Lecture 2: 回归, Lasso, RKHS

Tianjun Ke

Renmin University of China

Introduction to basic statistical learning, April 2023

Table of Contents

Table of Contents

引入

一套完整的数据分析流程:

1. Q: 数据? $\Longrightarrow$ A: 概率分布（Lecture 1）
2. Q: 建模? $\Longrightarrow$ A: 统计学习（Lecture 2）
3. Q: 算法? $\Longrightarrow$ A: 优化
4. Q: 决策（预测）? $\Longrightarrow$ A: 统计推断

引入

本节课的内容主要基于 BST235: Advanced Regression and Statistical Learning 的 Lecture Notes, A Primer on Reproducing Kernel Hilbert Spaces[1]以及 CS229T/STAT231: Statistical Learning Theory 的 Lecture Notes[2]。

本节课内容包括:

- 回归（线性模型）
- Lasso
- 核方法与 RKHS（非线性模型）

[1] https://arxiv.org/pdf/1408.0952.pdf

[2] https://web.stanford.edu/class/cs229t/notes.pdf

引入

相比传统课程，我们更关注统计学习角度的内容：

- 模型的应用场景 (e.g., 什么时候用 Lasso)
- 模型的学习率 (learning rate)

Table of Contents

背景知识

在课程中，我们会用到以下的背景知识

- （中心化的）亚高斯随机变量。如果一个随机变量 $X \in \mathbb{R}$ 满足 $\mathbb{E}[X] = 0$ 且

$$\mathbb{E}[\exp(sX)] \le \exp\left(\frac{\sigma^2 s^2}{2}\right), \quad \forall s \in \mathbb{R}$$

  我们称这个随机变量服从参数（variance proxy）为 σ^2 的亚高斯分布。亚高斯随机变量可以看作是高斯随机变量的推广，它具有很好的薄尾性质。（作业题 1）

- O_P 记号。如果任取 $\epsilon > 0$, 存在 $C > 0$ 以及 $N > 0$ 使得对于所有的 $n > N$ 都有

$$\mathbb{P}\left(|X_n/a_n| > C\right) < \epsilon,$$

  则称 $X_n = O_P(a_n)$。

Table of Contents

回归

给定 $i=1,\ldots,n$ 时的响应变量 Y_i 和协变量 X_i，回归模型假设

$$Y_i = f(X_i) + \varepsilon_i, \text{ for all } i=1,\ldots,n.$$

其中 ε_i 是误差/噪声。通常我们假设误差项满足 $\mathbb{E}\varepsilon_i = 0$ 且 $\varepsilon_1,\ldots,\varepsilon_n$ 是独立的。

线性回归

线性回归模型是一种特殊的回归模型，其中我们假设 $f(x) = x^\top \beta$，$\beta \in \mathbb{R}^d$，因此回归模型变为

$$Y_i = X_i^\top \beta + \varepsilon_i, \text{ for all } i = 1, \ldots, n.$$

我们还要引入一些矩阵符号来更方便地表达线性回归问题。我们定义设计矩阵 $\mathbb{X} = (X_1^\top, \ldots, X_n^\top)^\top \in \mathbb{R}^{n\times d}$，响应向量 $Y = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ 以及噪声向量 $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$。我们可以将线性模型写作

$$Y = \mathbb{X}\beta + \varepsilon.$$

我们也将设计矩阵写为列 $\mathbb{X} = (\tilde{X}_1, \ldots, \tilde{X}_d)$，其中 $\tilde{X}_j$ 是 $\mathbb{X}$ 的第 j 列。

线性回归

线性回归有两种典型情形:

- 非随机：协变量 $X_1, \ldots, X_n$ 是确定的。
- 随机：协变量 $X_1, \ldots, X_n$ 是随机的，并且我们通常假设 ε 独立于 $\mathbb{X}$。

在本次讲座及回归分析中，我们都关注了非随机情形。如果 $\mathbb{X}$ 实际上是随机的，我们可以对 $\mathbb{X}$ 取条件并还原到非随机的情形。

线性回归

我们在统计学习理论中主要关心两件事情：预测与参数估计。

- 预测。我们可以用估计 $\widehat{f}$ 与真实函数 f^* 之间的均方误差（Mean Squared Error，MSE）来衡量预测准确性：

$$\mathrm{MSE}(\widehat{f}) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(X_i) - f^*(X_i))^2.$$

在线性回归场景下，我们可以将均方误差表示为：

$$\begin{aligned}\mathrm{MSE}(\mathbb{X}\widehat{\beta}) &= \frac{1}{n}\sum_{i=1}^{n}(X_i^\top\widehat{\beta} - X_i^\top\beta^*)^2 = \frac{1}{n}\|\mathbb{X}(\widehat{\beta} - \beta^*)\|^2 \\ &= (\widehat{\beta} - \beta^*)^\top\widehat{\Sigma}(\widehat{\beta} - \beta^*),\end{aligned}$$

其中 $\widehat{\Sigma} = \mathbb{X}^\top\mathbb{X}/n = \frac{1}{n}\sum_{i=1}^{n} X_i^\top X_i$ 是样本的协方差矩阵。

- 参数估计。我们关心估计 $\widehat{\beta}$ 与 β^* 的差距，也即 $\|\widehat{\beta} - \beta^*\|$ 的收敛速度。

线性回归

下面我们以最常见的最小二乘估计为例，阐释统计学习理论中一个很重要的概念——收敛速度。一般而言，收敛速度是刻画模型好坏最直观的统计结果。让我们首先引入最小二乘估计：

$$\widehat{\beta}^{\mathrm{LS}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i^\top \beta)^2 = \arg\min_{\beta} \| Y - \mathbb{X}\beta \|^2.$$

下面给出了普通最小二乘估计的闭式解。

最小二乘估计的闭式解

$$\widehat{\beta}^{\mathrm{LS}} = (\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top Y,$$

其中 $A^\dagger$ 是 A 的 Moore-Penrose 伪逆。

我们给出简单的证明。

证明

根据定义，最小二乘损失在 $\widehat{\beta}^{\mathrm{LS}}$ 处的临界点有

$$0 = \frac{\partial}{\partial \beta} \|Y - \mathbb{X}\beta\|^2 \bigg|_{\beta = \widehat{\beta}^{\mathrm{LS}}} = 2\mathbb{X}^\top (Y - \mathbb{X}\widehat{\beta}^{\mathrm{LS}}).$$

解上述方程，可得 $\mathbb{X}^\top \mathbb{X} \widehat{\beta}^{\mathrm{LS}} = \mathbb{X}^\top Y$，因此 $\widehat{\beta}^{\mathrm{LS}} = (\mathbb{X}^\top \mathbb{X})^\dagger \mathbb{X}^\top Y$。

线性回归

下面我们来讲一讲最小二乘估计得到的估计的收敛速度。

最小二乘估计下 MSE 的收敛速度

对独立的误差项 $\varepsilon_1 \ \ldots \ \varepsilon_n$ 而言，如果它们满足 $\mathbb{E}\varepsilon_i = 0$ 且服从参数 (variance proxy) 为 σ^2 的亚高斯 (sub-Gaussian) 分布，则

$$\mathbb{E}[\mathrm{MSE}(\mathbb{X}\widehat{\beta}^{\mathrm{LS}})] \lesssim \frac{\sigma^2 r}{n},$$

并且至少以 $1-\delta$ 的概率，

$$\mathrm{MSE}(\mathbb{X}\widehat{\beta}^{\mathrm{LS}}) \lesssim \frac{\sigma^2 r}{n} + \frac{\sigma^2}{n}\log(\frac{1}{\delta}).$$

其中 $\mathrm{rank}(\mathbb{X}) = r$，$a_n \lesssim b_n$ 表示存在一个与 n 无关的常数 C，使得对于所有 n，$a_n \leq C b_n$。

为了证明这个定理，我们先引入刻画亚高斯噪声性质的极大值不等式 (maximal inequality)。

ℓ_2 范数的极大值不等式

给定随机向量 $X \in \mathbb{R}^d$。如果对所有 $u \in \mathbb{R}^d$，$\langle u, X \rangle$ 都是参数为 $\sigma^2 \|u\|^2$ 的亚高斯随机变量，则有

$$\mathbb{E}\|X\| \leq 4\sigma\sqrt{d}.$$

并且以不低于 $1-\delta$ 的概率有

$$\|X\| \leq 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}.$$

由于该不等式的证明需要引入 ε 网与覆盖数的概念，时间关系，我们在此只给出结果[3]。

[3] 见 Theorem 1.19，`https://ocw.mit.edu/courses/18-s997-high-dimensional-statistics-spring-2015/a69e2f53bb2eeb9464520f3027fc61e6_MIT18_S997S15_Chapter1.pdf`

线性回归

证明

我们使用优化中的零阶条件作为证明的起点，也即 $\widehat{\beta}^{\mathrm{LS}}$ 是 $\|Y-\mathbb{X}\widehat{\beta}\|^2$ 的最优解。我们可以建立起 $\widehat{\beta}^{\mathrm{LS}}$ 与 $\widehat{\beta}^*$ 的联系

$$\|Y-\mathbb{X}\widehat{\beta}^{\mathrm{LS}}\|^2 \le \|Y-\mathbb{X}\beta^*\|^2 = \|\mathbb{X}\beta^*+\varepsilon-\mathbb{X}\beta^*\|^2 = \|\varepsilon\|^2.$$

另一方面，

$$\|Y-\mathbb{X}\widehat{\beta}^{\mathrm{LS}}\|^2 = \|\mathbb{X}\beta^*+\varepsilon-\mathbb{X}\widehat{\beta}^{\mathrm{LS}}\|^2 = \|\mathbb{X}(\widehat{\beta}-\beta^*)\|^2-2\langle\varepsilon,\mathbb{X}(\widehat{\beta}-\beta^*)\rangle+\|\varepsilon\|^2.$$

因此，将上述两个不等式结合起来，可得

$$\|\mathbb{X}(\widehat{\beta}-\beta^*)\|^2 \le 2\langle\varepsilon,\mathbb{X}(\widehat{\beta}-\beta^*)\rangle = 2\|\mathbb{X}(\widehat{\beta}-\beta^*)\|\langle\varepsilon,\frac{\mathbb{X}(\widehat{\beta}-\beta^*)}{\|\mathbb{X}(\widehat{\beta}-\beta^*)\|}\rangle. \tag{1}$$

(续)

下一步，我们将通过“sup-out”技巧对 $\langle \varepsilon, \frac{\mathbb{X}(\widehat{\beta}-\beta^*)}{\|\mathbb{X}(\widehat{\beta}-\beta^*)\|} \rangle$ 进行放缩。定义 $\mathcal{C}(\mathbb{X})$ 为 $\mathbb{X}$ 的列向量张成的线性空间。设 $\Phi = (\phi_1, \ldots, \phi_r) \in \mathbb{R}^{n\times r}$ 的列向量为 $\mathcal{C}(\mathbb{X})$ 的标准正交基，满足 $\Phi^\top \Phi = \mathrm{I}_r$。由于 $\mathbb{X}(\widehat{\beta}-\beta^*) \in \mathcal{C}(\mathbb{X})$，因此存在 $\nu = (\nu_1, \ldots, \nu_r)^\top \in \mathbb{R}^r$，使得 $\mathbb{X}(\widehat{\beta}-\beta^*) = \sum_{j=1}^r \nu_j \phi_j = \Phi\nu$。定义 $\widetilde{\varepsilon} = \Phi^\top \varepsilon \in \mathbb{R}^r$，可得

$$\langle \varepsilon, \frac{\mathbb{X}(\widehat{\beta}-\beta^*)}{\|\mathbb{X}(\widehat{\beta}-\beta^*)\|} \rangle = \langle \varepsilon, \frac{\Phi\nu}{\|\Phi\nu\|} \rangle = \frac{\varepsilon^\top \Phi\nu}{\|\nu\|} = \langle \Phi^\top \varepsilon, \frac{\nu}{\|\nu\|} \rangle \le \sup_{\|u\|\le 1} \langle \widetilde{\varepsilon}, u \rangle = \|\widetilde{\varepsilon}\|.$$

其中我们在上述第一个不等号中使用了“sup-out”技巧。将上述不等式与上页的 (1) 结合起来，

$$\mathrm{MSE}(\mathbb{X}\widehat{\beta}^{\mathrm{LS}}) = \frac{1}{n}\|\mathbb{X}(\widehat{\beta}-\beta^*)\|^2 \le \frac{4}{n}\langle \varepsilon, \frac{\mathbb{X}(\widehat{\beta}-\beta^*)}{\|\mathbb{X}(\widehat{\beta}-\beta^*)\|} \rangle^2 \le \frac{4\|\tilde{\varepsilon}\|^2}{n}.$$

（续）

因此我们可以求出 MSE 的期望的上界

$$\mathbb{E}\left[\mathrm{MSE}\left(\mathbb{X}\widehat{\beta}^{\mathrm{LS}}\right)\right] \leq \frac{4\mathbb{E}\|\widetilde{\varepsilon}\|^2}{n} = \frac{4}{n}\sum_{i=1}^{r}\mathbb{E}\left[\widetilde{\varepsilon}_i^2\right] \leq \frac{16\sigma^2 r}{n}.$$

其中我们用到了 $\mathbb{E}\left[\widetilde{\varepsilon}_i^2\right] = \mathbb{E}\left(\phi_i^\top \varepsilon\right)^2 \leq 4\sigma^2$。这是亚高斯的性质（作业题 2）。

线性回归

(续)

为了证明 MSE 的尾部概率不等式，我们需要 ℓ_2 范数的最大不等式。因此，我们需要验证对于任何 $u \in \mathbb{R}^r$，$\langle u, \tilde{\varepsilon} \rangle$ 是具有参数为 $\sigma^2\|u\|^2$ 的亚高斯分布:

$$\mathbb{E}e^{\lambda\langle u,\tilde{\varepsilon}\rangle} = \mathbb{E}e^{\lambda\left\langle u,\Phi^\top\varepsilon\right\rangle} = \mathbb{E}e^{\lambda\langle\Phi u,\varepsilon\rangle} \le e^{\frac{\lambda^2}{2}\|\Phi u\|^2\sigma^2} = e^{\frac{\lambda^2}{2}\sigma^2\|u\|^2}.$$

现在我们可以使用极大值不等式了。把结果代入“sup-out”得到的 MSE 上界，以至少 $1-\delta$ 的概率可得

$$\mathrm{MSE}\left(\mathbb{X}\widehat{\beta}^{\mathrm{LS}}\right) \le \frac{4\|\tilde{\varepsilon}\|^2}{n} \le \frac{4}{n}[4\sigma\sqrt{r}+2\sigma\sqrt{2\log(1/\delta)}]^2 \lesssim \frac{\sigma^2 r}{n}+\frac{\sigma^2}{n}\log\left(\frac{1}{\delta}\right).$$

Table of Contents

Lasso

仍然考虑带噪声的线性回归

$$Y = \mathbb{X}\beta + \varepsilon.$$

Lasso(Least Absolute Shrinkage and Selection Operator) 是通过求解

$$\min_{\beta} \frac{1}{2n} \|Y - \mathbb{X}\beta\|_2^2 + \lambda \|\beta\|_1$$

得到的估计。

Lasso

为什么要使用 Lasso？

- 实际使用的角度：限制模型的复杂度，使得模型更加稀疏（即某些参数为零），从而提高模型的泛化能力和可解释性，解决过拟合问题。同时它还可以进行特征选择
- 统计模型的角度：**稀疏性假设**

稀疏性假设

特征维度 d 可能很大，但只有少数特征真正发挥作用。

在线性回归的背景下，这意味着我们假设 β^* 是稀疏的，即有

$$\|\beta^*\|_0 = s << d.$$

因此，我们希望通过 Lasso 估计来复原真实的稀疏 β^*。那么自然地，我们会关注 Lasso 估计的效率，也即 $\|\widehat{\beta}^{\text{Lasso}} - \beta^*\|$ 的收敛速度。

Lasso

在高维情况下，最小二乘的损失函数 $\mathcal{L}_n(\beta) = \frac{1}{2n}\|Y - \mathbb{X}\beta\|_2^2$ 有一些不妙，因为它不太凸。这是因为 $\nabla^2\mathcal{L}_n(\widehat{\beta}) = \mathbb{X}^\top\mathbb{X}/n = \widehat{\Sigma}$，当 $d >> n$ 时，$\lambda_{\min}(\widehat{\Sigma}) = 0$，意味着 $\nabla^2\mathcal{L}_n(\widehat{\beta})$ 只是半正定的，不太行。

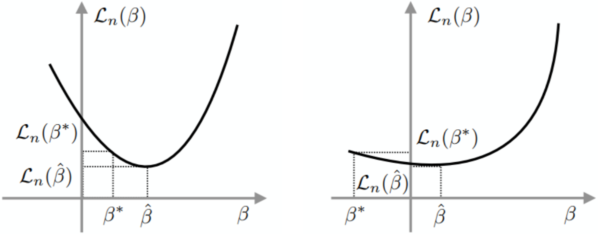


图: 左边比较凸，右边不太凸。不太凸的时候就会使 $\widehat{\beta}$ 和 β^* 离得比较远。

所以我们需要一些条件才能够对 Lasso 的收敛速度进行描述。

Lasso

RE 条件（Restricted Eigenvalue condition）

RE condition

定义 $S := \{j \mid \beta_j^* \neq 0\}$ 为 β^* 的支撑集合。如果在 $\mathbb{C}_\alpha(S) := \{\Delta \mid \|\Delta_{S^c}\|_1 \leq \alpha \|\Delta_S\|_1\}$ 中任取 $\Delta \in \mathbb{C}_\alpha(S)$ 都有

$$\frac{1}{n}\|\mathbb{X}\Delta\|_2^2 \geq \kappa\|\Delta\|_2^2,$$

则称 $\mathbb{X}$ 满足 $\mathrm{RE}(\kappa, \alpha)$ 条件。

由于样本协方差 $\widehat{\Sigma} = \mathbb{X}^\top\mathbb{X}/n$ 的最小特征值可以用下面的方法表示

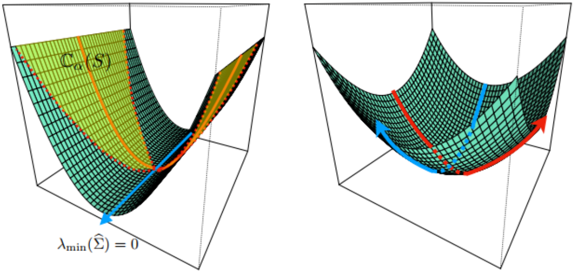
$$\lambda_{\min}(\widehat{\Sigma}) = \min_{\Delta} \frac{\Delta^\top\widehat{\Sigma}\Delta}{\|\Delta\|_2^2} = \min_{\Delta} \frac{\Delta^\top\mathbb{X}^\top\mathbb{X}\Delta}{n\|\Delta\|_2^2} = \min_{\Delta} \frac{1}{n}\frac{\|\mathbb{X}\Delta\|_2^2}{\|\Delta\|_2^2},$$

所以 RE 条件就是限制了它的最小特征值在锥中会大于等于 κ

$$\min_{\Delta\in\mathbb{C}_\alpha(S)} \frac{1}{n}\frac{\|\mathbb{X}\Delta\|_2^2}{\|\Delta\|_2^2} \geq \kappa.$$

Lasso

这个锥 $\mathbb{C}_\alpha(S)$ 的直观含义是什么呢？实际上就是在 S 能控制 S^c 的方向上，最小二乘损失是凸的。



(a) $\widehat{\Sigma} \succeq 0$ (b) $\widehat{\Sigma} \succ 0$

引入了这个条件，我们就不加证明地给出 Lasso 的收敛速度[4]。

Lasso 估计的收敛速度

如果模型满足：

- 噪声 $\varepsilon_1, \ldots, \varepsilon_n$ 是独立的，且对于所有 $i = 1, \ldots, n$，ε_i 是具有参数为 σ^2 的亚高斯随机变量
- 设计矩阵 $\mathbb{X}$ 已归一化，使得设计矩阵的第 j 列 $\mathbb{X}_j$ 的方差满足 $\frac{1}{n}\|\mathbb{X}_j\|_2^2 \le 1$，其中 $1 \le j \le d$
- $\mathbb{X}$ 满足 $\mathrm{RE}(\kappa, 3)$，且我们选择 $\lambda = \sigma\sqrt{\log(2d/\delta)/(2n)}$

则至少以 $1-\delta$ 的概率有

$$\|\widehat{\beta} - \beta^*\|_2 \le \frac{3\sigma}{2\kappa}\sqrt{\frac{2s\log(2d/\delta)}{n}}.$$

[4]可以参见这个 note 的 Theorem 15.2 的证明，只使用了 Hölder 不等式：https://www.stat.cmu.edu/~arinaldo/Teaching/36710/F18/Scribed_Lectures/Oct22.pdf

Lasso

这个结果意味着，如果我们的惩罚项系数选为 $\lambda = C\sqrt{\log d/n}$（$C$ 为某个充分大的常数），则 Lasso 估计器具有如下的收敛速度

$$\|\widehat{\beta} - \beta^*\|_2 = O_P\left(\sqrt{\frac{s \log d}{n}}\right).$$

只要 $s \log d/n = o(1)$，Lasso 估计就是相合的。还可以注意到，如果 s 固定，维数 d 可以以样本大小的指数增长速度增加。

Group Lasso

Lasso 估计器可用于从高维特征中选择变量。有时，这些特征是分组的，我们想要选择分组中的变量。例如，我们想要预测明天的 COVID 病例数。预测 Y 的协变量是分组的：

(1) 与过去病例数量相关的特征组：今天的病例数、昨天的病例数、过去一个月的病例数等
(2) 与天气相关的特征组：温度、降水等
(3) 与隔离相关的特征组：在家工作的人数、开放餐馆的数量等
(4) 与特朗普相关的特征组：特朗普的推文数量、特朗普从 COVID 中康复的天数等

我们可能期望 COVID 病例数与其中某个特征组相关。假设 $\beta \in \mathbb{R}^d$ 有 J 个组。我们将每个组表示为 $S_j \subset 1, \ldots, d$，$j = 1, \ldots, J$。因此，我们想要选择子向量 $\beta_{S_1}, \ldots, \beta_{S_J}$.

如果 β 是以组为单位稀疏的，那么向量 $\left(\|\beta_{S_1}\|_2, \|\beta_{S_2}\|_2, \ldots, \|\beta_{S_J}\|_2\right)^\top \in \mathbb{R}^J$ 是稀疏的。因此，可以考虑 Group Lasso 惩罚项

$$\left\|\left(\|\beta_{S_1}\|_2, \|\beta_{S_2}\|_2, \ldots, \|\beta_{S_J}\|_2\right)\right\|_1 = \sum_{j=1}^{J} \|\beta_{S_j}\|_2 .$$

Group Lasso 估计器可以表示如下：

$$\min_{\beta} \|Y - \sum_{j=1}^{J} \mathbb{X}_{S_j}\beta_{S_j}\|_2^2 + \lambda \sum_{j=1}^{J} \|\beta_{S_j}\|_2 .$$

Group Lasso 的一个重要应用场景是在稀疏加和模型（sparse additive model, spAM）[5]:

$$Y_i = \sum_{j=1}^{d} f_j(X_{ij}) + \varepsilon_i, \text{ for } i = 1, \ldots, n,$$

其中只有 s 个函数 f_j 是非零的。为了估计 f_j，我们用基函数将函数展开为:

$$f_j(x_j) = \sum_{k=1}^{\infty} \beta_{jk}^{*} \phi_k(x_j), \text{ for } j = 1, \ldots, d.$$

其中 $\{\phi_k\}_{k=1}^{\infty}$ 是我们选择的一种基函数, 比如多项式基 $\{x^k\}_{k=1}^{\infty}$, 三角基 $\{\sin(kx), \cos(kx)\}_{k=1}^{\infty}$, B 样条（B-splines）等等。

[5] https://arxiv.org/pdf/0711.4555.pdf

Group Lasso in spAM

因此，如果我们想要选出正确的 f_j，就等价于选择组内的基函数的系数 ${\beta^*_{jk}}_{k=1}^{\infty}$，其中 $j = 1, \ldots, d$。因此，我们就可以使用 Group Lasso 进行估计了。

spAM 的 Group Lasso 估计

$$\min_{\beta_{jk}} \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i - \sum_{j=1}^{d} \sum_{k=1}^{m} \beta_{jk} \phi_k(X_{ij}) \Big)^2 + \lambda \sum_{j=1}^{d} \Big(\sum_{k=1}^{m} \beta_{jk}^2 \Big)^{\frac{1}{2}},$$

其中 m 是我们选择用于近似真实函数的基函数的个数。

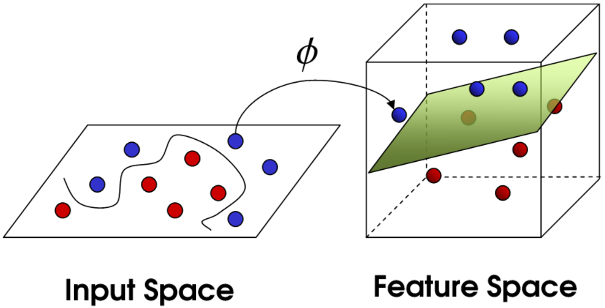
Table of Contents

核方法与 RKHS

在前面的回归中，我们基本上只关注了线性模型：$f(x) = x^\top \beta = \langle x, \beta \rangle$。然而它难以对非线性关系进行建模。但是我们可以巧妙地将 $\langle x, \beta \rangle$ 替换为 $\langle \phi(x), \beta \rangle$，其中 $\phi : \mathcal{X} \to \mathbb{R}^d$ 是任意的特征映射，例如：

- 对于 $x \in \mathbb{R}$，$\phi(x) = (1, x, x^2)$
- 对于一个字符串 x，$\phi(x) = (\text{出现的 a 的次数}, \ldots)$

因此，我们可以通过控制 $\phi(x)$ 来获得我们所需要的非线性特征。



然而 $\phi(x)$ 可能有很高的维度（甚至无穷维！），如果我们先把 x 映射到 $\phi(x)$ 再来求解 $f(x) = \langle \phi(x), \beta \rangle$ 会带来非常高的计算开销。但是如果我们可以用一些简单的运算“绕开”ϕ 的话，这个问题就会比较简单。让我们考虑回归的最小二乘损失函数。

$$L(\widehat{\beta}) = \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \langle \widehat{\beta}, \phi(X_i) \rangle)^2.$$

对 $\widehat{\beta}$ 求导，由一阶条件可得

$$\nabla L(\widehat{\beta}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle \widehat{\beta}, \phi(X_i) \rangle) \phi(X_i) = 0.$$

如果把 $\phi(X_i)$ 视为一组基的话，我们可以把 $\widehat{\beta}$ 分解为

$$\widehat{\beta} = \sum_{j=1}^{N} w_j \phi(X_j) + v,$$

其中 v 垂直于 $\mathrm{span}\{\phi(X_j), j = 1, \ldots, N\}$。代入上面的式子就有

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \langle \sum_{j=1}^{N} w_j \phi(X_j), \phi(X_i)\rangle)\phi(X_i) = 0.$$

我们发现，只要计算内积 $\langle \phi(X_i), \phi(X_j)\rangle$ 就可以求解上述问题。那我们只要找到一个对应的函数 k，使得计算 k 相对比较简单，就“绕开”了 ϕ。即找到下面这样的 k

$$k(X_i, X_j) = \langle \phi(X_j), \phi(X_i)\rangle.$$

这就是核技巧 (kernel trick)。

核方法与 RKHS

我们进一步引入核函数的定义:

核函数

函数 $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 是一个核函数当且仅当对于所有有限的点集 $x_1, \ldots, x_n \in \mathcal{X}$，由 $K_{ij} = k(x_i, x_j)$ 定义的核矩阵 $K \in \mathbb{R}^{n\times n}$ 是半正定的。

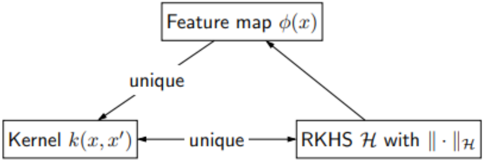
在有惩罚项的情况下，一定有 $v = 0$（由 representer theorem 得到），从而对于新的数据 X，我们有 $\widehat{f}(X) = \sum_{j=1}^{N} w_j \langle \phi(X_j), \phi(X) \rangle$。也就是说，预测函数 $\widehat{f}(X)$ 同样也可以用核函数表示。我们给出一些常见的核函数:

- 线性核： $k(x, x') = \langle x, x' \rangle$
- 多项式核： $k(x, x') = (\langle x, x' \rangle + c)^p$，其中 c 为某个常数，p 为多项式的指数。
- 高斯/rbf 核： $k(x, x') = \exp\left(\frac{-\|x-x'\|_2^2}{2\sigma^2}\right)$，最常用的核。

核方法与 RKHS

那么很自然的，我们想知道 $\phi, \widehat{f}$ 以及 k 之间的关系。尤其是对于 $\widehat{f}$，能否直接地进行刻画呢？这就需要引入 RKHS(reproducing kernel hilbert space)。

- 映射函数 ϕ：从一个数据点 $x \in \mathcal{X}$ 映射到一个内积空间 $\mathcal{H}$ 中的无穷维向量。
- 核函数 k：将一对数据点 $x, x' \in \mathcal{X}$ 映射到 $\mathbb{R}$。它刻画了某种内积关系（也即刻画了一对数据点之间的相似性）。
- RKHS $\mathcal{H}$：定义了内积 $\|\cdot\|_{\mathcal{H}}$ 的函数 $f: \mathcal{X} \to \mathbb{R}$ 的集合（函数空间）。RKHS 描述了预测函数 $\widehat{f}$ 的性质。



核方法与 RKHS

我们给出严格的定义:

RKHS

先定义 Hilbert Space: Hilbert Space 是带有内积 $\langle\cdot,\cdot\rangle : \mathcal{H}\times\mathcal{H}\to\mathbb{R}$ 的完备向量空间，其中内积满足:

- 对称性: $\langle f,g\rangle = \langle g,f\rangle$
- 线性: $\langle \alpha_1 f_1+\alpha_2 f_2, g\rangle = \alpha_1\langle f_1,g\rangle+\alpha_2\langle f_2,g\rangle$
- 正定性: $\langle f,f\rangle \geq 0$，且只在 $f=0$ 的时候取等

然后定义 RKHS: 对 $f:\mathcal{X}\to\mathbb{R}$ 定义的 Hilbert Space，RKHS 满足对所有的 $x\in\mathcal{X}$，评估泛函 (evaluation functional) $L_x := f\mapsto f(x)$ 有界。

例子: 对于 $\mathcal{X}=\mathbb{R}^d$ 以及 $\mathcal{H}=\left\{f_c : c\in\mathbb{R}^d\right\}$，其中 $f_c(x)=\langle c,x\rangle$ 是线性函数, 则 evaluation functional 为 $L_x(f_c)=\langle c,x\rangle$。
如何理解这个定义: Hilbert Space **定义了内积**，而 RKHS 使得任何在 $\mathcal{H}$ 中的函数 f 在数据点 $x\in\mathcal{X}$ 上有良好的定义，也就是说我们可以算 $f(x)$ 了。

核方法与 RKHS

我知道了 RKHS 使得 $f(x)$ 有良好的定义，可是为什么叫RKHS 呢？

再生核

对于包含 $f: \mathcal{X} \to \mathbb{R}$ 的 RKHS $\mathcal{H}$，它的再生核 $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ 满足对于所有 $f \in \mathcal{H}$ 以及 $y \in \mathcal{X}$，有 $\langle f, k(\cdot, y) \rangle = f(y)$。这里，$k(\cdot, y)$ 是指函数 $x \mapsto k(x, y)$，且它是 $\mathcal{H}$ 中的一个元素。给定 $x \in \mathcal{X}$，可以进一步证明这个 $k(x, \cdot) \in \mathcal{H}$ 是唯一的 (Riez representation theorem)

如何理解 $k(x, \cdot)$？我们可以把它理解为 f 的“坐标”。
在欧式空间中，坐标是一个这样的东西：对于某个属于 $\mathbb{R}^n$ 的元素 $(x_1, \cdots, x_n)$，我们可以用 x_i 表示它第 i 维的坐标。也就是说坐标函数 $\pi_i: \mathbb{R}^n \to \mathbb{R}$ 把 $(x_1, \cdots, x_n)$ 送去了 x_i，而且它是连续的。那么对于 RKHS 来说，根据定义我们知道 $L_x(f) = f(x) = \langle f, k(\cdot, x) \rangle$，其中 L_x 是我们所说的 evaluation functional。因此，我们也有这样的坐标函数 $L_x: \mathcal{H} \to \mathbb{R}$ 把 f 送去了 $f(x)$，而且它也是连续的[6]。

[6]线性泛函有界和连续等价，而根据定义，L_x 在 RKHS 上有界

核方法与 RKHS

坐标真的很炫酷！正如 $\mathbb{R}^n$ 中我们可以用坐标表示所有元素，在 RKHS 中我们也可以用坐标表示所有元素[7]（更严格地说，借助核函数 k 指定的坐标来构建 RKHS）。
任取 $n \in \mathbb{N}$，我们可以用 $f := \sum_{i=1}^{n} \alpha_i k(x_i, \cdot)$ 定义 RKHS 的元素（坐标的有限线性组合，$\alpha_i \in \mathbb{R}$），并定义内积：$\langle f, g \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha'_j k(x_i, x_j)$。
完备化后即可得到一个 k 对应的 RKHS[8]。

[7] 关于坐标系统的讨论，见`https://arxiv.org/pdf/1408.0952.pdf`的 1.3

[8] Moore-Aronszajn theorem，见 `https://web.stanford.edu/class/cs229t/notes.pdf`的 Theorem 22 或者`https://en.wikipedia.org/wiki/Reproducing_kernel_Hilbert_space`

现在我们能够描述它们之间的关系了!

- ϕ 确定 k：给定 $\phi : \mathcal{X} \to \mathcal{H}$，$k(x, x') := \langle \phi(x), \phi(x') \rangle$ 是核函数
- k 确定 ϕ：给定 k，存在一个 Hilbert Space $\mathcal{H}$ 和映射函数 $\phi : \mathcal{X} \to \mathcal{H}$ 使得 $k(x, x') = \langle \phi(x), \phi(x') \rangle$
- RKHS 确定 k：每个 RKHS $\mathcal{H}$ 都有唯一一个再生核 $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$
- k 确定 RKHS：对所有的核函数 k，都存在唯一一个再生核为 k 的 RKHS $\mathcal{H}$
- RKHS 确定预测函数 $\widehat{f}$：representer theorem!

Representer theorem

令 $\mathcal{H}$ 为核函数 k 对应的 RKHS，$\|f\|_{\mathcal{H}}$ 表示空间 $\mathcal{H}$ 中的函数 f 的范数。$\forall$ 单调递增函数 $\Omega:[0,\infty]\to\mathbb{R}$ 和 $\forall$ 非负损失函数 $\ell:\mathbb{R}^n\to[0,\infty]$ 优化问题

$$\min_{f\in\mathcal{H}} L(f)=\Omega\left(\|f\|_{\mathcal{H}}\right)+\ell\left(f(x_1),\ldots,f(x_n)\right)$$

的解总可以写成

$$f^*(x)=\sum_{i=1}^{n}\alpha_i k\left(x,x_i\right)$$

一方面，当我们给定了一个核 k 之后，预测函数 $\widehat{f}$ 一定会在 k 对应的 RKHS 里，从而 RKHS 描述了 $\widehat{f}$ 的所有性质。另一方面，representer theorem 给出了现实中求解 kernel 相关优化问题的方法，见下一个例子。

例子:
核岭回归 (kernel ridge regression):

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} \frac{1}{2} \left(f(x_i) - y_i\right)^2 + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2$$

用 representer theorem，我们等价于解决以下的问题:

$$\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^{n} \frac{1}{2} \left(\sum_{j=1}^{n} \alpha_j k(x_i, x_j) - y_i \right)^2 + \frac{\lambda}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j).$$

定义 $K \in \mathbb{R}^{n \times n}$ 为核矩阵，$Y \in \mathbb{R}^n$ 为向量形式的响应变量，则有

$$\min_{\alpha \in \mathbb{R}^n} \frac{1}{2} \|K\alpha - Y\|_2^2 + \frac{\lambda}{2} \alpha^\top K \alpha.$$

核方法与 RKHS

使用优化的一阶条件（对 α 求导置 0）:

$$K(K\alpha - Y) + \lambda K\alpha = 0.$$

得到解为

$$\alpha = (K + \lambda I)^{-1} Y.$$

对于一个新的输入 x，怎么获得 $\widehat{f}(x)$ 呢？(作业题 3)

总结

我们最后通过一些问题回顾本节课的内容

- 统计学习如何描述一个模型的性质？
- 统计角度下，为什么我们热爱 Lasso？
- 什么时候可以用 Group Lasso？
- 为什么引入核方法？
- 我们已经有 kernel 了，RKHS 有什么用？
- 用 kernel 有什么缺陷呢？
- kernel 如何在深度学习中登场？(Neural tangent kernel, Deep kernel learning, etc.)

作业

1. 对于一个服从参数为 σ^2 的亚高斯分布的随机变量，证明任取 $t>0$，都有 $\mathbb{P}(X>t) \le \exp\left(-\frac{t^2}{2\sigma^2}\right)$[9]。
2. 证明 P20 的 $\mathbb{E}\left[\widetilde{\varepsilon}_i^2\right] = \mathbb{E}\left(\phi_i^\top \varepsilon\right)^2 \le 4\sigma^2$ [10]。
3. 写出 P46 的 $\widehat{f}(x)$ 的具体形式。
4. 利用 representer theorem 和 RKHS 的性质，解释为什么 P46 中 $\|f\|_{\mathcal{H}} = \alpha^\top K\alpha$。进一步的，定义 n-norm 为 $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n f(x_i)$，给出它的矩阵形式。

[9]Hint：Chernoff bound $\mathbb{P}(X>t) \le \mathbb{P}\left(e^{sX} > e^{st}\right)$ + Markov's inequality

[10]Hint：利用 $\mathbb{E}\left[|X|^k\right] = \int_0^\infty \mathbb{P}\left(|X|^k > t\right) \mathrm{d}t$