核方法（3）-核方法的稳定性分析

Y.Q. Wang

2016 年 2 月 28 日

1 摘要

本章将重点分析核方法的稳定性：如何分析模型的抗噪性，以及抵抗训练数据的噪音使得模型不会出现过拟合现象。“集中度” 是分析具体函数的抗噪性的评价方法。通过对某一随机变量做随机扰动，观察具体函数的误差上界。McDiarmid 不等式是集中度评价的常用方法。本章会详细介绍 McDiarmid 不等式，并提供对应的证明方法以便理解。推论 Hoeffding 不等式是 Mcdiarmid 不等式的简化分析分析方法。通过采用核方法的质心估计这一应用问题，希望能够较为形象地帮助理解集中度分析的作用。进一步的，知道一类函数对应用问题的 capacity（即抗噪性，函数对测试集的鲁棒性）也是一个关键问题。在传统的机器学习中，利用 VC 维分析是计算 capacity 的一种常见方法。本章介绍 Rademacher 复杂度，并用 Rademacher 复杂度来评价一类函数的 capacity。利用 Rademacher 复杂度，可以直接从给定的训练集上计算函数的 capacity。接着以分类问题为例，考虑 kernel-based linear function 作为分类函数，具体分析这类函数的 Rademacher 复杂度。根据计算结果，提高函数的 capacity 可以通过两类基本途径：正则化或增加训练样本数量。训练样本通常不能掌控，因此正则化是一种更为有效的途径。在核方法中，正则化一般指限制 $\alpha' K\alpha$ 的规模，对其引入范数进行约束。对具体的问题，核方法的稳定性分析会视其场景变化而稍有不同（未来就具体问题逐个分析）。最后本章介绍了 Rademacher 复杂度的 7 个基本性质，以便扩展到具体应用问题中去。另外，掌握核方法的稳定性分析并不是应用核方法的必要环节，如无特殊需求，只需记住其中的若干结论即可:-)。

2 集中度不等式（Concentration inequality）

在机器学习中，确定任一通过有限的训练集所获得的固定函数是否是稳定的，一个关键的方法就是对数据做变换（与由变换前的数据同源）使用同一函数，判断函数输出的前后差异。Concentration（集中度）是这类评价方法的一个重要属性：对某一随机变量做变换，考察函数的前后差异。在集中度测试中，我们需要对选定的随机变量做一个非常小的扰动（通常可以是一个服从

指数族分布的扰动），并认为这种扰动所带来的数据变换能够保证新数据依旧与原数据是同源的。集中度测试结果所满足的不等式情况称之为 concentration inequation。

下面就集中度评价中最为出名的 McDiarmid inequality 做展开说明：

[McDiarmid's inequality] 令 $X_1, \ldots, X_n$ 为从集合 $\mathcal{X}$ 中取值的 n 个独立的随机变量。假设函数 $f : \mathcal{X}^n \to R$ 满足

$$\sup_{x_1,\ldots,x_n,\hat{x}_i \in \mathcal{X}} |f(x_1, \ldots, x_n) - f(x_1, \ldots, \hat{x}_i, x_{i+1}, \ldots, x_n)| \leq c_i, 1 \leq i \leq n. \tag{1}$$

则对所有的 $\epsilon > 0$ 都有

$$Pr[f - E[f] \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m} c_i^2}\right). \tag{2}$$

证明：证明该不等式需要掌握以下几点内容。

[Markov's inequality] 对任一非负随机变量 X，$Pr[X \geq t] \leq E[X]/t$。（可以由期望的定义 $E[X] = \sum_x xPr[X = x]$ 推导证明）

[Law of iterated expectation] 对随机变量 X, Y, Z，$E[E[X|Y, Z]|Z] = E[X|Z]$。（可以由期望的定义以及条件概率公式直接证明）

[Hoeffding's lemma] 令 X 为取值区间为 $[a, b]$ 的随机变量，其期望 $E[X] = 0$。则对 $t > 0$，满足 $E[e^{tX}] \leq \exp\left(t^2(b-a)^2/8\right)$。（证明见 wikipedia：Hoeffding's lemma）

正式开始：

令 $\bar{X}^i$ 是随机变量 $X_1, \ldots, X_i$ 的序列表示，定义随机变量 $Z_i = Z_i(X_1, \ldots, X_i) = E[f(X)|\bar{X}^i]$。则 $Z_0 = E[f], Z_n = f(X)$。

令 $U_i = \sup_u Z_i(X_1, \ldots, X_{i-1}, u), L_i = \inf_l Z_i(X_1, \ldots, X_{i-1}, l)$，可知 $L_i \leq Z_i(X_1, \ldots, X_i) \leq U_i$。则有

$$\begin{aligned}
|U_i - L_i| =& |E[f|X_1, \ldots, X_{i-1}, X_i = u] - E[f|X_1, \ldots, X_{i-1}, X_i = l]| \\
=& \Big| \int_{\mathcal{X}^{n-i}} f(X_1, \ldots, X_{i-1}, u, x_{i+1}, \ldots, x_n) dP_{i+1}(x_{i+1}) \ldots dP_n(x_n) \\
& - \int_{\mathcal{X}^{n-i}} f(X_1, \ldots, X_{i-1}, l, x_{i+1}, \ldots, x_n) dP_{i+1}(x_{i+1}) \ldots dP_n(x_n) \Big| \\
\leq & \int_{\mathcal{X}^{n-i}} dP_{i+1}(x_{i+1}) \ldots dP_n(x_n) \, |f(X_1, \ldots, X_{i-1}, u, x_{i+1}, \ldots, x_n) \\
& - f(X_1, \ldots, X_{i-1}, l, x_{i+1}, \ldots, x_n)| \\
\leq & |c_i|
\end{aligned}$$

因此，$E\left[\exp(t(Z_i - Z_{i-1}))|\hat{X}_{i-1}\right] \leq \exp(t^2 c_i^2/8)$（Hoeffding's lemma）。

则

$$\begin{aligned}
Pr[f - E[f] \geq \epsilon] &= Pr[e^{t(f-E[f])} \geq e^{t\epsilon}] \leq e^{-t\epsilon} E[e^{t(f-E[f])}] \ \textit{(Markov's inequality)} \\
&= e^{-t\epsilon} E[e^{t\sum_{i=1}^{m}(Z_i - Z_{i-1})}] \\
&= e^{-t\epsilon} E\left[E[e^{t\sum_{i=1}^{m}(Z_i - Z_{i-1})}|\bar{X}^{m-1}]\right] \ \textit{(Iterative expectation)} \\
&= e^{-t\epsilon} E\left[e^{t\sum_{i=1}^{m-1}(Z_i - Z_{i-1})} E[e^{t(Z_m - Z_{m-1})}|\bar{X}^{m-1}]\right] \\
&\leq e^{-t\epsilon} e^{\frac{t^2 c_m^2}{8}} E\left[e^{t\sum_{i=1}^{m-1}(Z_i - Z_{i-1})}\right]
\end{aligned}$$

重复上述步骤后可得，$Pr[f - E[f] \geq \epsilon] \leq \exp\left(-t\epsilon + \frac{t^2}{8}\sum_{i=1}^{m} c_i^2\right)$。

优化 t 使得 $-t\epsilon + \frac{t^2}{8}\sum_{i=1}^{m} c_i^2$ 取得最小值，则最小值情况为 $-2\epsilon^2 / \sum_{i=1}^{m} c_i^2$，代入上式可得，

$$Pr[f - E[f] \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m} c_i^2}\right).$$

证毕。

另外，当函数形式简化为 $f = \frac{1}{n}\sum_{i=1}^{n} X_i$ 时，可以得到 McDiarmid's inequality 的推论—Hoeffding's inequality。

[Hoeffding's inequality] 对 n 个独立的随机变量 $X_i \in [a_i, b_i]$，在 McDiarmid's inequality 的条件中，令 $f = \frac{1}{n}\sum_{i=1}^{n} X_i, c_i = \frac{b_i - a_i}{n}$，则

$$Pr[f - E[f] \geq \epsilon] \leq \exp\left(\frac{-2\epsilon^2 m^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right). \tag{3}$$

应用：质心估计

在 Hoeffding's inequality 中，函数形式 f 是 n 个随机变量的平均值。倘若将输入空间的 X 通过 ϕ 映射到高维空间，$\phi(X)$ 的均值是否仍然保持原函数的性质？以下将做质心估计来评价这一问题。

令训练集 $S = \{x_1, \ldots, x_n\}$，其在高维空间中真实的期望质心为 $E_{\mathbf{x}}[\phi(\mathbf{x})] = \int_X \phi(\mathbf{x}) dP(\mathbf{x})$。根据训练集 S 所得实际质心为 $\phi_S = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)$。则实际质心与真实的期望质心的差异为

$$g(S) = \|\phi_S - E_{\mathbf{x}}[\phi(\mathbf{x})]\| \tag{4}$$

应用 McDiarmid inequality，计算

$$\begin{aligned}\left|g(S)-g(\hat{S})\right| &= \left|\|\phi_S - E_{\mathbf{x}}[\phi(\mathbf{x})]\| - \left\|\phi_{\hat{S}} - E_{\mathbf{x}}[\phi(\mathbf{x})]\right\|\right| \\ &\leq \left\|\phi_S - \phi_{\hat{S}}\right\| = \frac{1}{n}\left\|\phi(\mathbf{x}) - \phi(\hat{\mathbf{x}})\right\| \leq \frac{2R}{n},\end{aligned}$$

其中 $R = \sup_{\mathbf{x}} \|\phi(\mathbf{x})\|$（$\hat{\mathbf{x}}$ 是 $\mathbf{x}$ 的微小扰动，因此可认为两者上界相同）。则

$$Pr\left[g(S) - E_S[g(S)] \geq \epsilon\right] \leq \exp(-\frac{2n\epsilon^2}{4R^2}). \tag{5}$$

上式，期望的估计误差 $E_S[g(S)]$ 对问题无用，进一步的需要将其消除。计算 $E[g(S)]$

$$\begin{aligned}E_S[g(S)] &= E_S[\|\phi_S - E_{\mathbf{x}}\phi(\mathbf{x})\|] = E_S[\|\phi_S - E_{\bar{S}}\phi(\bar{S})\|] \\ &= E_S\left[\|E_{\bar{S}}[\phi_S - \phi_{\bar{S}}]\|\right] \leq E_{S\bar{S}}\left[\|\phi_S - \phi_{\bar{S}}\|\right] \ \textit{triangle inequation} \\ &= E_{\sigma S\bar{S}}\left[\frac{1}{n}\left\|\sum_{i=1}^{n}\sigma_i(\phi(\mathbf{x}_i) - \phi(\bar{\mathbf{x}}_i))\right\|\right] \\ &\leq 2E_{\sigma S}\left[\frac{1}{n}\left\|\sum_{i=1}^{n}\sigma_i\phi(\mathbf{x}_i)\right\|\right] \ \textit{triangle inequation} \\ &= \frac{2}{n}E_{\sigma S}\left[\left\langle\sum_{i=1}^{n}\sigma_i\phi(\mathbf{x}_i), \sum_{i=1}^{n}\sigma_i\phi(\mathbf{x}_i)\right\rangle^{1/2}\right] \\ &\leq \frac{2}{n}E_{\sigma S}\left[\sum_{i,j=1}^{n}\sigma_i\sigma_j k(\mathbf{x}_i, \mathbf{x}_j)\right]^{1/2} \quad \textit{Jensen's inequation} \\ &= \frac{2}{n}E_S\left[\sum_{i,j=1}^{n}k(\mathbf{x}_i, \mathbf{x}_j)\right]^{1/2} \\ &\leq \frac{2R}{\sqrt{n}}.\end{aligned}$$

其中，计算第三行所引入的变量 σ 是Rademacher 变量（以概率 0.5 的方式取 +1 或 -1 的 Rademacher 分布变量）；第四行中使用到了 “样本 S 和 $\bar{S}$ 的产生方式一致” 的假设；倒数第二行用到了核矩阵的对称性质；最后一行的不等式用到了定义 $R = \sup_{\mathbf{x}} \|\phi(\mathbf{x})\|$。

假设 δ 等于公式 5 的右侧，并将 $E_S[g(S)]$ 一并代入公式 (5) 中可知，在质心估计问题中，能够以概率 $1-\delta$ 的方式获得误差不大于

$$g(S) \leq \frac{R}{\sqrt{n}}\left(2 + \sqrt{2\ln\frac{1}{\delta}}\right) \tag{6}$$

的估计效果。

从公式 (6) 中可知，在质心估计问题中，估计的好坏与输入变量的维度无关（特征空间维度）。

3 Capacity

简单的理解：当训练数据多到足够反应数据的分布时，可以较为简单的获得对数据分布的估计。而当数据体量不足时，尽力拟合数据往往会导致过拟合问题。因此定义 capacity：一类函数对不同数据的拟合能力。在机器学习中，通常采用 VC 维（Vapnik-Chervonenkis dimension）来估计。这里引入 Rademacher 复杂度来估计 capacity。

[Rademacher complexity] 样本 $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ 是由定义在集合 X 上的分布 $\mathcal{D}$ 所产生，对于定义在 X 上的实值函数 $\mathcal{F}$ 而言，其经验的 Rademacher 复杂度为

$$\hat{R}_n(\mathcal{F}) = E_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(\mathbf{x}_i) \right| \middle| \mathbf{x}_1, \ldots, \mathbf{x}_n \right], \tag{7}$$

其中 σ_i 是相互独立的 Rademacher 变量。函数 $\mathcal{F}$ 的 Rademacher 复杂度为

$$R_n(\mathcal{F}) = E_S[\hat{R}_n(\mathcal{F})] = E_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i f(\mathbf{x}_i) \right| \middle| \mathbf{x}_1, \ldots, \mathbf{x}_n \right], \tag{8}$$

通过 Rademacher 复杂度引出对函数估计误差精确界的形式化。

定理 1： 固定 $\delta \in (0, 1)$，令 $\mathcal{F}$ 为一类从输入空间 Z 到 $[0, 1]$ 的函数。令 $(\mathbf{z}_i)_{i=1}^n$ 是从分布 $\mathcal{D}$ 独立采样所得。则随机采样 n 次，函数 $f \in \mathcal{F}$ 以不小于 $1 - \delta$ 的概率满足

$$\begin{aligned} E_{\mathcal{D}}[f(\mathbf{z})] &\leq \hat{E}[f(\mathbf{z})] + R_n(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2n}} \\ E_{\mathcal{D}}[f(\mathbf{z})] &\leq \hat{E}[f(\mathbf{z})] + \hat{R}_n(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}. \end{aligned} \tag{9}$$

其中 $\hat{E}$ 指在某一采样上的经验期望值（由 Rademacher 复杂度以及 Mcdiarmid's inequality 证明，无力完全理解，从略……）。

虽然没有给出定理 1 的证明，但定理 1 明确了函数的 capacity 可以直接由 Rademacher 复杂度从给定的训练集上估计获得。接下来将定理 1 应用到基于核的线性函数类（Kernel-based linear classes）中。

4 正则化（Regularization）

以分类问题为例分析基于核的线性函数类的稳定性。令分类函数为

$$\mathcal{L}(\mathbf{x}, y) = \mathcal{H}(-yg(\mathbf{x})), \tag{10}$$

其中 $\mathcal{H}$ 为越阶函数（Heaviside function）

$$\mathcal{H}(z) = \begin{cases} 1, & \text{if } z > 0; \\ 0, & \text{otherwise} \end{cases}, \tag{11}$$

g 是输入为 x 的分类函数，y 为取值为 $\{\pm 1\}$ 的分类标签。因此，可以定义函数类 $\hat{\mathcal{F}}$

$$\hat{\mathcal{F}} = \{(\mathbf{x}, y) \to -yg(\mathbf{x}) : g \in \mathcal{F}\}. \tag{12}$$

定义这类函数的损失为

$$E_{\mathcal{D}}[\mathcal{H}(-yg(\mathbf{x}))] = E_{\mathcal{D}}[\mathcal{H}(f(\mathbf{x}, y))] = Pr_{\mathcal{D}}(y \neq h(\mathbf{x})). \tag{13}$$

由此确定目标，考察 $\mathcal{H} \circ \hat{\mathcal{F}} = \{\mathcal{H} \circ f : f \in \hat{\mathcal{F}}\}$ 的 Rademacher 复杂度。

首先，引入辅助损失函数 $\mathcal{A}$，满足 $\mathcal{H}(f(\mathbf{x}, y)) \leq \mathcal{A}(f(\mathbf{x}, y))$。辅助损失函数 $\mathcal{A}$ 是一个满足 Lipschitz 条件的函数，其定义如下：

定义 1： Lipschitz 函数 $\mathcal{A}$:$R \to [0, 1]$ 满足

$$|\mathcal{A}(a) - \mathcal{A}(a')| \leq L|a - a'| \forall a, a' \in R. \tag{14}$$

其中 L 称为 Lipschitz 常量。

记 $(.)_+$ 为函数

$$(x)_+ = \begin{cases} \text{x}, & x \geq 0; \\ 0, & \text{otherwise} \end{cases}. \tag{15}$$

可以定义函数 $\mathcal{A}$ 为 hinge loss$\mathcal{A}(f(\mathbf{x}, y)) = (1 + f(\mathbf{x}, y))_+ = (1 - yg(\mathbf{x}))_+$。

考虑 kernel-based linear function，定义具体的分类函数为

$$\left\{\mathbf{x} \to \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, x) : \alpha' K \alpha \leq B^2\right\} \subseteq \mathbf{x} \to \langle \mathbf{w}, \phi(\mathbf{x}) : \|w\| \leq B\rangle = \mathcal{F}_B. \tag{16}$$

其中 B 为边界范数（bounded norm）。**注意：**公式16的左侧集合是定义在具体的训练集上的。而 $\mathcal{F}_B$ 是不针对具体训练集的函数类。

定理 2：若 $k : X \times X \to R$ 是一个核，$S = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ 是从 X 中采样的点集，则函数类 $\mathcal{F}_B$ 的经验 Rademacher 复杂度满足

$$\hat{\mathcal{R}}_n(\mathcal{F}_B) \leq \frac{2B}{n}\sqrt{\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{x}_i)} = \frac{2B}{n}\sqrt{tr(K)}. \tag{17}$$

证明：

$$\begin{aligned}
\hat{\mathcal{R}}_n(\mathcal{F}_B) &= E_\sigma\left[\sup_{f\in\mathcal{F}_B}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(\mathbf{x}_i)\right|\right] \\
&= E_\sigma\left[\sup_{w\leq B}\left|\left\langle w, \frac{2}{n}\sum_{i=1}^{n}\sigma_i\phi(\mathbf{x}_i)\right\rangle\right|\right] \\
&\leq \frac{2B}{n}E_\sigma\left[\left\|\sum_{i=1}^{n}\sigma_i\phi(\mathbf{x}_i)\right\|\right] \\
&= \frac{2B}{n}E_\sigma\left[\left(\sum_{i=1}^{n}\sigma_i\sigma_j k(\mathbf{x}_i, \mathbf{x}_j)\right)^{1/2}\right] \\
&\leq \frac{2B}{n}E_\sigma\left[\left(\sum_{i=1}^{n}\sigma_i\sigma_j k(\mathbf{x}_i, \mathbf{x}_j)\right)\right]^{1/2} \\
&= \frac{2B}{n}(\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{x}_i))^{1/2}
\end{aligned} \tag{18}$$

证毕。

由定理 2 可知，要提高基于核的线性函数类 capacity，可以通过减小范数 B 或是增加样本数 n 达到。减小范数 B 这一方法称作正则化（regularization）。

推广：（不做详述）介绍一下经验 Rademacher 复杂度的 7 条性质。

定义 conv($\mathcal{F}$) 是由属于向量空间 $\mathcal{F}$ 的元素的凸组合（convex combinations）所构成的集合。令 $\mathcal{F}, \mathcal{F}_1, \ldots, \mathcal{F}_l$ 和 $\mathcal{G}$ 都是实函数类。则

- 若 $\mathcal{F} \subseteq \mathcal{G}$，则 $\hat{R}_n(\mathcal{F}) \leq \hat{R}_n(\mathcal{G})$；
- $\hat{R}_n(\mathcal{F}) = \hat{R}_n(\text{conv}(\mathcal{F}))$；
- 对所有 $c \in R$，$\hat{R}_n(c\mathcal{F}) = |c|\hat{R}_n(\mathcal{F})$；
- 若 $\mathcal{A} : R \to R$ 是常量为 L 的 Lipschitz 函数，满足 $\mathcal{A} = 0$，则 $\hat{R}_n(\mathcal{A} \circ \mathcal{F}) \leq 2L\hat{R}_n(\mathcal{F})$；
- 对任意函数 h，$\hat{R}_n(\mathcal{F} + h) \leq \hat{R}_n(\mathcal{F}) + 2\sqrt{\hat{E}[h^2]/n}$；

- 对任意 $1 \leq q < \infty$，令 $\mathcal{L}_{\mathcal{F},h,q} = \{|f-h|^q | f \in \mathcal{F}\}$。若 $\|f-h\|_\infty \leq 1$ 对所有 $f \in \mathcal{F}$ 成立，则 $\hat{R}_n(\mathcal{L}_{\mathcal{F},h,q}) \leq 2q\left(\hat{R}_n(\mathcal{F}) + 2\sqrt{\hat{E}[h^2]/n}\right)$；
- $\hat{R}_n(\sum_{i=1}^l \mathcal{F}_i) \leq \sum_{i=1}^l \hat{R}_n(\mathcal{F}_i)$。