left(\frac{3M}{\epsilon}\right)^m, \\ \mathcal{P}(\mathcal{H}, \epsilon, ||\cdot||_\infty) &\leq \left(\frac{3M}{\kappa\epsilon}\right)^m. \end{aligned}$$

使用引理*12.4*，我们可以得到对覆盖数的界限

$$\mathcal{N}(\mathcal{H}, \epsilon, ||\cdot||_\infty) \leq \left(\frac{3M}{\kappa\epsilon}\right)^m.$$

我们已经证明了对于$\mathcal{H} \subset \mathcal{C}(\mathcal{X})$这是紧致的，我们有一致收敛性。 这个要求太严格了，无法确定必要条件。 这些条件不适用于一大类函数，如指示函数 $f(x)\in\{0,1\}$。

12.5. 指示函数假设空间的泛化界限

在本节中，我们推导出一致收敛的必要和充分条件，从而得到指示函数的泛化界限，$f(x)\in\{0,1\}$。

与紧致函数的情况一样，我们将取一个指示函数类$\mathcal{H}$并将其缩减为一些有限的函数集。 对于指示函数的情况，我们通过增长函数的概念来完成，现在我们来定义它。

定义。 给定一组 n个点 $\{x_i\}^{ni}{}_{=1}$和一个指示函数类 $\mathcal{H}$如果函数 $f\in \mathcal{H}$能够选择出 $\{x_i\}^{ni}{}_{=1}$的某个子集，则我们称该函数 $f\in$ H选择出了该子集。. 可以选择出的子集数量的基数被称为增长函数：

$$\triangle个(\mathcal{H}, \{x_i\}^{ni}{}_{=1}) = \#\{f\cap \{x_i\}^{ni}{}_{=1} : f\in \mathcal{H}\}。$$

现在我们将陈述一个引理，它看起来很像紧致或有限维情况的概括结果。

引理。假设 $\mathcal{H}$是一个指示函数类， $S = \{z_i\}^{ni}{}_{=1}$是独立同分布抽取的样本，则对于经验最小化函数 f_S，至少以概率 $1 - e^{-t/8}$ $(t > 0)$成立。

$$\mathrm{IE}_{x,y} I_{\{f_S(x)=y\}} < n^{-1}\sum_{i=1}^{n} I_{\{f_S(x_i)=y_i\}} + \sqrt{\frac{(8\log 8\triangle_n(\mathcal{H}, S) + t)}{n}},$$

其中 $\triangle_n(\mathcal{H}, S)$是给定 S观测的增长函数。

注意：上述结果取决于通过增长函数抽取的$2n$样本的特定抽取。 我们将很快消除这种依赖关系。

我们首先证明两个有用的引理。定义

$$D(f, S) := \mathrm{IE}_{x,y} I_{\{f(x)=y\}} - n^{-1}\sum_{i=1}^{n} I_{\{f(x_i)=y_i\}}。$$

第一个引理基于对称化的思想，将经验误差和期望误差之间的偏差替换为两个经验误差之间的差异。

引理。 给定两个独立的数据副本 $S = \{z_i\}^{ni}{}_{=1}$和 $S = \{z_i'\}^{ni}{}_{=1}$
对于任意固定的 $f\in \mathcal{H}$如果 $n \geq 2/\epsilon^2$

$$\mathrm{IP}\left(|D(f, S)| > \epsilon\right) \leq 2\,\mathrm{IP}\left(|D(f, S) - D(f, S')| > \epsilon/2\right),$$

其中

$$|D(f, S) - D(f, S')| = n^{-1}\sum_{i=1}^{n} I_{\{f(x_i)=y_i\}} - n^{-1}\sum_{i=1}^{n} I_{\{f(x_i')=y_i'\}}.$$

证明。 我们首先假设

$$\mathrm{IP}(|D(f, S')| \leq \epsilon/2 \,|\, S) \geq 1/2,$$

我们已经给出了条件 S. 由于 S和 S'是独立的，我们可以积分消去

$$1/2\,\mathrm{IP}(|D(f, S')| > \epsilon) \leq \mathrm{IP}(|D(f, S')| \leq \epsilon/2, |D(f, S')| > \epsilon).$$

根据三角不等式 $|D(f, S)|> \epsilon$和 $|D(f, S')| \leq \epsilon/2$ 可以推出

$$|D(f, S) - D(f, S')| \geq \epsilon/2,$$

所以

$$\mathrm{IP}(|D(f, S')| \leq \epsilon/2, |D(f, S')| > \epsilon) \leq \mathrm{IP}(|D(f, S) - D(f, S')| \geq \epsilon/2).$$

为了完成证明，我们需要证明我们的初始假设成立。 由于 $\mathcal{H}$是一个指示函数类，和式中的元素是二项随机变量，其中的元素的方差最多为$1/4$ n。 因此，根据 Bie$n$aym´e-Chebyshev不等式

$$\mathrm{IP}(|D(f, S')| > \epsilon/2) \geq 1/4n\epsilon^2,$$

这意味着当 $n \geq 2/\epsilon^2$时，初始假设成立。 □

通过对称化，我们现在有一个只依赖于样本的项。 问题在于它不仅依赖于我们拥有的样本，还依赖于一个独立的复制。
通过第二步的对称化，这个麻烦被消除了。

引理。 设 σ_i为一个$R\mathbf{a}d\mathbf{ema}ch\mathbf{e}r$随机变量（$\mathrm{IP}(\sigma_i = \pm 1) = 1/2$）则

$$\mathrm{IP}\left(|D(f,S) - D(f,S')| > \epsilon/2\right) \leq 2\mathrm{IP}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i)=y_i\}}\right| > \epsilon/4\right).$$

证明。

$$\begin{aligned}
\mathrm{IP}\left(|D(f,S) - D(f,S')| > \epsilon/2\right) &= \mathrm{IP}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i)=y_i\}} - n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i')=y_i'\}}\right| > \epsilon/2\right) \\
&\leq \mathrm{IP}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i)=y_i\}}\right| > \epsilon/4\right) + \\
&\quad \mathrm{IP}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i')=y_i'\}}\right| > \epsilon/4\right) \\
&\leq 2\mathrm{IP}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i)=y_i\}}\right| > \epsilon/4\right). \ \square
\end{aligned}$$

第二个对称化步骤允许我们基于仅在经验数据上计算的数量来限制经验和期望误差之间的偏差。

我们现在证明引理12.5。
通过对称化引理，对于 $n \geq 8/\epsilon^2$

$$\mathrm{IP}(|D(f,S)| > \epsilon) \leq 4\,\mathrm{IP}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i)=y_i\}}\right| > \epsilon/4\right).$$

通过Hoeffdings不等式的Rademacher版本

$$\mathrm{IP}\left(\left|n^{-1}\sum_{i=1}^{n}\sigma_i I_{\{f(x_i)=y_i\}}\right| > \epsilon\right) \leq 2e^{-2\epsilon^2}.$$

将上述结果结合起来对于单个函数

$$\mathrm{IP}(|D(f,S)| > \epsilon) \leq 8e^{-\epsilon^2/8}.$$

给定数据 S，增长函数描述了可以“挑选出”的子集的基数，这是可能的标记或可实现函数的数量的上界， $\ell = \triangle_n(\mathcal{H}, S)$. 我们通过 f_j索引可能的标记，其中$j = 1, ..., \ell$.

我们现在按照有限个函数的情况进行

$$\begin{aligned}\mathrm{IP}\left(\sup_{f\in\mathcal{H}}|D(f,S)|\geq\epsilon\right) &= \mathrm{IP}\left(\bigcup_{f\in\mathcal{H}}|D(f,S)|\geq\epsilon\right)\\ &\leq \sum_{i=1}^{\ell}\mathrm{IP}\left(|D(f_j,S)|\geq\varepsilon\right)\\ &\leq 8\triangle_n(\mathcal{H},S)e^{-n\epsilon^2/8}.\end{aligned}$$

设置 $e^{-t/8}= 8\triangle_n(\mathcal{H},S)e^{-n\epsilon^2/8}$完成证明。 □

这个界限不是均匀的，因为增长函数取决于数据 S。 我们可以通过定义一个均匀的增长函数来使界限变得均匀。

定义。 均匀增长函数是

$$\triangle_n(\mathcal{H})=\max_{x_1,\ldots,x_n\in\mathcal{X}}\triangle_n(\mathcal{H},\{x_i\}_{i=1}^n).$$

推论。 设 $\mathcal{H}$为指示函数的类别， $S=\{z_i\}^{ni}{}_{=1}$独立同分布地抽取，那么对于经验最小化器，以至少 $1-e^{-t/8}$ $(t>0)$的概率，有 f_S,

$$\mathrm{IE}_{x,y}I_{\{f_S(x)=y\}}<n^{-1}\sum_{i=1}^{n}I_{\{f_S(x_i)=y_i\}}+\sqrt{\frac{(\log 8\triangle_n(\mathcal{H})+t)}{n}},$$

其中 $\triangle_n(\mathcal{H})$是均匀增长函数。

推论。对于一类指示函数，ERM是一致的，当且仅当对于所有$\varepsilon>0$

$$\lim_{n\to\infty}\frac{8\log\triangle_n(\mathcal{H})}{n}=0.$$

我们现在描述一下均匀增长函数多项式增长的条件。 为了做到这一点，我们需要一些定义。

定义。 一个假设空间 $\mathcal{H}$，如果它的每个 2_n子集都可以被 $\mathcal{H}$ "挑出"，则它破碎了集合 $\{x_1,...,x^n\}$。. *Vap***ni***k-Cer***v***onen***k***is (VC)*维度， $v(\mathcal{H})$，是假设空间中最大的 n，使得所有大小为 n的集合都被 $\mathcal{H}v(\mathcal{H})=\sup\{n:\triangle_n(\mathcal{H})=2^n\}$，如果不存在

这样的 n，则VC维度是无穷大。

定义。指示函数的假设空间 $\mathcal{H}$是一个VC类，当且仅当它具有有限的VC维度。

例子。

VC维度通过以下引理控制增长函数。

引理。 对于具有VC维度 d的假设空间 $\mathcal{H}$和$n>d$

$$\triangle_n(\mathcal{H})\leq\sum_{i=1}^{d}\binom{n}{i}.$$

证明。

证明将通过对 $n+d$进行归纳。我们定义 $\binom{n}{i} := 0$，如果$i<0$或$i>n$。此外，可以验证

$$\binom{n}{i} = \binom{n-1}{i-1} + \binom{n-1}{i}.$$

当 $d=0$ $|\mathcal{H}|=1$，因为不能破坏任何点，所以对于所有 n

$$\triangle_n(\mathcal{H}) = 1 = \binom{n}{0} = \Phi_0(n)。$$

当 $n=0$时，只有一种方法可以标记0个例子，所以

$$\triangle_0(\mathcal{H}) = 1 = \sum_{i=1}^{d} \binom{0}{i} = \Phi_d(0).$$

假设引理对于 n', d'成立，使得 $n'+d' < n+d$.

我们现在定义三个假设空间 $\mathcal{H}_0, \mathcal{H}_1,$和 $\mathcal{H}_2$:

$$\begin{aligned}
\mathcal{H}_0 &:= \{f_i : i = 1, ..., \triangle_n(\mathcal{H}, S)\} \\
\mathcal{H}_1 &:= \{f_i : i = 1, ..., \triangle_{n-1}(\mathcal{H}, S_n)\} \\
\mathcal{H}_2 &:= \mathcal{H}_0 - \mathcal{H}_1,
\end{aligned}$$

其中，集合 $\mathcal{H}_0$中的每个 f_i都是 S通过 $\mathcal{H}$的可能标记，集合 $\mathcal{H}_1$中的每个 f_i都是 S_n通过 $\mathcal{H}$的可能标记。

对于集合 $\mathcal{H}_1$上的 S_n： $n_1 = n-1$，因为样本减少了一个，且 $v(\mathcal{H}_1) \le d$，因为减少假设的数量不能增加VC维度。

对于集合 $\mathcal{H}_2$上的 S_n： $n_1 = n-1$，因为样本减少了一个，且$v(\mathcal{H}_2) \le d-1$。如果 $S' \subseteq S_n$被 $\mathcal{H}_2$击碎，则 S'的所有标记必须同时出现在 $\mathcal{H}_1$和 $\mathcal{H}_2$中，但在 x_n上的标记不同。因此， $S' \cup \{x_n\}$的基数为 $|S'|+1$，被 $\mathcal{H}$击碎，因此 $|S'|$不能超过 $d-1$。

通过归纳 $\triangle_{n-1}(\mathcal{H}_1, S_n) \le \Phi_d(m-1)$ 和 $\triangle_{n-1}(\mathcal{H}_2, S_n) \le \Phi_{d-1}(m-1)$。

通过构造

$$\begin{aligned}
\triangle_n(\mathcal{H}, S) &= |\mathcal{H}_1| + |\mathcal{H}_2| = \triangle_{n-1}(\mathcal{H}_1, S_n) + \triangle_{n-1}(\mathcal{H}_2, S_n) \\
&\le \Phi_d(n-1) + \Phi_{d-1}(n-1) \\
&= \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} \\
&= \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d} \binom{n-1}{i-1} \\
&= \sum_{i=0}^{d} \left[\binom{n-1}{i} + \binom{n-1}{i-1} \right] \\
&= \sum_{i=0}^{d} \binom{n}{i}. \ \square
\end{aligned}$$

引理。对于 $n \ge d \ge 1$

$$\sum_{i=1}^{d} \binom{n}{i} < \left(\frac{en}{d}\right)^d.$$

证明。

对于 $0 \le i \le d$和 $n \ge d$

$$(m/d)^d (d/m)^i \ge 1,$$

所以

$$\sum_{i=1}^{d} \binom{n}{i} \le (n/d)^d \sum_{i=1}^{d} \binom{n}{i} (d/n)^i \le (n/d)^d (1 + d/n)^n < (ne/d)^d. \ \square$$

现在我们可以用VC维度来表达泛化界限。

定理。 假设 $\mathcal{H}$是一个具有VC维度 d的指示函数类， $S = \{z_i\}^{ni}_{=1} d\mathbf{r}awn\ i.i.d.$，那么对于经验最小化函数 f_S，至少有概率 $1 - e^{-t/8}$ $(t > 0)$满足以下条件：

$$\mathrm{IE}_{x,y} I_{\{f_S(x)=y\}} < n^{-1} \sum_{i=1}^{n} I_{\{f_S(x_i)=y_i\}} + 2 \sqrt{\frac{(8d \log(8en/d) + t)}{n}}.$$

证明。 从引理的证明我们有

$$\mathrm{IP}\left(\sup_{f \in \mathcal{H}} |D(f, S)| \ge \epsilon \right) \le 8 \triangle_n(\mathcal{H}, S) e^{-n\epsilon^2/8},$$

因此，由于 $\triangle_n(\mathcal{H}, S) \le \binom{en}{d}^d$, 我们有

$$\mathrm{IP}\left(\sup_{f \in \mathcal{H}} |D(f, S)| \ge \epsilon \right) \le 8 \left(\frac{en}{d} \right)^d e^{-n\epsilon^2/8},$$

并且设置 $e^{-t/8} = 8 \binom{en}{d}^d e^{-n\epsilon^2/8}$ 给我们

$$\mathrm{IE}_{x,y} I_{\{f_S(x)=y\}} < n^{-1} \sum_{i=1}^{n} I_{\{f_S(x_i)=y_i\}} + \sqrt{\frac{(8d \log(8en/d) + t + 8 \log 8)}{n}},$$

对于$n > 2$ 和$d > 1$ $8 \log 8 < 8d \log(en/d)$ 所以

$$\sqrt{\frac{(8d \log(8en/d) + t + 8 \log 8)}{n}} < 2 \sqrt{\frac{(8d \log(8en/d) + t)}{n}},$$

这证明了定理。 $\square$

定理。 对于一类指示函数ERM，以下是等价的

(1)ERM是一致的

(2)对于所有$\varepsilon > 0$

$$\lim_{n \to \infty} \frac{8 \log \triangle_n(\mathcal{H})}{n} = 0.$$

(3) VC维度 $v(\mathcal{H})$是有限的。

12.6. 科尔莫哥洛夫链

在本节中，我们介绍科尔莫哥洛夫链，这是一种更高效的构建覆盖的方法。 在这个过程中，我们推导出达德利的熵积分。我们首先定义一个经验范数。

定义。 给定 $S = \{x_1, ..., x_n\}$，经验 $\ell_2\mathbf{no}rm$ 是

$$\rho_S(f, g) = \left(n^{-1} \sum_{i=1}^{n} (f(x_i) - g(x_i))^2 \right)^{1/2}.$$

我们可以根据经验范数定义一个覆盖

定义。给定假设空间 $\mathcal{H}$和上述范数，覆盖数
$\mathcal{N}(\mathcal{H}, \epsilon, \rho_S)$是最小的数 $\ell \in \mathrm{IN}$，使得对于每个 $f\in \mathcal{H}$都存在
函数 $\{g_i\}_{i=1}^{\ell}$ *such that*

$$\rho_S(f, g_i) \leq \epsilon \text{对于某个 } i.$$

以下定理的证明与引理12.5的证明完全相同。

定理. 给定平方损失和函数 $\mathcal{H}$，使得 $-1 \leq f(x) \leq 1$，
$y \in [-1, 1]$， $S = \{z_i\}^{ni}{}_{=1}$独立同分布地抽取，则对于经验最小化器fS，至少以概率 $1 - e$ $(t > 0)$ 成立,

$$\mathrm{IE}_{x,y}(f_S(x) - y)^2 < n^{-1} \sum_{i=1}^{n} (f_S(x_i) - y_i)^2 + \sqrt{\frac{(8 \log \mathcal{N}(\mathcal{H}, \varepsilon/8M, \rho_S) + t)}{n}},$$

其中 $\mathcal{N}(\mathcal{H}, \varepsilon/8M, \rho_S)$是经验覆盖.

引理12.5和上述定理证明的关键思想是

$$\mathrm{IP}(|D(f, S)| > \epsilon) \leq 4\,\mathrm{IP}\left(\left| n^{-1} \sum_{i=1}^{n} \sigma_i f(x_i) \right| > \epsilon/4 \right),$$

其中

$$D(f, S) := \mathrm{IE}_{x,y}(f(x) - y)^2 - n^{-1} \sum_{i=1}^{n} (f(x_i) - y_i)^2,$$

并且 σ_i是一个Rademacher随机变量。

我们现在证明链式定理。

定理。 给定一个假设空间 $\mathcal{H}$，对于所有的 $f\in \mathcal{H}$ $-1 \leq f(x) \leq 1$，如果我们定义

$$R(f) = n^{-1} \sum_{i=1}^{n} \sigma_i f(x_i),$$

则

$$\mathrm{IP}\left(\forall f \in \mathcal{H},\ \ R(f) \leq \frac{2^{9/2}}{\sqrt{n}} \int_0^{d(0,f)} \log^{1/2} \mathcal{P}(\mathcal{H}, \varepsilon, \rho_S) d\varepsilon + 2^{7/2} d(0, f) \sqrt{\frac{u}{n}} \right) \geq 1 - e^{-u},$$

其中 $\mathcal{P}(\mathcal{H}, \varepsilon, \rho_S)$是经验包装数

$$\int_0^{d(0,f)} \log^{1/2} \mathcal{P}(\mathcal{H}, \varepsilon, \rho_S) d\varepsilon$$

是达德利的熵积分。

证明。

不失一般性，我们假设零函数 $\{0\}$在 $\mathcal{H}$中。
我们将构造一系列嵌套的函数集合

$$\{0\} = \mathcal{H}_0 \subseteq \mathcal{H}_1 \subseteq \mathcal{H}_2 ... \subseteq \mathcal{H}_j \subseteq ... \mathcal{H}。$$

这些子集具有以下特性

(1) $\forall f, g \in \mathcal{H}_j \;\; \rho_S(f, g) > 2^{-j}$
(2) $\forall f \in \mathcal{H} \;\; \exists \;\; f \in \mathcal{H}_j$ such that $\rho_S(f, g) \leq 2^{-j}$.

给定一个集合 $\mathcal{H}_j$，我们可以通过以下步骤构建 $\mathcal{H}_{j+1}$：

(1) $\mathcal{H}_{j+1} := \mathcal{H}_j$
(2) 找到所有 $f \in \mathcal{H}$，使得对于所有 $g \in \mathcal{H}_{j+1}$， $\rho_S(f, g) > 2^{-(j+1)}$
(3) 将上述 f添加到 $\mathcal{H}_{j+1}$中。

现在我们定义一个投影操作 $\pi_j : \mathcal{H} \to \mathcal{H}_j$，其中给定 $f \in \mathcal{H}$

$$\pi_j(f) := g，\text{其中 } g \in \mathcal{H}_j，\text{使得 } \rho_S(g, f) \leq 2^{-j}。$$

对于所有 $f \in \mathcal{H}$，以下链接成立

$$\begin{aligned} f &= \pi_0(f) + (\pi_1(f) - \pi_0(f)) + (\pi_2(f) - \pi_1(f)) + ... \\ &= \sum_{j=1}^{\infty} (\pi_j(f) - \pi_{j-1}(f)), \end{aligned}$$

和

$$\begin{aligned} \rho_S(\pi_{j-1}(f), \pi_j(f)) &\leq \rho(\pi_{j-1}(f), f) + \rho_S(\pi_j(f), f) \\ &\leq 2^{-(j-1)} + 2^{-j} = 3 \cdot 2^{-j} \leq 2^{-j+2}. \end{aligned}$$

$R(f)$ 是一个线性函数, 所以

$$R(f) = \sum_{j=1}^{\infty} (\pi_j(f) - \pi_{j-1}(f)).$$

两个层级之间链路的集合定义如下

$$L_{j-1,j} := \{f - g : f \in \mathcal{H}_j, \, g \in \mathcal{H}_{j-1} \text{ and } \rho_S(f, g) \leq 2^{-j+2}\}.$$

对于一个固定的链接 $\ell \in L_{j-1,j}$

$$R(l) = n^{-1} \sum_{i=1}^{n} \sigma_i \ell(x_i),$$

且 $|\ell(x_i)| \leq 2^{-j+2}$ 所以根据Hoeffding不等式

$$\begin{aligned} \text{IP}(R(\ell) \geq t) &\leq e^{-nt^2/(\frac{2}{n}\sum_{i=1}^{n} \ell^2(x_i))} \\ &\leq e^{-nt^2/(2 \cdot 2^{-2j+4})}. \end{aligned}$$

链接集合的基数为

$$|L_{j-1,j}| \leq |\mathcal{H}_j| \cdot |\mathcal{H}_{j-1}| \leq (|\mathcal{H}_j|)^2.$$

所以

$$\text{IP}\left(\forall \ell \in L_{j-1,j}, \, R(\ell) \leq t\right) \geq 1 - (|\mathcal{H}_j|)^2 e^{-nt^2/2^{-2j+5}},$$

设置

$$t = \sqrt{\frac{2^{-2j+5}}{n}(4 \log |\mathcal{H}_j| + u)} \leq \sqrt{\frac{2^{-2j+5}}{n} 4 \log |\mathcal{H}_j|} + \sqrt{\frac{2^{-2j+5} u}{n}},$$

给我们

$$\text{IP}\left(\forall \ell \in L_{j-1,j}, \, R(\ell) \leq \frac{2^{7/2} 2^{-j} \log^{1/2} |\mathcal{H}_j|}{\sqrt{n}} + 2^{5/2} 2^{-j} \sqrt{\frac{u}{n}}\right) \geq 1 - \frac{1}{|\mathcal{H}_j|} e^{-u}.$$

如果 $\mathcal{H}_{j-1} = \mathcal{H}_j$，那么

$$\pi_{j-1}(f) = \pi_j(f) \text{ 并且 } L_{j-1,j} = \{0\}.$$

所以在所有级别和链接上，概率至少为$1 - \sum_{j=1}^{\infty} \frac{1}{|\mathcal{H}_j|} e^{-u}$

$$\text{对于所有的} j \geq 1，\text{对于所有的} \ell \in L_{j-1,j}，\ R(\ell) \leq \frac{7/2 2^{-j} \log^{1/2} |\mathcal{H}_j|}{\sqrt{n}} + 2^{5/2} 2^{-j} \sqrt{\frac{u}{n}},$$

和

$$1 - \sum_{j=1}^{\infty} \frac{1}{|\mathcal{H}_j|} e^{-u} \geq 1 - \sum_{j=1}^{\infty} \frac{1}{j^2} e^{-u} = 1 - \left(\frac{\pi^2}{6} - 1 \right) e^{-u} \geq 1 - e^{-u}.$$

对于某个层级 k

$$2^{-(k+1)} \leq d(0, f) \leq 2^{-k}$$

和

$$0 = \pi_0(f) = \pi_1(f) = \cdots = \pi_k(f).$$

所以

$$\begin{aligned} R(f) &= \sum_{j=k+1}^{\infty} R(\pi_j(f) - \pi_{j-1}(f)) \\ &\leq \sum_{j=k+1}^{\infty} \left(\frac{2^{7/2} 2^{-j}}{\sqrt{n}} \log^{1/2} |\mathcal{H}_j| + 2^{5/2} 2^{-j} \sqrt{\frac{u}{n}} \right) \\ &\leq \sum_{j=k+1}^{\infty} \left(\frac{2^{7/2} 2^{-j}}{\sqrt{n}} \log^{1/2} \mathcal{P}(\mathcal{H}, 2^{-j}, \rho_S) \right) + 2^{5/2} 2^{-k} \sqrt{\frac{u}{n}}. \end{aligned}$$

由于$2^{-k} < 2d(f,0)$，我们得到定理中的第二项

$$2^{7/2} d(0, f) \sqrt{\frac{u}{n}}.$$

对于第一项

$$\begin{aligned} \sum_{j=k+1}^{\infty} \frac{2^{7/2} 2^{-j}}{\sqrt{n}} \log^{1/2} \mathcal{P}(\mathcal{H}, 2^{-j}, \rho_S) &\leq \frac{2^{9/2}}{\sqrt{n}} \sum_{j=k+1}^{\infty} 2^{-(j+1)} \log^{1/2} \mathcal{P}(\mathcal{H}, 2^{-j}, \rho_S) \\ &\leq \frac{2^{9/2}}{\sqrt{n}} \int_0^{2^{-(k+1)}} \log^{1/2} \mathcal{P}(\mathcal{H}, \varepsilon, \rho_S) d\varepsilon \\ &\leq \frac{2^{9/2}}{\sqrt{n}} \int_0^{d(0,f)} \log^{1/2} \mathcal{P}(\mathcal{H}, \varepsilon, \rho_S) d\varepsilon \end{aligned}$$

，上述数量是达德利熵积分。 □

12.7. 覆盖数和VC维度

在这一部分中，我们将展示如何通过VC维度来限制覆盖数。
我们介绍的覆盖数通常适用于实值函数，而不是指示函数。

VC维度和VC类的概念可以扩展到各种映射中的实值函数。 最常见的扩展是VC子图类的概念。

定义。 函数 $f(x)$的子图，其中 $f : \mathcal{X} \to \mathrm{IR}$是集合

$$\mathcal{F}_f = \{(x, t) \in \mathcal{X} \times \mathrm{IR} : 0 \le t \le f(x) \ or \ f(x) \le t \le 0\}。$$

定义。 函数类 $\mathcal{H}$的子图是集合

$$\mathcal{F} = \{\mathcal{F}_f : f \in \mathcal{H}\}。$$

定义。如果 $\mathcal{F}$是一个VC类的集合，那么 $\mathcal{H}$是一个VC子图类的函数
且 $v(\mathcal{H}) = v(\mathcal{F})$。

我们现在证明，我们可以用VC维度的函数来上界覆盖数目用经验$\ell_1$norm来上界覆盖数目，对于具有有限VC维度的假设空间。

定理。给定一个VC子图类 $\mathcal{H}$其中 $-1 \le f(x) \le 1 \ \forall f \in \mathcal{H}$和 $x \in \mathcal{X}$
与 $v(\mathcal{H}) = d$和 $\rho_S(f, g) = n^{-1} \sum_{i=1}^{n} |f(x_i) - g(x_i)|$那么

$$\mathcal{P}(\mathcal{H}, \varepsilon, \rho_s) \le \left(\frac{8e}{\varepsilon} \log \frac{7}{\varepsilon}\right)^d .$$

上述定理中的界限可以改进为

$$\mathcal{P}(\mathcal{H}, \varepsilon, \rho_s) \le \left(\frac{K}{\varepsilon}\right)^d ,$$

然而，证明更加复杂，因此我们证明了较弱的陈述。
证明。

设 $m = \mathcal{P}(\mathcal{H}, \varepsilon, \rho_s)$，因此 $\{f_1, ..., f_m\}$是 ε分离的，每个函数 f_k都有其相应的子图 $\mathcal{F}_{f_k}$.

从 $\{x_1, ..., x_n\}$ 中均匀采样，其中k 属于 $\{z_1, ...z\ k\}$ 并且在 $[-1, 1]$ k elements $\{t_1, ..., t_k\}$上均匀采样。

现在我们限制两个 ε分离函数的子图选择不同的子集的概率IP $(\mathcal{F}_f$

$_k$ 和 $\mathcal{F}_{f_l}$ 挑选出不同的子集 $\{(z_1, t_1), ..., (z_k, t_k)\})$
$=$ IP（至少有一个（z_i, t_i）被 $\mathcal{F}_f$挑选出来） $_k$ 或者 $\mathcal{F}_f$ 但不是另一个）
$=$ $1 -$IP（所有（z_i, t_i）都被两者或者都没有挑选出来）.

（z_i, t_i）被两者中的任意一个挑选出来的概率 $\mathcal{F}_f$ $_k, \mathcal{F}_{f_l}$ 或者两者都没有挑选出来

12.8. 对称化和Rademacher复杂度

在之前的讲座中，我们考虑了各种复杂度度量，比如覆盖数。 但是对于我们提出的学习问题，什么是正确的复杂度概念呢？ 暂时考虑一下覆盖数。 取一个小的函数类并取其凸包。 得到的类可能非常大。

然而，期望误差和经验误差之差的上确界将在顶点处达到，即在基类处。 从某种意义上说，类的“内部”并不重要。 覆盖数考虑了整个类，并且对于凸包来说会变得非常大，尽管基本复杂度是基类的复杂度。 这表明覆盖数不是理想的复杂度度量。 在本讲座中，我们介绍另一个概念（Rademacher平均数），可以称为“正确”的概念。特别是，

凸包的Rademacher平均数将等于基类的Rademacher平均数。
这种复杂度概念还具有其他良好的性质。

我们将采用更长的路线，展示这些平均数是如何产生的，而不是直接跳到Rademacher平均数的定义。 关于这个主题的结果可以在经验过程理论中找到，因此我们将给出一些定义。

设 $\mathcal{F}$为函数类。 如果对于任意的i，$(Z_i)_{i\in\mathcal{I}}$是一个以 $\mathcal{F}$为索引的随机过程，那么 $Z\ i(f)$对于任意的i都是一个随机变量。

与之前一样， μ是Ω上的概率测度，数据 $x_1, ..., x_n \sim \mu$。 那么 μ_n是支撑在 $x_1, ..., x_n$上的经验测度：

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}。$$

定义 $Z_i(\cdot) = (\delta_{x_i} - \mu)(\cdot)$，即

$$Z_i(f) = f(x_i) - \mathrm{IE}_\mu(f)。$$

那么 $Z_1, ..., Z_n$是一个均值为0的独立同分布过程。

在之前的讲座中，我们研究了以下数量：

$$\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathrm{IE}f \right|, \tag{12.1}$$

可以写成 $n \sup_{f\in\mathcal{F}} |\sum_{i=1}^{n} Z_i(f)|$。

回想一下，(12.1)的困难在于我们不知道 μ，因此无法计算IEf。 覆盖 $\mathcal{F}$并使用并集边界的经典方法太松散了。

命题。对称化：如果 $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$接近 IE$f$对于数据 $x_1, ..., x_n$，那么 $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ 接近 $\frac{1}{n}\sum_{i=1}^{n} f(x'_i)$，对于 $x'_1, ..., x'_n$的经验平均值(对于 $x_1, ..., x_n$的一个独立副本)。 因此，如果两个经验平均值相距很远，则经验误差与预期误差相距很远。

现在修复一个函数 f。 令 $\epsilon_1, ..., \epsilon_n$为独立同分布的Rademacher随机变量（以1/2的概率取值为0或1）。那么

$$\begin{aligned}
\mathrm{IP}\left[\left|\sum_{i=1}^{n}(f(x_i) - f(x'_i))\right| \geq t\right] &= \mathrm{IP}\left[\left|\sum_{i=1}^{n}\epsilon_i(f(x_i) - f(x'_i))\right| \geq t\right] \\
&\leq \mathrm{IP}\left[\left|\sum_{i=1}^{n}\epsilon_i f(x_i)\right| \geq t/2\right] + \mathrm{IP}\left[\left|\sum_{i=1}^{n}\epsilon_i f(x'_i)\right| \geq t/2\right] \\
&= 2\mathrm{IP}\left[\left|\sum_{i=1}^{n}\epsilon_i f(x_i)\right| \geq t/2\right]
\end{aligned}$$

结合对称化，这表明控制IP（$|\sum_{i=1}^{n}\epsilon_i f(x_i)| \geq t/2$）足以控制IP $\left(\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathrm{IE}f\right| \geq t\right)$。 当然，这是一个非常简单的例子。 我们能否对类别中均匀分布的量进行相同的操作?

定义。 经验过程的上确界：

$$Z(x_1, ..., x_n) = \sup_{f\in\mathcal{F}} \left[\mathrm{IE}f - \frac{1}{n}\sum_{i=1}^{n} f(x_i) \right].$$

定义。 $Rademacher$过程的上确界：

$$R(x_1, ..., x_n, \epsilon_1, ..., \epsilon_n) = \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right].$$

命题。 $Rademacher$过程的期望值限制了经验过程的期望值：

$$\mathrm{IE}Z \leq 2\mathrm{IE}R^1.$$

证明。

$$\begin{aligned}
\mathrm{IE}Z &= \mathrm{IE}_x \sup_{f \in \mathcal{F}} \left[\mathrm{IE}f - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right] \\
&= \mathrm{IE}_x \sup_{f \in \mathcal{F}} \left[\mathrm{IE}_{x'} \left(\frac{1}{n} \sum_{i=1}^{n} f(x_i') \right) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right] \\
&\leq \mathrm{IE}_{x,x'} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(x_i') - f(x_i)) \\
&= \mathrm{IE}_{x,x',\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f(x_i') - f(x_i)) \\
&\leq \mathrm{IE}_{x,x',\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i') + \mathrm{IE}_{x,x',\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (-\epsilon_i) f(x_i) \\
&= 2\mathrm{IE}R \quad \square.
\end{aligned}$$

正如我们之前讨论的，我们希望对经验过程进行界定 Z因为这将对 $\mathcal{F}$中的任何函数都意味着“泛化”。 我们将通过Rademacher平均值$\mathrm{IE}R$来界定 Z，我们将看到它具有一些很好的性质。

定理。 如果 $\mathcal{F}$中的函数在 $[a, b]$之间均匀有界，则以概率 $1 - e^{-u}$

$$Z \leq 2\mathrm{IE}R + \sqrt{\frac{2u(b-a)}{n}}.$$

证明。 这个不等式涉及两个步骤

(1) 集中度 Z在其均值$\mathrm{IE}Z$周围

(2) 应用界定$\mathrm{IE}Z \leq 2\mathrm{IE}R$

我们将使用麦克迪尔密德不等式作为第一步。 我们定义以下两个变量 $Z := Z(x_1, ..., x_i, ..., x_n)$ 和 $Z_i := Z(x_1, ..., x'^i, ..., x_n)$。 由于对于所有的$f \in \mathcal{F}$，有 $a \leq f(x) \leq b$：

$$\begin{aligned}
\left| Z^i - Z \right| &= \left| \sup_{f \in \mathcal{F}} |\mathrm{IE}f - n^{-1} \sum_{j=1}^{n} f(x_j) + \left(n^{-1} f(x_i) - n^{-1} f(x_i') \right)| - \sup_{f \in \mathcal{F}} |\mathrm{IE}f - n^{-1} \sum_{j=1}^{n} f(x_j)| \right| \\
&\leq \sup_{f \in \mathcal{F}} \frac{1}{n} |f(x_i) - f(x_i')| \leq \frac{b-a}{n} = c_i.
\end{aligned}$$

1数量$\mathrm{IE}R$被称为$Rademacher$平均。

这限制了经验过程的鞅差。 给定差异边界，McDiarmid不等式说明

$$\text{IP}\left(Z - \text{IE}Z > t\right) \leq \exp\left(\frac{-t^2}{2\sum_{i=1}^{n}\frac{(b-a)^2}{n^2}}\right) = \exp\left(\frac{-nt^2}{2(b-a)^2}\right).$$

因此，至少有$1 - e^{-u}$的概率,

$$Z - \text{IE}Z < \sqrt{\frac{2u(b-a)}{n}}.$$

因此，随着样本数量的增加， n, Z越来越集中在IEZ周围。

应用对称化证明了定理。 至少有$1-e^{-u}$的概率。

$$Z \leq \text{IE}Z + \sqrt{\frac{2u(b-a)}{n}} \leq 2\text{IE}R + \sqrt{\frac{2u(b-a)}{n}}. \quad \square$$

McDiamid不等式不包含方差的概念，因此可以使用Talagrand不等式来获得更尖锐的不等式，用于经验过程的上确界。

现在我们需要限制Rademacher平均。 在前一讲中，Kolmogorov链中隐含了这样一个界限。 在我们重新陈述该结果并给出一些例子之前，我们先陈述一些关于Rademacher平均的好用且有用的性质。

性质。 让 $\mathcal{F}$, $\mathcal{G}$是实值函数的类。. 那么对于任意的 n,

(1) 如果$\mathcal{F} \subseteq \mathcal{G}$, 那么 $\text{IE}R(\mathcal{F}) \leq \text{IE}R(G)$

(2) $\text{IE}R(\mathcal{F}) = \text{IE}R(conv\mathcal{F})$

(3) 对于任意的$c \in \text{IR}$, $\text{IE}R(c\mathcal{F}) = |c|\text{IE}R(\mathcal{F})$

(4) 如果 ϕ: IR $\to$ IR是 L-*Lipschitz*函数且 $\phi(0) = 0$, 那么 $\text{IE}R(\phi(\mathcal{F})) \leq 2\,L\,\text{IE}R(\mathcal{F})$

(5) 对于*RKHS*球, $c\left(\sum_{i=1}^{\infty}\lambda_i\right)^{1/2} \leq \text{IE}R(\mathcal{F}_k) \leq C(\sum_{i=1}^{\infty}\lambda_i)^{1/2}$, 其中 λ_i是*RKHS*中相应线性算子的特征值。定理. *Rademacher*平均值由*Dudley*的熵积分界定

$$\text{IE}_\epsilon R \leq c\frac{1}{\sqrt{n}}\int_0^D \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, L_2(\mu_n))}d\epsilon,$$

其中 $\mathcal{N}$表示覆盖数。

例子。 设 $\mathcal{F}$为具有有限*VC*维度 V的类。那么

$$\mathcal{N}(\epsilon, \mathcal{F}, L_2(\mu_n)) \leq \left(\frac{2}{\epsilon}\right)^{kV},$$

对于某个常数 k。 上述熵积分被限制为

$$\begin{aligned}\int_0^1 \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, L_2(\mu_n))}d\epsilon &\leq \int_0^1 \sqrt{kV\log 2/\epsilon}\; d\epsilon \\ &\leq k'\sqrt{V}\int_0^1 \sqrt{\log 2/\epsilon}\; d\epsilon \leq k\sqrt{V}.\end{aligned}$$

因此， $\text{IE}_\epsilon R \leq k\sqrt{\frac{V}{n}}$ 对于某个常数 k。

讲座 13

混合模型和潜在空间模型

我们现在考虑具有额外未观察变量的模型。 这些变量被称为潜在变量或状态变量，而这些模型的一般名称为状态空间模型。一个经典的例子来自遗传学/进化，可以追溯到1894年，即螃蟹的甲壳是来自一个正态分布还是来自两个正态分布的混合。

我们将从一个潜在空间模型的常见例子开始，混合模型。

13.1. 混合模型

13.1.1. 高斯混合模型（GMM）

混合模型利用潜在变量来模拟不同组（或簇）的数据点的不同参数。 对于一个点 x_i，将其所属的簇标记为 z_i；其中 z_i是潜在的，或者未观察到的。 在这个例子中（尽管可以扩展到其他似然函数），我们假设可观测特征 $\mathbf{x}_i$服从高斯分布，因此均值和方差将根据点 x_i所关联的簇来选择。 然而，在实践中，几乎可以选择任何分布作为可观测特征。 对于d维高斯数据 x_i，我们的模型是：

$$\begin{aligned} z_i \mid \boldsymbol{\pi} &\sim \mathrm{Mult}(\boldsymbol{\pi}) \\ \mathbf{x}_i \mid z_i = k &\sim \mathcal{N}_D(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \end{aligned}$$

其中 $\boldsymbol{\pi}$是K维单纯形上的一个点，所以 $\boldsymbol{\pi} \in \mathbb{R}\ K$ 满足以下性质：$\sum^{Kk}_{=1} \pi_k = 1$，并且对于所有的$k \in \{1, 2, .., K\} : \pi_k \in [0,1]$。 变量 π被称为混合比例，而簇特定的分布被称为混合成分。 变量 $\boldsymbol{\mu}_k$和 $\boldsymbol{\Sigma}_k$分别是第 k个簇的均值和协方差参数（$k \in \{1, 2, .., K\}$）。 图1是一个GMM的示例，其中 $d = 1$且$^K = 3$。

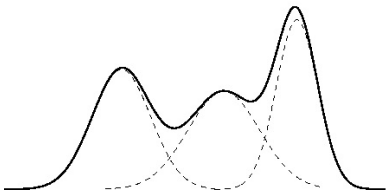


图1。 具有三个高斯混合成分的一维高斯混合模型的密度，每个成分都有自己的均值和方差项（$K=3$， $d=\mathbf{1}$）。. [来源：http://p**r**at**ee**kvjosh**i.**com]

给定参数的条件下，x的似然是：

$$\begin{aligned} p(\mathbf{x}_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\pi}) &= \sum_{k=1}^{K} p(x_i, z_i = k \mid \pi, \theta) \\ &= \sum_{k=1}^{K} p(z_i = k|\pi_k) p(x_i \mid z_i = k, \theta) \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}_d(\mathbf{x_i}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}), \end{aligned}$$

这是一个高斯混合模型的加权线性和，这样我们可以很好地解释π_k为我们模型中每个聚类$_k$的概率权重。

13.1.2. 用于聚类的混合模型

混合模型经常用于聚类；这是一个生成模型，因为我们明确地建模了$p(z)$和$p(x|z)$。 对于一般的参数θ=$\{\pi, \mu, \Sigma\}$，将点x_i分配给聚类k的后验概率由贝叶斯规则给出：

$$r_{ik} \triangleq p(z_i = k \mid \mathbf{x}_i, \theta) \propto p(\mathbf{x}_i \mid z_i = k, \theta)\ p(z_i = k \mid \pi).$$

计算每个数据点 x_i的后验概率分布，即已知为一个簇的概率
被称为软聚类.硬聚类是将最佳簇 z_i^*分配给数据点 x_i^*
使得

$$z_i^* = \arg\max_k (r_{ik}) = \arg\max_k \ \log\left(p\left(\mathbf{x}_i \mid z_i = k, \theta\right)\right) + \log\left(p\left(z_i = k \mid \pi\right)\right).$$

硬聚类在簇之间引入了线性边界，将数据点分配给单个簇，而软聚类计算每个点从每个簇生成的概率。

13.1.3. GMM的估计尝试

估计所有参数和聚类的一种可能方法是通过MLE过程，给定我们观察到的 $\mathbf{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$]. 我们希望估计的参数是 $\theta = \{\pi, \mu, \Sigma\}$. 为了简化符号，我们将写出 $z_i = k$作为向量 $\mathbf{z}_i$，其中 $z_{ij} = \mathbf{1}(j = k)$，表示向量中有一个单独的1，表示多项式 z_i 的取值。这被称为多项式向量表示。所以 $\mathbf{z}^k{}_i$被定义为布尔值

在 $\mathbf{z}_i$向量的 k位置的值

$$p\left(\mathbf{z}_i \mid \pi\right)=\prod_{k=1}^{K} \pi_k^{\mathbf{z}_i^k}$$

$$p\left(\mathbf{x}_i \mid \mathbf{z}_i, \theta\right)=\prod_{k=1}^{K} \mathcal{N}\left(\mu_k, \Sigma_k\right)^{\mathbf{z}_i^k}$$

因此，GMM的对数似然函数可以写成

$$\begin{aligned}\ell(\theta ; \mathbf{D}) &= \sum_{i=1}^{N} \log p(\mathbf{x}_i \mid \theta)=\sum_{i=1}^{N} \log \left[\sum_{\mathbf{z}_i=1}^{K} p(\mathbf{x}_i, \mathbf{z}_i \mid \theta)\right] \\ &= \sum_{i=1}^{N} \log \Big[\sum_{\mathbf{z}_i \in Z} \prod_{k=1}^{K} \pi_k^{\mathbf{z_i^k}} \mathcal{N}\left(\mu_k, \Sigma_k\right)^{\mathbf{z}_i^k}\Big]\end{aligned}$$

请注意 $\sum_{\mathbf{z}_i \in Z}$ 表示对 $\mathbf{z}_i$的所有可能的分类值求和

这并不解耦似然函数，因为对数不能被“推入”求和符号内部；然而，如果我们观察到每个 $\mathbf{z}_i$，那么这个问题会解耦吗？ 假设我们观察到每个 $\mathbf{z}_i$，那么现在 $\mathbf{D}=\{(\mathbf{x}_1\mathbf{z}_1), \ldots,(\mathbf{x}_N\mathbf{z}_N)]$。现在的对数似然函数变为：

$$\begin{aligned}\ell(\theta ; \mathbf{D}) &= \log \prod_{i=1}^{N} p\left(\mathbf{z}_i \mid \pi\right) p\left(\mathbf{x}_i \mid \mathbf{z}_i, \theta\right) \\ &= \sum_{i=1}^{N} \log p\left(\mathbf{z}_i \mid \pi\right)+\log p\left(\mathbf{x}_i \mid \mathbf{z}_i, \theta\right) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{K}\left[\mathbf{z}_i^k \log \pi_k+\mathbf{z}_i^k \log \mathcal{N}\left(\mu_k, \Sigma_k\right)\right]\end{aligned}$$

我们的参数估计现在是解耦的，因为我们可以分别估计 π_k和μ_k, Σ_k。 事实上，我们对每个参数都有一个单峰后验分布 $p\left(\theta \mid \mathbf{D}\right)$，因此我们说我们有可识别的参数。

当我们没有 θ的唯一最大似然估计或最大后验估计时，就会出现不可识别的参数的问题，这意味着我们的后验分布具有多个模式。 换句话说，可能有多个 θ的值产生相同的似然度。

对于我们的GMM中未观察到的 $\mathbf{z}_i$的情况，我们无法计算出唯一的MLE估计值，因为后验概率依赖于未观察到的 $\mathbf{z}_i$。

混合模型需要考虑的另一个问题是标签切换：如果我们将聚类A、B、C排序，然后再次运行，可能会得到聚类C、B、A，它们与聚类A、B、C具有相同的似然度。这在比较不同模型或给定模型的不同运行时需要考虑。

在没有观测到 z的情况下，没有绝对真实值，所以我们只能通过特征分布来猜测隐藏的原因。

13.2. 期望最大化（EM）

EM算法是一种在模型依赖于未观察到的或潜在变量的情况下进行参数估计的通用方法，由Dem**s**te**r**、La**i**r**d**和Rub**i**n于1977年发表。**EM**算法交替估计模型参数（从某个初始猜测开始）和估计潜在变量的值。 每次迭代包括一个**E**步（期望步骤）和一个**M**步（最大化步骤）。设 $X = \{x_1, \ldots, x_n\}$为一组观察变量， $Z = \{z_1, \ldots, z_n\}$为一组潜在变量。 回顾边际对数似然：

$$\begin{aligned} \ell\left(\theta; Z, X\right) &= \sum_{i=1}^{N} \log p(x_i \mid \theta) \\ &= \sum_{i=1}^{N} \log \left[\sum_{z_i} p(x_i, z_i \mid \theta) \right]. \end{aligned}$$

正如第2.3节所讨论的，这很难优化，因为我们在对数内部有求和，这不允许我们的潜在变量和参数在这个方程中分离，导致多个可能的模式。 相反，让我们定义完全数据对数似然，或者假设我们有完整观测（即潜在变量和观测变量都有）时的数据似然：

$$\ell_C\left(\theta; Z, X\right) = \sum_{i=1}^{N} \log p(x_i, z_i \mid \theta)$$

这很容易处理，因为我们在对数内部没有任何求和，并且它很好地解耦。 但它仍然依赖于未知的潜在状态。 为此，我们将看到EM中的E步骤来估计潜在状态的实现。 从完全对数似然中，我们可以根据当前参数的值对潜在变量进行期望：

$$\begin{aligned} Q\left(\theta^{(t)}\right) &= \mathrm{IE}_{\theta^{(t)}}\left[\ell_C\left(\theta; Z, X\right) \mid \mathbf{D}, \theta^{(t-1)}\right] \\ &= E\left[\sum_{i=1}^{n}\sum_{k=1}^{K} z_i^k \log \pi_k + z_i^k \log \mathcal{N}(\mathbf{x_i}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})\right] \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} E[z_i^k] \log \pi_k + E[z_i^k] \log \mathcal{N}(\mathbf{x_i}; \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}). \end{aligned}$$

这个方程 Q被称为期望完全对数似然或者辅助函数。 这个函数表示我们模型的期望充分统计量。为了解决 $Q\left(\theta^{(t)}\right)$，EM算法由两个主要步骤组成：

- E步骤：计算 $\mathbf{z}_i$的期望充分统计量：

$$r_{ik}^t = \mathrm{IE}_{\theta^t}\left[\mathbf{z}_i^k\right]$$

在这个高斯混合模型中，我们将计算 $p\left(\mathbf{z}_i = k \mid \theta, \mathbf{x}_i\right)$

- M步骤：更新 θ^t为 θ^{t+1}

$$\theta^{t+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta^t)$$

我们可以计算MAP而不是计算ML。然后，

$$\theta^{(t+1)} = \underset{\theta}{\operatorname{argmax}} \left(Q(\theta^{(t)}) + \log(p(\theta^{(t)}) \right)$$

因为这些变量是分离的，我们知道对于给定的期望值集合，ECLL对于每个参数都是凹的。 因此，M步骤涉及通过对每个变量求导数，将其设为零并解决，从而最大化ECLL，就像我们的标准MLE过程一样。

13.2.1. GMM的EM算法

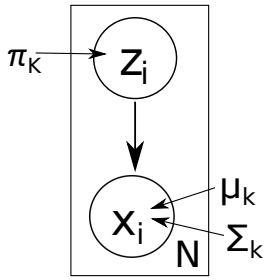


图2. GMM的图模型

在这里，我们将展示一个使用EM算法来估计我们一直在讨论的GMM的例子（图5和第2.3节）。我们希望估计的参数是 $\theta = \{\pi_k, \mu_k, \Sigma_k\}$。 我们期望的完整数据对数似然函数可以分解为：

$$\begin{aligned}
Q(\theta^t) &= \mathrm{IE}_{\theta^t}(\ell_C(\theta; Z, X)) \\
&= \mathrm{IE}(\sum_{i=1}^{n} \log p(\mathbf{x}_i, \mathbf{z}_i|\theta)) \\
&= \sum_{i=1}^{N} \mathrm{IE}_{\theta^t} \left[\log p(z_i|\pi) p(\mathbf{x}_i \mid \mathbf{z}_i, \theta)\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathrm{IE}_{\theta^t} \left[\mathbf{z}_i^k \log(\pi_k p_k(\mathbf{x}_i \mid \mu_k^t, \Sigma_k^t))\right] \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathrm{IE}_{\theta^t} \left[\mathbf{z}_i^k\right] \log(\pi_k p_k(\mathbf{x}_i \mid \mu_k^t, \Sigma_k^t)) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} \mathrm{IE}_{\theta^t} \left[\mathbf{z}_i^k\right] \log(\pi_k) + \mathrm{IE}_{\theta^t} \left[\mathbf{z}_i^k\right] \log(p_k(\mathbf{x}_i \mid \mu_k^t, \Sigma_k^t)) \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} r_{ik}^t \log(\pi_k) + r_{ik}^t \log(p_k(\mathbf{x}_i \mid \mu_k^t, \Sigma_k^t))
\end{aligned}$$

现在对于每次迭代，我们按照以下步骤计算E步和M步。

- E步骤：

$$r_{ik}^{t+1} = p\left(\mathbf{z}_i = k \mid \theta, \mathbf{x}_i\right) = \frac{\pi_k^t \, \mathcal{N}(\mathbf{x}_i \mid \mu_k^t, \Sigma_k^t)}{\sum_{j=1}^{K} \pi_j^t \, \mathcal{N}(\mathbf{x}_i \mid \mu_j^t, \Sigma_j^t)}$$

r_{ik}^{t+1} 是每个数据点分配给 k 的后验概率。

- M 步骤：在计算完 $r_i{}^t$之后，最大似然更新与朴素贝叶斯中的更新相同，即通过相应的 $Q(\theta^t)$的偏导数来更新每个参数。

$$\begin{aligned}
\pi_k^{t+1} &= \frac{\sum_{i=1}^{N} r_{ik}^t}{N} \\
\mu_k^{t+1} &= \frac{\sum_{i=1}^{N} r_{ik}^t \mathbf{x}_i}{\sum_{i=1}^{N} r_{ik}^t} \\
\Sigma_k^{t+1} &= \frac{\sum_{i=1}^{N} r_{ik}^t \left(\mathbf{x}_i - \mu_k^{t+1}\right)\left(\mathbf{x}_i - \mu_k^{t+1}\right)^T}{\sum_{i=1}^{N} r_{ik}^t}
\end{aligned}$$

13.3. K-Means

K-means算法是一种简单的聚类方法，旨在使用硬聚类将n个观测值划分为K个簇，其中每个观测值被分配到最近的均值/质心所在的簇。 正如我们将看到的，EM算法为每个簇分配了观测值属于该簇的概率（或权重）。 另一方面，K-means没有潜在的概率模型，而是以硬聚类的方式将每个观测值分配到特定的簇中。 这就是为什么我们称聚类均值为质心：强调没有潜在的（高斯）概率模型。 以下步骤实现了K-means算法。

(1) 设置 K 并选择初始质心, η_k (通常可以从 K数据点中选择这些)。

(2) 将每个数据点分配给其最近的质心:

$$z_{ik} = \begin{cases} 1, & \text{if } k = \arg\min_k \|x_i - \eta_k\|^2 \\ 0, & \text{否则} \end{cases} \tag{13.1}$$

(3) 通过计算分配给它的所有点的平均值来更新每个聚类中心

$$\eta_k = \frac{\sum_{i=1}^{N} z_{ik} x_i}{\sum_{i=1}^{N} z_{ik}}$$

(4) 重复第二步和第三步，直到聚类分配在迭代之间不发生变化。

不合适的 K 选择可能导致结果不佳，因此重要的是变化 K 并运行诊断检查。 欧几里德距离（步骤2中定义点 x_i 和质心 η_k 之间的距离的术语）也不一定是最小化的度量；其他度量，如马哈拉诺比斯距离，也可以使用。距离度量可以根据您正在使用的特定空间和特征集进行定制。图4提供了K-均值算法的可视化。本质上，K-均值假装我们观察到潜在状态 z并根据这个（假装的）观察更新聚类特定的质心。下一个

我们应该问的问题是如何在概率框架下调整K-means算法?
我们将看到EM算法是如何实现这一点的.

与广义线性模型一样，我们可以使用基于梯度的方法来找到似然函数的局部最小值（通过边缘化潜在变量z）。相反，我们将按照概率化K-means的思路进行。

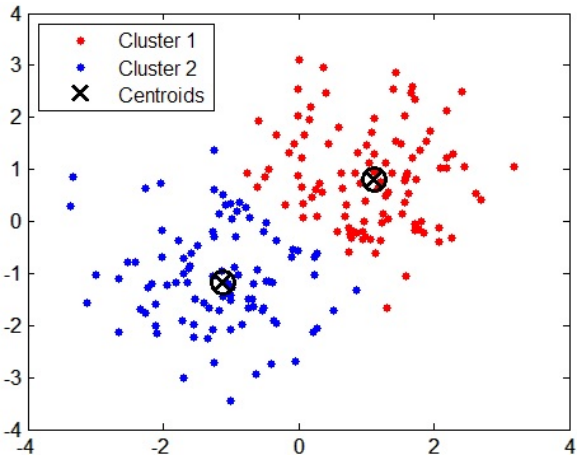


图3. K-means算法结果示例。. 来源：www.mathworks.com。

讲座 14
潜在狄利克雷分配

我们研究的生物问题是通过遗传数据推断种群结构。 这是种群遗传学/生物学中的一个重要问题，既可以了解种群的遗传历史，也可以在研究全基因组关联研究（GWAS）时控制种群结构。 事实证明，同样的模型可以用于建模文档，在文本分析领域被称为主题建模。

这些想法中使用了几个统计学概念，包括混合模型、吉布斯采样和共轭先验。 多项式数据的混合模型的一般形式被称为潜在狄利克雷分配。

无混合的种群结构推断

我们首先考虑观察到的个体没有混合的情况。 这意味着每个个体都是从一个来自k=1, ..., K个祖先种群的等位基因分布中抽取的。 然后我们将考虑混合的情况，每个个体的基因组可以来自K个祖先种群的混合。

定义问题的量是

(a) $\{X_1, ..., X_n\}$– 个体的基因型。 这些是观察到的变量，对于每个个体，我们有$x_\ell^{(\mathrm{i,a})}$ $\equiv(x(\backslash_\ell^{i,1)}, x^{(\backslash_\ell^{i,2)}}) =$ 第i个个体在第ℓ个位点的基因型，其中i=1, ..., n，$\ell = 1, ..., L$。

(b) $\{Z_1, ..., Z_n\}$– 第^{i}个个体的原始种群， $z_i =$ 第^{i}个个体的来源种群，其中 $z_i = \{1, 2, ...K\}$。(c) $p_{k\ell j}=$ 种群 k中位点 j的频率，其中 j= 1, ..., J ℓ是位点$_\ell$上可能的等位基因数， k= 1, ..., K。 注意$p_{z(\ i)\ell j}= \Pr(x^{(i,a)}_\ell = j \mid Z, P)$

观察到基因型 X。
原始种群 Z是隐藏的，必须推断出来。 频率变量
P也必须推断出来。

从条件概率的角度来看，我们希望计算基因型的似然模型的后验分布（以及 Z和 P的先验分布）

$$\Pr(Z, P \mid X) \propto \Pr(Z) \times \Pr(P) \times \mathrm{Lik}(X; Z, P), \quad \mathrm{Lik}(X; Z, P) \equiv \Pr(X \mid Z, P).$$

我们现在从最简单的情况到一般情况逐步建立问题没有混合。 假设只有一个祖先种群 $K=1$在一个位点 ℓ上只有两种可能的等位基因 $J_\ell=2$。 在位点 ℓ上的n个个体的似然函数是

$$\mathrm{Lik}(X_\ell^{1,a},\ldots,X_\ell^{(n,a)};p)\propto\prod_{i=1}^{n}p^{X_\ell^{(i,a)}}(1-p)^{1-X_\ell^{(i,a)}},$$

这是一个二项分布。 从 $J_\ell=2$ 推广到 $J_\ell>2$ 或 $X_\ell^{(i,a)}=\{0,1,2,\ldots,J_\ell\}$对应于从二项分布到多项分布的转变

$$\mathrm{Lik}(X_\ell^{(1,a)},\ldots,X_\ell^{(n,a)};\{p_{\ell 1},\ldots,p_{\ell J_\ell}\})\propto\prod_{i=1}^{n}\left[\prod_{j=1}^{J_\ell}p_{\ell j}^{I(X_\ell^{(i,a)},j)}\right]=\prod_{j=1}^{J_\ell}p_{\ell j}^{S_{\ell j}}$$

其中 $p_{\ell j}$是位点 ℓ上第 j个等位基因的概率，满足 $\sum_{j=1}^{J_\ell}p_{\ell j}=1,\ p_{\ell j}\geq 0$, $S_{\ell j}=\#\{X_\ell^{(i,a)}=j\}$是在位点 ℓ上具有等位基因 j的个体数量，而 $I(X^{(i,a)}_\ell,j)=1$如果 $X_\ell^{(i,a)}$是第 j个等位基因，否则为0（这被称为指示函数）。参数 $P_\ell=\{p_{\ell 1},\ldots,p_{\ell J_\ell}\}$是不确定的。这些参数也可以用概率分布来建模。我们首先看一下当$J_\ell=2$ 时的情况，即二项式情况，其中我们有一个参数 p。用来建模 p的自然概率分布是具有参数$\alpha,\beta>0$ 的贝塔分布

$$f(p;\alpha,\beta)\propto p^{\alpha-1}(1-p)^{\beta-1},\quad 0\leq p\leq 1$$

当 $\alpha=\beta=1$ 时，返回均匀分布。 如果我们使用贝塔分布来设置 p的先验，并使用二项式似然函数，我们得到以下关于 p的后验分布，给定我们的数据

$$\begin{aligned}\Pr(p\mid X_\ell^{(1,a)},\ldots,X_\ell^{(n,a)}) &\propto \mathrm{Lik}(X_\ell^{(1,a)},\ldots,X_\ell^{(n,a)};p)\times f(p;\alpha,\beta)\\ &= \left[\prod_{i=1}^{n}p^{X_\ell^{(i,a)}}(1-p)^{1-X_\ell^{(i,a)}}\right]p^{\alpha-1}(1-p)^{\beta-1}\\ &= \left[p^{S_\ell}(1-p)^{n-S_\ell}\right],\quad S_\ell=\#\{X_\ell^{1,a}=1\}\\ &= p^{S_\ell+\alpha-1}(1-p)^{n+\beta-S_\ell-1}\\ &= \mathrm{Beta}(S_\ell+\alpha,\,n-S_\ell+\beta),\end{aligned}$$

所以后验分布是一个beta分布。 beta分布和二项分布是共轭分布。 在多项式的情况下，自然分布是 $\{p_{\ell 1},\ldots,p_{\ell J_\ell}\}$ 由一个狄利克雷分布给出

$$f(\{p_{\ell 1},\ldots,p_{\ell J_\ell}\};\{\alpha_1,\ldots,\alpha_{J_\ell}\})\propto\prod_{j=1}^{J_\ell}p_{\ell j}^{\alpha_j-1}.$$

使用狄利克雷作为先验，我们可以证明参数的后验分布 $(\{p_{\ell 1}, ..., p_{\ell J_\ell}) $ 在给定基因型的情况下也是狄利克雷分布

$$\begin{aligned}\Pr(p \mid X_\ell^{(1,a)}, ..., X_\ell^{(n,a)}) &\propto \text{Lik}(X_\ell^{(1,a)}, ..., X_\ell^{(n,a)}; \{p_{\ell 1}, ..., p_{\ell J_\ell}\}) \times f(\{p_{\ell 1}, ..., p_{\ell J_\ell}\}; \{\alpha_1, ..., \alpha_{J_\ell}\}) \\ &= \left[\prod_{j=1}^{J_\ell} p_{\ell j}^{S_{\ell j}}\right]\left[\prod_{j=1}^{J_\ell} p_{\ell j}^{\alpha_j - 1}\right] \\ &= \prod_{j=1}^{J_\ell} p_{\ell j}^{S_{\ell j}+\alpha_j - 1} \\ &= \text{Dir}(S_{\ell 1} + \alpha_1, S_{\ell 2} + \alpha_2, ..., S_{\ell J_\ell} + \alpha_{J_\ell})。\text{所以在}\end{aligned}$$

这一点上，我们知道如果只有一个祖先种群，我们如何推断等位基因频率的后验分布，它由狄利克雷分布给出。
显然，这并不那么有趣，因为这种情况违背了推断人口结构的目的。

现在我们扩展到$K>2$ 的情况，其中我们有真实的人口结构。 我们引入一个潜在变量 $Z^{(i)}$，它为每个个体分配一个原始人口。 这种添加变量的方法有时被称为增广。 如果我们知道 $Z^i = k$，我们可以写出等位基因频率的后验分布$p_{k\ell j}$，它们是群体 k在位点$_\ell$上等位基因频率 $j= 1, ...J$ ℓ的频率。

$$\begin{aligned}\Pr(p_{k\ell 1}, ..., p_{k\ell J_\ell} \mid Z^{(i)} = k, X^{(1,a,)}, ..., X^{(n,a)}) &\propto \left[\prod_{j=1}^{J_\ell} p_{k\ell j}^{S_{k\ell j}}\right]\left[\prod_{j=1}^{J_\ell} p_j^{\alpha_j - 1}\right], \\ &= \text{Dir}(\alpha_1 + S_{k\ell 1}, ..., \alpha_{J_\ell} + S_{k\ell J_\ell}).\end{aligned}$$

这给我们提供了从后验分布$\Pr(P \mid Z, X)$中抽样的方法。我们将展示我们也可
以从$\Pr(Z \mid P, X)$中抽样。 我们可以使用贝叶斯规则写成

$$\Pr(Z^{(i)} = k \mid X, P) = \frac{\Pr(X^{(i)} \mid P, z^{(i)} = k)}{\sum_{k'} \Pr(X^{(i,a)} \mid P, z^{(i)} = k')},$$

其中

$$\Pr(X^{(i,a)} \mid P, Z^{(i)} = k) = \prod_{\ell=1}^{L} p_{k\ell x^{(i,1)}} p_{k\ell x^{(i,2)}}.$$

在这一点上，我们知道如何绘制$Pr(\ Z\ \mid P, X)$和$Pr(\ P\ \mid Z, X)$。 问题是我们想要绘制$Pr(\ Z, P\ \mid X)$。 在许多情况下，可以使用一种称为Gibb's采样的过程来实现这一点。 Gibb's采样背后的思想是，如果我想要从联合分布$Pr(Z, X)$中采样，但只能计算条件概率$Pr(Z \mid P)$和$Pr(\ P\ \mid Z)$，那么我可以使用以下迭代过程来采样联合分布：(1) 猜测一个 $Z_{(0)}$

(2) 对于 $t= 1$ 到 T
  (a) 采样 $P_{(t)} \mid Z_{(t-1)}$
  (b) 采样 $Z_{(t)} \mid P_{(t)}$
(3) 移除前 t_0 pairs的$(\ P_{(t)}, Z_{(t)})$，这被称为burn-in
(4) 保留剩余的每 a对$(P_{(t)}, Z_{(t)})$，这被称为稀疏化
(5) 现在我们从$\Pr(Z, P)$中获得了 a个独立同分布的样本

STRUCTURE中的过程按以下方式调整了上述算法

(1) 对于 i从1到 n： $Z_{(0)}^{(i)} \overset{iid}{\sim} \text{Uni}(1, ..., K)$
(2) 对于 t从1到 T
(a) 对于每个k，ℓ

$$P_{k\ell\cdot}^{(t)} \mid X, Z_{(t-1)} \sim \text{Dir}(\lambda_1 + n_{k\ell 1}, ..., \lambda_{J_\ell} + n_{k\ell J_\ell}),$$

其中 $n_{k\ell j} = \#\{(i, a) : X_\ell^{i,a} = k, Z_{(t-1)}^{(i)} = k\}$
(b) 对于每个 i

$$\Pr\left(Z_{(t)}^{(i)} = k \mid X, P^{(t)}\right) = \frac{\Pr(X^{(i)} \mid P^{(t)}, z^{(i)} = k)}{\sum_{k'} \Pr(X^{(i,a)} \mid P^{(t)}, z^{(i)} = k')},$$

其中

$$\Pr\left(X^{(i,a)} \mid P^{(t)}, Z^{(i)} = k\right) = \prod_{\ell=1}^{L} p_{k\ell x^{(i,1)}} p_{k\ell x^{(i,2)}}.$$

用混合进行人口结构推断

在存在混合的情况下，每个个体不一定来自一个祖先种群。他们的基因组来自祖先种群的混合。这就是混合。为了对此建模，我们需要引入一个新变量Q并调整我们之前的变量 Z。我们的变量 P和 X与之前相同。新的或调整后的变量为：

(a) $\{Q_1, ..., Q_n\}$– 每个个体的混合比例向量

$q_k^{(i)}$ = 第 i个个体基因组中来自种群k的比例

(b) $\{Z\}$– 等位基因拷贝 $X_\ell^{(i,a)}$来自未知种群 $Z_\ell^{(i,a)}$

$z_\ell^{(i,a)}$ = 等位基因拷贝 $X_\ell^{(i,a)}$的起源种群之前我们只需要每个个体的一个 Z。现在我们观察到

$$\Pr(X_\ell^{(i,a)} = j \mid Z, P, Q) = p_{z_\ell^{(i,a)} \ell j},$$

和

$$\Pr(z^{(i,a)} = k \mid P, Q) = q_k^{(i)},$$

我们可以给出先验分布

$$q^{(i)} \sim \text{Dir}(\alpha, ..., \alpha).$$

我们很快就会看到我们可以如何写出以下条件语句

$$P, Q \mid X, Z, \quad Z \mid X, P, Q.$$

这使我们能够写出以下吉布斯采样器。

(1) 对于每个i, a: $Z_{(0)}^{(i,a)} \overset{iid}{\sim} \text{Uni}(1, ..., K)$
(2) 对于 t从1到 T
(a) 对于每个k，ℓ

$$P_{k\ell\cdot}^{(t)} \mid X, Z_{(t-1)} \sim \text{Dir}(\lambda_1 + n_{k\ell 1}, ..., \lambda_{J_\ell} + n_{k\ell J_\ell}),$$

其中 $n_{k\ell j} = \#\{(i, a) : X_\ell^{i,a} = k, Z_{(t-1)}^{(i,a)} = k\}$

(b) 对于每个 i

$$q_{(t)}^{(i)} \mid X,\, Z_{(t-1)}^{(i,a)} \sim \mathrm{Dir}(\alpha + m_1^{(i)}, ..., \alpha + m_k^{(i)},$$

其中

$$m_k^{(i)} = \#\{(\ell, a) : z_\ell^{(i,a)} = k\}.$$

(c) 对于每个$i,\, a,\, \ell$

$$\Pr\left(Z_{(t)}^{(i,a)} = k \mid X,\, P^{(t),Q_{(t)}}\right) = \frac{q_k^{(i)} \Pr(X_\ell^{(i)} \mid P^{(t)},\, z^{(i)} = k)}{\sum_{k'} q_k'^{(i)} \Pr(X_\ell^{(i,a)} \mid P^{(t)},\, z^{(i)} = k')},$$

其中

$$\Pr\left(X_\ell^{(i,a)} \mid P^{(t)},\, Z^{(i)} = k\right) = p_{k\ell x^{(i,a)}}.$$

讲座 15
马尔可夫链蒙特卡罗

在许多情况下，后验计算是一项挑战，因为没有后验分布的闭式表达式。 马尔可夫链蒙特卡罗方法是一种从后验分布中采样的通用方法。 为了解释MCMC，我们需要介绍一些一般的马尔可夫链理论。 然而，首先我们先证明吉布斯采样，这可以在不使用任何马尔可夫链理论的情况下完成。

基本问题是我们想要从中生成样本

$$\pi(\theta) \equiv f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{f(x)} \equiv \frac{w(\theta)}{Z},$$

这里的归一化常数 $Z = \int f(x, \theta) d\theta$通常是难以处理的。 我们的MCMC算法的目标是从 $\pi(\theta)$中采样，而无需计算 Z。 计算 $w(\theta)$通常是可处理的，因为评估似然和先验通常是解析操作。

15.1. Gibbs采样器

Gibbs采样器的思想是要从联合后验分布中采样。 我们无法获得联合的闭式形式，但我们对条件概率有解析形式。 考虑一个联合后验分布$f(\theta \mid x)$，其中 x是数据，$\theta = \{\theta_1, \theta_2\}$。 为了简化表示，我们将写成$\pi(\theta_1, \theta_2) \equiv f(\theta_1, \theta_2 \mid x)$。 Gibbs采样算法的思想是以下过程将提供从联合分布 $\pi(\theta_1, \theta_2)$中的样本：

1) 设置 $\theta_{22} \sim$均匀分布[θ_2的支持]
2) 从$\pi(\ \theta_1' \mid \theta_2)$中抽取 θ_1'
3) 从$\pi(\ \theta_2' \mid \theta_1')$中抽取 θ_2'
4) 设置 θ_2为 θ_2'
5) 转到步骤2

我们现在展示为什么上述算法中的抽样 θ_1', θ_2'是从 $\pi(\theta_1, \theta_2)$中得到的。
以下链条以以下方式开始，即由以下指定的联合分布开始

上述算法并继续展示它是联合分布

$$\begin{aligned} p(\theta_1', \theta_2') &= \int \pi(\theta_1, \theta_2)\pi(\theta_1' \mid \theta_2)\pi(\theta_2' \mid \theta_1')d\theta_1 d\theta_2 \\ &= \int \pi(\theta_1, \theta_2)\frac{\pi(\theta_1', \theta_2)}{\pi(\theta_2)}\frac{\pi(\theta_1', \theta_2')}{\pi(\theta_1')}d\theta_1 d\theta_2 \\ &= \int \pi(\theta_1 \mid \theta_2)\pi(\theta_2 \mid \theta_1')\pi(\theta_2', \theta_1')d\theta_1 d\theta_2 \\ &= \pi(\theta_1', \theta_2')\left[\int \pi(\theta_1 \mid \theta_2)\pi(\theta_2 \mid \theta_1')d\theta_1 d\theta_2\right] \\ &= \pi(\theta_1', \theta_2')\left[\int \pi(\theta_1, \theta_2 \mid \theta_1')d\theta_1 d\theta_2\right] \\ &= \pi(\theta_1', \theta_2'). \end{aligned}$$

我们还可以从更一般的Metropolis-Hastings算法中推导出Gibbs采样器。 我们将把它作为一个练习留给你。

15.2. 马尔可夫链

在我们讨论马尔可夫链蒙特卡罗方法之前，我们首先必须定义一些马尔可夫链的基本属性。 我们讨论的马尔可夫链始终是离散时间的。 链在任意时间 $t = 1, ..., T$可以采用离散或连续的状态空间或值。 我们将状态空间表示为 $\mathcal{S}$，并且在几乎所有的例子中，我们将考虑一个有限的离散状态空间 $\mathcal{S} = \{s_1, ..., s_L\}$，这样我们可以使用线性代数而不是算子理论进行所有的分析。

在贝叶斯推断的背景下，自然地将状态空间 $\mathcal{S}$看作是参数值 Θ 的空间，而状态 s_i 对应于参数值θ_i。

对于离散状态的马尔可夫链，我们可以定义一个马尔可夫转移矩阵 $\mathbf{P}$，其中每个元素

$$\mathbf{P}_{s_t \to s_{t+1}} = \Pr(s_t \to s_{t+1}),$$

是在任意时间 t从一个状态转移到另一个状态的概率. 我们将概率向量 ν定义为一个由 L个数构成的向量，满足 $\sum_\ell \nu_\ell = 1$ 且 $\nu_\ell \geq 0$. 我们要求我们的链混合并具有唯一的稳定分布。 这个要求将通过两个标准来捕捉：不变性和不可约性或遍历性。

我们从不变性开始：我们希望链具有以下特性

$$\lim_{T\to\infty} \mathbf{P}^T \nu = \nu^*, \quad \forall \nu$$

极限 ν^*被称为不变测度或极限分布。 存在唯一的不变分布满足以下一般平衡条件

$$\sum_{s'} \mathbf{P}(s' \to s)\, \nu^*(s') = \nu^*(s).$$

给定过渡矩阵 $\mathbf{P}$，可以通过计算

$$\mathbf{P} = U^T \Lambda U$$来简单检验不变性，

其中，如果我们将特征值 λ_ℓ从大到小排序，我们知道最大的特征值 $\lambda_1 = 1$。 我们还知道所有特征值不能小于 -1。

所以我们现在考虑

$$\lim_{N\to\infty}\left[\mathbf{P}^N=\left(\sum_{\ell}\lambda_{\ell}^N u_{\ell}u_{\ell}^T\right)\right]$$

只要没有特征值 $\lambda_\ell=-1$，它将收敛。 此外，所有特征值 $\lambda_\ell\in(-1,1)$ 不会对极限产生影响。

链的遍历性或不可约性意味着以下内容：

存在一个 Nsuch that $\mathbf{P}^N(s'\to s)>0$ 对于所有的 s'和 $\nu^*(s)>0$。

另一种陈述上述内容的方式是整个状态空间可以从状态空间中的任意点到达。 我们可以再次使用线性代数来检查不可约性。 我们首先定义链的生成器 $L=\mathbf{P}-I$。 现在我们来看 L的特征值，从最小到最大排序。 我们知道最小的特征值$\lambda_1=0$，对应的特征向量为 $\mathbf{1}$。 如果第二个特征值 $\lambda_2>0$，那么链是不可约的， $\lambda_2-\lambda_1=\lambda_2$被称为谱间隙。对于具有唯一的遍历不变测度的马尔可夫链，以下的混合速率成立sup

$$\|\nu^*-\mathbf{P}\nu\|=O((1-\lambda)^N).$$

ν

我们希望我们的链条能够混合。

从算法角度来看，我们将设计满足所谓详细平衡的马尔可夫链：

$$\mathbf{P}(s'\to s)\nu^*(s')=\mathbf{P}(s\to s')\nu^*(s),\quad\forall s,\,s'.$$

详细平衡在算法中很容易检查，而详细平衡加上遍历性意味着链条混合。 在下一节中，我们将看到为什么详细平衡对于最常见的MCMC算法Metropolis-Hastings很容易验证。

15.3. Metropolis-Hastings算法

我们首先介绍一些符号，我们将马尔可夫转移概率或马尔可夫转移核定义为

$$Q(s';s)\equiv f(s'\mid s),$$

作为条件概率 $s'\mid s$，在有限状态空间的情况下，这些值由马尔可夫转移矩阵给出。 我们还有一个状态概率

$$p(s)\equiv\frac{w(s)}{Z},$$

在这里，我们可以使用先验和似然度来评估 $w(s)$。 请注意，无论何时我们写下 $\frac{p(s)}{p(s')}$ 我们可以使用计算 $\frac{w(s)}{w(s')}$ 作为替代。

以下是Metropolis-Hastings算法

1) $t=1$
2) $s^{(t)}\sim\text{Unif[support of }s]$
3) 从$Q(\,s';s^{(t)})$中抽取 s'
4) 计算接受概率 α

$$\alpha=\text{最小值}\left(1,\frac{p(s')Q(s;s')}{p(s)Q(s';s)}\right)$$

5) 以概率 α接受s': $u\sim\text{Unif}[0,1]$, 如果 $u\le\alpha$则 $\begin{cases}t=t+1\\ s^{(t)}=s'\end{cases}$

6) 如果$t<T$则转到步骤3，否则停止

Metropolis-Hastings算法旨在从后验分布 $p(s)$中生成样本$(s^{(1)}, ..., s^{(T)})$。 我们很快将证明该算法满足细致平衡。 在此之前，我们将陈述上述算法的一个性质。 常见的提议 $Q(s'; s)$ 是一个随机游走提议 $s' \sim \mathrm{N}(s, \sigma^2)$。 如果 σ^2非常小，那么接受率 α通常接近1，但在这种情况下，两个连续的抽样 $s^{(t)}, s^{(t+1)}$将有条件依赖关系。 如果 σ^2非常大，那么接受率 α接近0，但在这种情况下，两个连续的抽样 $s^{(t)}$, $s^{(t+1)}$将是独立的。 关于步长的局部/全局性以及良好的接受率 α的技巧，一些理论建议 $\alpha = .25$是最优的。此外，首 T_0个样本通常不是从平稳分布中抽取的，平稳分布尚未启动。因此，通常不包括首 T_0个样本，这称为燃烧期。

我们现在展示详细平衡。 首先观察到 $P(s \to s') = \alpha Q(s'; s)$。我们从这里开始

$$\begin{aligned} P(s \to s')\nu^*(s) &= Q(s'; s) \min\left(1, \frac{\nu^*(s')Q(s; s')}{\nu^*(s)Q(s'; s)}\right)\nu^*(s) \\ &= \min\left(\nu^*(s)Q(s'; s), \nu^*(s')Q(s; s')\right) \\ &= Q(s; s') \min\left(1, \frac{\nu^*(s)Q(s'; s)}{\nu^*(s')Q(s; s')}\right)\nu^*(s') \\ &= P(s' \to s)\nu^*(s'). \end{aligned}$$

讲座 16
隐藏马尔可夫模型

隐藏马尔可夫模型（HMM）的概念是马尔可夫链的扩展。
基本形式是我们有两个变量 $X_1, ..., X_T$ 被观察到
和 $Z_1, ..., Z_T$ 是隐藏状态，它们具有以下条件
依赖结构

$$\begin{aligned} x_{t+1} &= f(x_t; \theta_1) \\ z_{t+1} &= g(x_{t+1}; \theta_2), \end{aligned}$$

在这里我们将 t看作时间， $f(\cdot)$ 和 $g(\cdot)$ 是条件分布。 在这种情况下，我们将时间视为离散的。 通常在HMM中，我们认为隐藏状态是离散的，还有更一般的状态空间模型，其中隐藏变量和可观察变量都是连续的。 条件分布 $g(\cdot)$ 的参数通常被称为转移概率，观测分布 $g(x_{t+1}; \theta_2)$ 的参数通常被称为发射概率。 我们经常使用符号 $x_{1:t} \equiv x_1, ..., x_t$。

通常使用HMM提出的问题包括：

- 过滤：给定观测值 $x_1, ..., x_t$我们想要知道隐藏状态 $z_1, ..., z_t$所以我们想要推断 – $p(z_{1:t} \mid x_{1:t})$。
- 平滑：给定观测值 $x_1, ..., x_T$我们想要知道隐藏状态 $z_1, ..., z_t$其中$t < T$。在这里，我们使用过去和未来的观测来推断隐藏状态 – $p(z_{1:t})$
- 后验采样： $z_{1:T} \sim p(z_{1:T} \mid x_{1:T})$

HMM中的隐藏变量是推断的挑战所在。 我们从写下联合（完全）似然开始

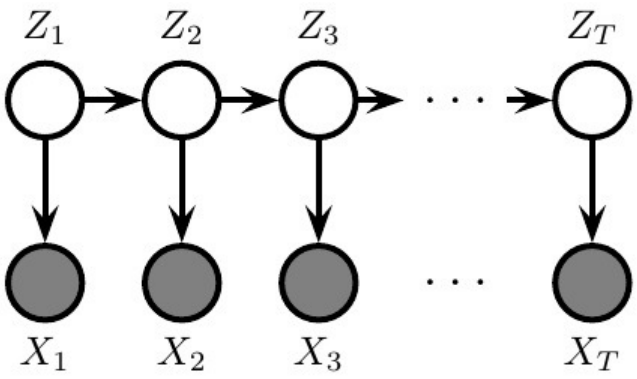
$$\text{Lik}(x_1, ..., x_T, z_1, ..., z_T; \theta_1, \theta_2) = \pi(z_1) \prod_{t=2}^{T} f(z_{t+1} \mid z_t, \theta_1) \prod_{t=1}^{T} g(x_t \mid z_t, \theta_2),$$

这里 $\pi(\cdot)$ 是初始状态的概率。 通过边缘化，可以得到观测数据的似然

$$\text{Lik}(x_1, ..., x_T; \theta_1, \theta_2) = \sum_{z_1, ..., x_T} \left(\pi(z_1) \prod_{t=2}^{T} f(z_{t+1} \mid z_t, \theta_1) \prod_{t=1}^{T} g(x_t \mid z_t, \theta_2) \right).$$

天真地说，上述求和是残酷的，因为它包含了所有可能的隐藏轨迹。
如果我们假设 N 个隐藏状态，那么我们将有 N^T 个可能的轨迹。我们

将看到马尔可夫结构在减少计算方面给我们带来了很大的好处。



16.1. EM算法

我们从完整的对数似然开始

$$\begin{aligned}\ell_c(z,\, x;\theta) &= \log[\mathrm{Lik}(z,\, x \mid \theta)] \\ &= \log\left\{p(z_1)\left[\prod_{t=1}^{T} p(z_t \mid z_{t-1})\right]\left[\prod_{t=1}^{T} p(x_t \mid z_t)\right]\right\} \\ &= \log \pi(z_1) + \sum_{t=1}^{T-1} \log a_{z_t, z_{t+1}} + \sum_{t=1}^{T} \log p(x_t \mid z_t, \theta_2).\end{aligned}$$

然后我们写出期望的完整对数似然

$$\begin{aligned}\mathrm{IE}\ell_c(z,\, x;\theta) &= \mathrm{IE}\, \log[\mathrm{Lik}(z,\, x \mid \theta)] \\ &= \mathrm{IE}\, \log\left\{p(z_1)\left[\prod_{t=1}^{T} p(z_t \mid z_{t-1})\right]\left[\prod_{t=1}^{T} p(x_t \mid z_t)\right]\right\} \\ &= \sum_{k=1}^{N} \mathrm{IE}[z_1^k] \log \pi_k + \log \pi(z_1) + \sum_{t=1}^{T-1} \sum_{j,k=1}^{K} \mathrm{IE}[z_t^j z_{t+1}^k] \log a_{jk} \\ &\quad + \sum_{t=1}^{T} \mathrm{IE}[\log p(X_t \mid Z_t, \theta_2)],\end{aligned}$$

其中 ${z_t}^k$表示在时间 t时处于第 k个状态。

对于EM算法的E步骤，我们需要计算

$$\mathrm{IE}[Z_1^k] = \mathrm{IE}[Z_1^k \mid X_{1:T}, \theta] = p(Z_1^k = 1 \mid X_{1:T}, \theta)$$

这是我们所期望的，因为 Z_1遵循多项式分布，所以其期望值就是后验概率的向量。我们还需要计算

$$\mathrm{IE}[Z_t^j,\, Z_{t+1}^k] = \mathrm{IE}[Z_t^j,\, Z_{t+1}^k \mid X_{1:T}, \theta] = \sum_{t=1}^{T-1} p(Z_t^j Z_{t+1}^k \mid X_{1:T}, \theta)$$

需要注意的是，直观上，$\mathrm{IE}[{Z_t}^j,\, {Z_t}^k{}_{+1}]$计算了我们观察到的转移对的频率。

我们现在陈述前向后向算法，这是一种计算上述期望的高效方法。 我们想要计算 $p(z_1 \mid x_{1:T})$，所以我们开始写

$$\begin{aligned} p(z_t \mid x_{1:T}) &= \frac{p(z_t, x_{1:T})}{p(y_{1:T})} \\ p(z_t, x_{1:T}) &= p(x_{1:T} \mid z_t)p(z_t) \\ &= p(x_{1:t}, z_t)p(x_{t+1:T} \mid z_t) \\ &= \alpha(z_t)\beta(z_t), \end{aligned}$$

其中 $\alpha(z_t)$ 回顾过去， $\beta(z_t)$ 展望未来。 两者都可以递归计算。

对于 α:

$$\begin{aligned} \alpha(z_t) &= p(x_{1:t}, z_t) \\ &= \sum_{z_{t-1}} p(x_{1:t}, z_t, z_{t-1}) \\ &= \sum_{z_{t-1}} p(x_{1:t-1}, z_{t-1})p(x_t, z_t \mid x_{1:t-1}, z_{t-1}) \\ &= \sum_{z_{t-1}} p(x_{1:t-1}, z_{t-1})p(z_t \mid z_{t-1})p(x_t \mid z_t) \\ &= \sum_{z_{t-1}} \alpha(z_{t-1})p(z_t \mid z_{t-1})p(x_t \mid z_t), \end{aligned}$$

请注意，给定参数模型，上述计算很容易，因为 $p(x_t \mid z_t)$是发射概率，而 $p(z_t \mid z_{t-1})$ 是状态转移概率。注意我们可以将 α初始化为 $\alpha(z_1) = p(x_1, z_1) = p(z_1)p(x_1 \mid z_1)$.

对于 β:

$$\begin{aligned} \beta(z_t) &= p(x_{t+1:T} \mid z_t) \\ &= \sum_{z_{t+1}} p(x_{t+1:T}, z_{t+1} \mid z_t) \\ &= \sum_{z_{t+1}} p(x_{t+1:T} \mid z_{t+1}, z_t)p(z_{t+1} \mid z_t) \\ &= \sum_{z_{t+1}} p(x_{t+2:T} \mid z_{t+1})p(x_{t+1} \mid y_{t+1})p(z_{t+1} \mid z_t) \\ &= \sum_{z_{t+1}} \beta(z_{t+1})p(x_{t+1} \mid z_{t+1})p(z_{t+1} \mid z_t) \end{aligned}$$

请注意，给定参数模型，上述计算很容易，因为 $p(x_{t+1} \mid z_{t+1})$是发射概率，而 $p(z_{t+1} \mid z_t)$是状态转移概率。 请注意我们可以将 β初始化为 $\beta(z_{T-1}) = p(x_T \mid z_{T-1}) = \sum_{z_T} p(x_T \mid z_T)p(z_T \mid z_{T-1})$。

这导致了一个具有两个阶段的算法

前向阶段： $\alpha(z_t) = p(x_t \mid z_t) \sum_{z_{t-1}} p(z_t \mid z_{t-1})\alpha(z_{t-1})$

反向阶段: $\beta(z_t) = \sum_{z_{t+1}} p(x_{t+1} \mid z_{t+1})p(z_{t+1} \mid z_t)\beta(z_{t+1})$.

我们还观察到

$$p(z_t \mid x_{1:T}) = \frac{p(z_1 \mid x_{1:T})}{p(x_{1:T})} \propto \alpha(z_t)\beta(z_t).$$

回想一下，在E步骤中，我们需要计算

$$\mathrm{IE}[Z_1^k] = p(z_1^k \mid x_{1:T}) \propto \alpha(z_1)\beta(z_1),$$

并且

$$\begin{aligned}
\mathrm{IE}[Z_t^j Z_{t+1}^k] &= p(z_t^j z_{t+1}^k \mid x_{1:T}) \\
&\propto p(z_t^j z_{t+1}^k, x_{t+1:T}) \\
&\propto p(x_{t+2:T} \mid z_{t+1}^k) p(x_{t+1} \mid z_{t+1}^k) p(z_{t+1}^k \mid z_t^j) p(z_t^j \mid x_{1:t})) \\
&= \beta(z_{t+1}^k)\, p(x_{t+1} \mid z_{t+1}^k)\, p(z_{t+1}^k \mid z_t^j)\, \alpha(z_t^j).
\end{aligned}$$

上述方程给出了我们对$\mathrm{IE}[Z_1^k]$和$\mathrm{IE}[Z_t{}^j Z_t{}^k{}_{+1}]$的估计，给出了当前的模型参数和 α 和 β的计算。

我们现在指定M步骤。 为了表示方便，我们将转移概率的参数表示为 $a_{jk} = p(z_t{}^j \mid z_t{}^k{}_{+1})$，初始概率表示为$\pi_i$，再次是一个多项式的发射概率的参数表示为$\eta_{jk} = p(x^j{}_t \mid z_t{}^k)$。带有参数的完整对数似然可以表示为

$$\sum_{i=1}^{N} E[Z_1^i] \log \pi_i + \sum_{t=1}^{T} \sum_{i,j=1}^{N} E[Z_t^i Z_t^j] \log a_{ij} + \sum_{t=1}^{T} \sum_{i,j=1}^{N,O} \mathrm{IE}[Z_t^i X_t^j] \log \eta_{ij},$$

我们假设了 N个隐藏状态和 O个可观察状态。 为了简化表示，我们定义以下术语 $\hat{z}_t{}^i = E[Z_t{}^i], z_t{}^{ij} = E[Z_t{}^i Z_t{}^j]$。现在我们写下足够统计量

$$z_1^i, \quad m_{ij} = \sum_{t=1}^{T} z_t^{ij}, \quad n_{ij} = \sum_{t=1}^{T} \hat{z}_t^i x_t^j.$$

给定充分统计量和参数，我们在约束条件下最小化完全对数似然

$$\sum_i \pi_i = 1, \quad \sum_{j=1}^{N} a_{ij} = 1, \quad \sum_{i=1}^{O} n_{ij} = 1.$$

使用拉格朗日乘子我们得到

$$\begin{aligned}
\hat{\pi}_i &= z_1^i \\
\hat{a}_{ij} &= \frac{m_{ij}}{\sum_{k=1}^{N} m_{ik}} \\
\hat{\eta}_{ij} &= \frac{n_{ij}}{\sum_{k=1}^{O} n_{ik}}.
\end{aligned}$$

讲座 17
谱方法和流形学习

在高维数据分析中的一个关键思想是数据的底层结构是低维的，所以即使 $X \subseteq \mathrm{IR}^p$数据的底层自由度或数据的支持是低维的，比如说， $d \ll p$。 关键问题是如何找到这个低维结构，如何利用这个低维结构，以及从数据中提取这个低维结构时做出了什么假设。 我们将用谱方法来解决这些问题。

17.1. 一般的谱方法

在本讲座中，通过谱方法，我们指的是对一个正半定对称算子进行特征分解。 我们将从数据中构建不同的算子，这些不同的算子将对数据做出不同的假设或者保留不同的数据特征。