中国海洋大学

水产学院

统计学习要素

Author:
韩方成

Student ID:
18060013010

统计学习要素笔记

留一份不足，可得无限美好

2021 年 7 月 20 日

目录

1 监督学习导论

1.1 两种简单的预测方法：最小二乘法和最近邻法

线性模型和最小二乘法

我们将截距 β_0 加入到系数 β 中，我们得到：

$$Y = \beta X \tag{1.1}$$

那我们模型的误差平方和为：

$$\mathrm{RSS}(\beta) = (y - X\beta)^T (y - X\beta) \tag{1.2}$$

我们的目的是让误差平方和最小，所以我们对其求导，在导数为 0 的点即取到最小值：

$$X^T (y - X\beta) = 0 \tag{1.3}$$

如果 $X^T X$ 是可逆的，那么我们就可以得到：

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{1.4}$$

最近邻法

最近邻法从训练集中挑选离输入空间 x 中最近的值 y 来作为预测值，即

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i \tag{1.5}$$

k 近邻算法的有效参数个数为 N/k，一般会大于最小二乘法的参数个数 p，并且随着 k 的增大而减小。但为什么有效参数个数是 N/k 呢？因为假设用离样本点最近的 k 个点来决定分类，如果相邻点是不重叠的，那么最多可以分为 N/k 个区域，KNN 算法对这 N/k 个区域每个都独立地估计一个响应值，共需要估计 N/k 个独立的值。

1.2 统计决策理论

假设 $X \in \mathbb{R}^p$ 为随机输入向量，$Y \in \mathbb{R}$ 为输出随机变量，他们的联合概率密度函数为 $\Pr(X, Y)$。我们使用平方均值误差作为损失函数，则

$L(Y, f(X)) = (Y - f(X))^2$。这就给了我们一个衡量 f 好坏的一个标准：

$$\begin{aligned} \mathrm{EPE}(f) &= E(Y - f(X))^2 \\ &= \int\int [(y - f(x))^2] \Pr(x, y) dx dy \end{aligned} \tag{1.6}$$

称为期望 (平方) 预测误差。通过用条件概率改写公式，我们可以将公式1.6写为

$$\begin{aligned} \mathrm{EPE}(f) &= E(Y - f(X))^2 \\ &= \int\int [(y - f(x))^2] \Pr(x, y) dx dy \\ &= \int\int [(y - f(x))^2 | x] \Pr(y|x) \Pr(x) dy dx \\ &= \int E_{Y|X}([Y - f(X)]^2 | X) \Pr(x) dx \\ &= E_X E_{Y|X}([Y - f(X)]^2 | X) \end{aligned} \tag{1.7}$$

对于某一点，我们的估计值为：

$$f(x) = \arg\min_c E_{Y|X}([Y - c]^2 | X = x) \tag{1.8}$$

因为在上式中 x 已经知道，因此我们可以将式子简化为

$$\begin{aligned} E_{Y|X}([Y - c]^2 | X = x) &= \int [(y - c)^2 | x] \Pr(y|x) dy \\ &= \int [(y - c)^2] \Pr(y) dy \end{aligned} \tag{1.9}$$

对上式进行求导

$$\int (y - c) \Pr(y) dy = 0 \tag{1.10}$$

得

$$c = \int y \Pr(y) dy = E(Y) \tag{1.11}$$

所以结果为

$$f(x) = E(Y | X = x) \tag{1.12}$$

得到的条件期望函数，也被成为回归函数。因此当用平方均值误差函数作为损失函数时，Y 的最有估计为 $X = x$ 是的条件期望。

最近邻试图使用训练数据直接实现该方法。我们直接用最接近 x 的 k 个点的 y 的均值来近似 y：

$$\hat{f}(x) = \text{Ave}(y_i|x_i \in N_k(x)) \tag{1.13}$$

在联合概率分布函数 $\Pr(X,Y)$ 的限制比较小时，可以证明当 $N,k \to \infty$ 并且 $k/N \to 0$ 时，$\hat{f}(x) \to E(Y|X=x)$。但是我们的数据通常没有那么多，并且随着数据数量和维数的增多，算法收敛的速度会越来越慢。

但是为什么线性回归能在此模式下工作呢？因为对于线性模型我们可以通过导数的知识来求解：

$$\beta = [E(XX^T)]^{-1}E(XY) \tag{1.14}$$

还有一些比较复杂的模型比如相加模型：

$$f(X) = \sum_{j=1}^{p} f_j(X_j) \tag{1.15}$$

它保持了线性模型的相加性，同时 f_j 也是任意的。我们可以利用 K 近邻的方法同时对每个 f_j 进行单参数的估计。

我们之前使用的损失函数都是 L_2 损失函数，如果我们使用 $L_1 : E|Y - f(X)|$ 作为损失函数呢？这种情况下的估计值为：

$$\hat{f}(x) = \text{median}(Y|X=x) \tag{1.16}$$

那为什么是中位数呢，由于不能直接求导，我们考虑两种情况，当 $y > c$ 时导数为 1，当 $y < c$ 时导数为 -1，所以当 c 为中位数是大于小于的数一样多，最后相加为 0。

如果我们的响应变量是一个分类变量 G 怎么办。方法与上文相似。我们的损失函数可以是一个 $K \times K$ 的矩阵 L，$L(k,l)$ 是 k 被分为 l 的代价。但是大多数情况下我们还是采用 $0-1$ 损失函数。期望预测误差为：

$$\text{EPE} = E[L(G, \hat{G}(x))] \tag{1.17}$$

同样的，EPE 可以写为

$$\text{EPE} = E_X \sum_{k=1}^{K} [\mathcal{G}_k, \hat{G}(X)] \Pr(\mathcal{G}_k|X) \tag{1.18}$$

同样的在某一个点的损失为

$$\hat{G}(x) = \arg\min_{g\in\mathcal{G}} \sum_{k=1}^{K} L(\mathcal{G}_k, g) \Pr(\mathcal{G}_k|X = x) \tag{1.19}$$

我们采用 $0-1$ 函数作为损失函数：

$$\hat{G}(x) = \arg\min_{g\in\mathcal{G}} [1 - \Pr(g|X = x)] \tag{1.20}$$

或者简化为

$$\hat{G}(X) = \mathcal{G}_k \text{如果} \Pr(\mathcal{G}_k|X = x) = \max_{g\in\mathcal{G}} \Pr(g|X = x) \tag{1.21}$$

这个就称为贝叶斯分类器。

1.3 高维问题的局部方法

我们考虑输入在 p 维单位超立方体均匀分布的最近邻过程 (图1.1)。假设我们在某个目标点构造超立方体的邻域来捕获观测值的一小部分 r。因为这个邻域对应单位体积的比例 r，则边长的期望值为 $e_p(r) = r^{1/p}$。在十维空间下 $e_{10}(0.01) = 0.63, e_{10}(0.1) = 0.80$，而全部范围为 1.0。所以选取 1% 或 10% 的数据形成局部均值，我们必须在每个输入变量上覆盖到 63% 或者 80%。这样的邻域不再是局部的，显著降低 r 也没有任何作用，因为我们选取的观测值越小，我们拟合的方差也会越大。

高维下的稀疏采样的另外一个后果就是所有的样本点离样本的某一边很近。考虑在 p 维以原点为中心的单位球中均匀分布的 N 个数据点。假设我们考虑原点处的最近邻估计。距离原点最近的数据点距离的中位数由下式给出：

$$d(p, N) = (1 - (\frac{1}{2})^{1/N})^{1/p} \tag{1.22}$$

证明，我们将数据点与原点的距离看作是随机变量 X，因为数据均匀分布，则 X 的分布函数为

$$F(X < x) = x^p, x \in [0, 1]$$

设距离原点最近的数据点距离的中位数为 r，D_0 为距离原点最近的点，则

$$\mathrm{P}(D_0 \geq r) = \frac{1}{2}$$

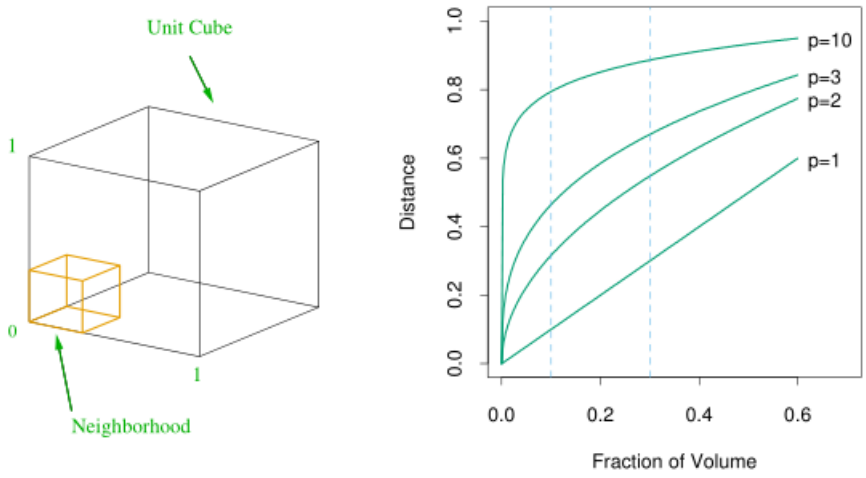

图 1.1: 超立方体下的维数灾难

距离原点最近的点距离大于 r，则所有的点的距离都大于 r，即 $(1-r^p)^N = \frac{1}{2}$ 解的：

$$d(p,N) = (1-(\frac{1}{2})^{1/N})^{1/p}$$

对于 $N = 500, p = 10, d(p,N) \approx 0.52$，超过了半径。

另一个问题是采样的密度与 $N^{1/p}$ 成正比。因此，如果 $N_1 = 100$ 代表一维情况下的取样密度，则对于 10 维的输入，样本量为 $N_{10} = 100^{10}$ 才能得到相同的样本密度。

让我们构建另一个均匀分布的例子。假设我们有 1000 个从 $[-1,1]^p$ 上的均匀分布产生的样本 x_i。假设 X 与 Y 之间的关系是

$$Y = f(X) = e^{-8||X||^2}$$

没有测量误差。我们使用 1 近邻在测试数据 $x_0 = 0$ 处预测 y_0。定义训练数据为 $\mathcal{T}$。我们可以计算在 x_0 点的期望预测误差，即 1000 个测试数据的平

均误差。当这个问题确定后，这就是在 $f(0)$ 点的平方均值误差 (MSE):

$$\begin{aligned}\mathrm{MSE}(x_0) &= E_{\mathcal{T}}[f(x_0) - \hat{y}_0]^2 \\ &= E_{\mathcal{T}}[\hat{y}_0 - E_{\mathcal{T}}]^2 + [E_{\mathcal{T}}(\hat{y}_0) - f(x_0)]^2 \\ &= \mathrm{Var}_{\mathcal{T}}(\hat{y}_0) + \mathrm{Bias}^2(\hat{y}_0)\end{aligned} \tag{1.23}$$

我们将 MSE 分为了两部分：方差和偏差平方。除非最近邻位置为 0，否则 $\hat{y}(0)$ 将会小于 $f(0)$，因此平均估计值就会减小 (图1.2)。当 $N = 1000$ 并在低维条件下，最近邻非常接近 0，此时偏差和方差都很小。当维数逐渐增加时，相邻点逐渐远离目标点，偏差和方差都会增大。当 $p = 10$ 时，超过 99% 的点距离原点的距离超过 0.5。因此随着 p 的增加，估计值趋近于 0(距离原点越远，最近邻的值就越趋近于 0)，此时 MSE 的值稳定在 1 附近 ($\mathrm{MSE} = E_{\mathcal{T}}[f(x_0) - \hat{y}_0]^2 = (1-0)^2 = 1$)，偏差也是，由于所有的数据都分布在边缘，逐渐靠近，并且函数在边缘处的导数接近于 0，数值差异很小，故所以方差开始减小。

但偏差项在 1− 最近邻中并不总是占主导地位，如图1.3，函数值只与少数维数有关，方差占主导地位。

另一方面，假设我们知道 Y 与 X 之间的关系是线性的：

$$Y = X^T\beta + \epsilon \tag{1.24}$$

在这里 $\epsilon \sim N(0, \sigma^2)$，我们采用最小二乘法来训练模型。对于一个任意的测试数据 x_0，我们有 $\hat{y}_0 = x_0^T\hat{\beta}$，我们可以将其写为 $\hat{y}_0 = x_0^T\beta + \sum_{i=1}^N \ell_i(x_0)\epsilon_i$，$\ell_i(x_0)$ 为 $X(X^TX)^{-1}x_0$ 的第 i 个元素。因为在这个模型下，最小二乘法是

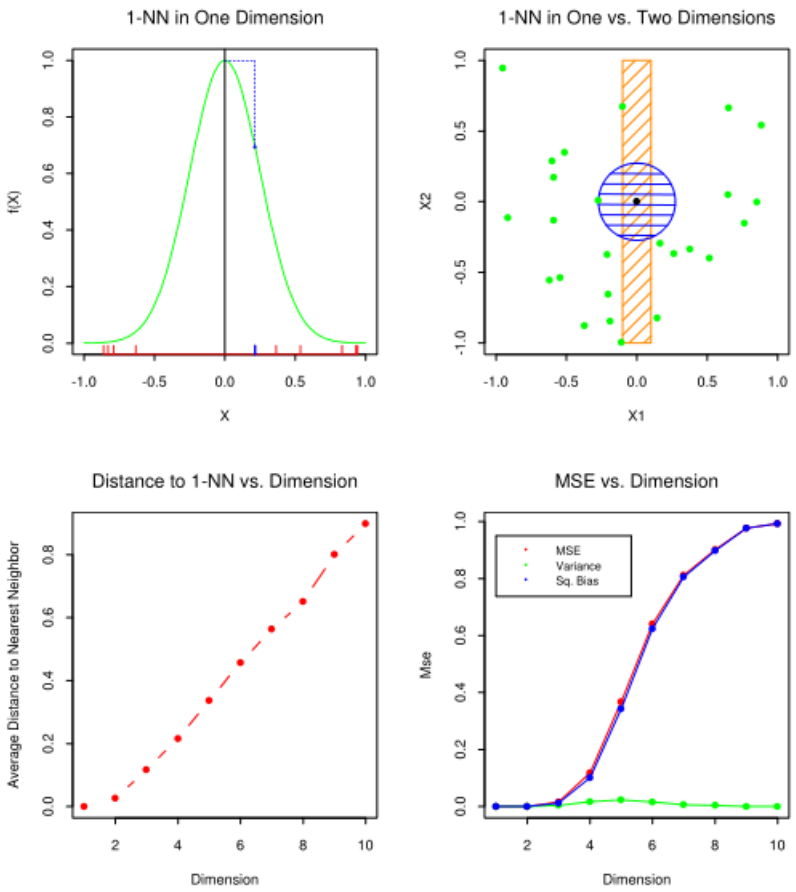

图 1.2: 一个模拟的例子，证明维数灾难以及其在 MSE，偏差和方差的影响。输入的特征在 $[-1,1]^p, p = 1, \cdots, 10$ 上的均匀分布。左上角显示了在 $\mathbb{R}$ 上的目标函数 (无噪声):$f(X) = e^{-8||X||^2}$，而且展示了 1− 最近邻估计 $f(0)$ 时的误差。训练点用蓝色的记号表示。右上角展示了为什么 1− 最近邻的半径随着维数 p 的增加而增加。

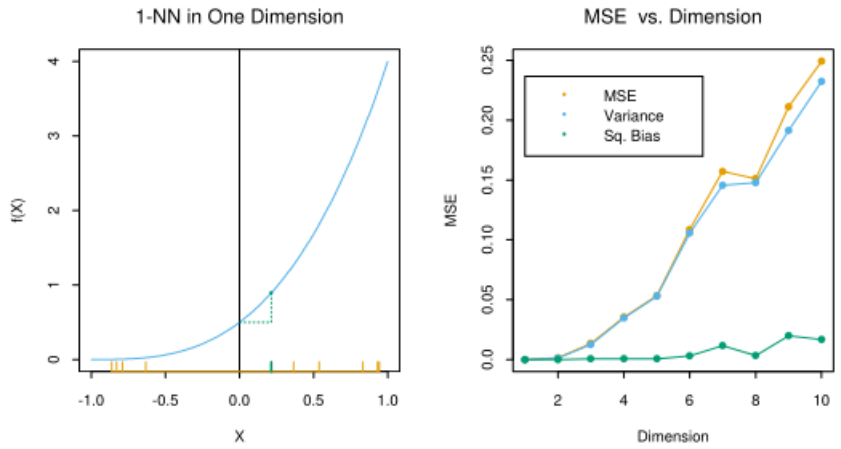

图 1.3: 一个与图1.2类似的例子。在这里函数是确定的并且函数值只与一个变量有关:$F(X) = \frac{1}{2}(X_1 + 1)^3$，此时方差占优势。

无偏估计，我们有

$$\begin{aligned}
\mathrm{EPE}(x_0) &= E_{y_0|x_0} E_{\mathcal{T}} (y_0 - \hat{y}_0)^2 \\
&= E_{y_0|x_0} E_{\mathcal{T}} (x_0^T \beta + \epsilon - \hat{y}_0)^2 \\
&= E_{y_0|x_0} E_{\mathcal{T}} (x_0^T \beta - \hat{y}_0)^2 + \epsilon^2 - 2\epsilon(x_0^T \beta - \hat{y}_0) \\
&= E_{y_0|x_0} \epsilon^2 + E_{\mathcal{T}} (x_0^T \beta - \hat{y}_0)^2 \\
&= \mathrm{Var}(y_0|x_0) + E_{\mathcal{T}} (x_0^T \beta - E_{\mathcal{T}} \hat{y}_0 + E_{\mathcal{T}} \hat{y}_0 - \hat{y}_0)^2 \\
&= \mathrm{Var}(y_0|x_0) + E_{\mathcal{T}} [\hat{y}_0 - E_{\mathcal{T}} \hat{y}_0]^2 + [E_{\mathcal{T}} \hat{y}_0 - x_0^T \beta]^2 \\
&= \mathrm{Var}(y_0|x_0) + \mathrm{Var}_{\mathcal{T}} (\hat{y}_0) + \mathrm{Bias}^2 (\hat{y}_0) \\
&= \sigma^2 + E_{\mathcal{T}} x_0^T (X^T X)^{-1} x_0 \sigma^2 + 0^2
\end{aligned} \tag{1.25}$$

对于 $\hat{y}_0 = x_0^T\beta + \sum_{i=1}^N \ell_i(x_0)\epsilon_i$ 的补充
我们有

$$\hat{\beta} = (X^TX)^{-1}X^Ty = (X^TX)^{-1}X^T(X\beta+\epsilon) = \beta + (X^TX)^{-1}X^T\epsilon$$

因此

$$\hat{y}_0 = x_0^T\hat{\beta} = x_0^T\beta + x_0^T(X^TX)^{-1}X^T\epsilon$$

因此这就得到了：

$$\hat{y}_0 = x_0^T\beta + \sum_{i=1}^N \ell_i(x_0)\epsilon_i$$

$\ell_i(x_0)$ 是 N 维列向量 $X(X^TX)^{-1}x_0$ 的第 i 个元素。

对于倒数第二步到倒数第一步的推理：

$$\begin{aligned} E_{\mathcal{T}}[\hat{y}_0 - E_{\mathcal{T}}\hat{y}_0]^2 &= E_{\mathcal{T}}[E_{\mathcal{T}}(\hat{y}_0 - E_{\mathcal{T}}\hat{y}_0)^2|x_0] \\ &= E_{\mathcal{T}}(\mathrm{Var}(\hat{y}_0|x_0)) \\ &= E_{\mathcal{T}}x_0^T\,\mathrm{Var}(\hat{\beta})x_0 \\ &= E_{\mathcal{T}}x_0^T(X^TX)^{-1}\sigma^2x_0 \\ &= E_{\mathcal{T}}x_0^T(X^TX)^{-1}x_0\sigma^2 \end{aligned} \tag{1.26}$$

那么为什么 $\mathrm{Var}(\hat{\beta}) = (X^TX)^{-1}$？ 证明：我们已经得到系数 β 的估计为 $(X^TX)^{-1}X^TY$，将 $Y = X\beta + \epsilon$ 带入得到

$$\hat{\beta} = \beta + (X^TX)^{-1}X^T\epsilon \tag{1.27}$$

于是我们有：

$$\begin{aligned} \mathrm{Var}(\hat{\beta}) &= (X^TX)^{-1}X^T\,\mathrm{Var}(\epsilon)X(X^TX)^{-1} \\ &= \sigma^2(X^TX)^{-1} \end{aligned} \tag{1.28}$$

这里我们在预测误差中引入了额外的方差 σ^2，因为我们的目标是不确定的。这里没有偏差，并且方差是依赖于 x_0 的。如果 N 很大 (那么我们抽样得到的 $(X^TX)^{-1}$ 可以代表真正的，下文推理用得到) 并且 $\mathcal{T}$ 是随机挑选

的，假设 $E(X) = 0$，然后 $X^T X \to N\,\mathrm{Cov}(X)$ 并且

$$\begin{aligned} E_{x_0}\,\mathrm{EPE}(x_0) &\sim E_{x_0} x_0^T\,\mathrm{Cov}(X)^{-1} x_0 \sigma^2/N + \sigma^2 \\ &= \mathrm{trace}[\mathrm{Cov}(X)^{-1}\,\mathrm{Cov}(x_0)]\sigma^2/N + \sigma^2 \\ &= \sigma^2(p/N) + \sigma^2 \end{aligned} \tag{1.29}$$

对于倒数第二步到最后一步我们用到结论：

$$\begin{aligned} E[x^T A x] &= E[\mathrm{tr}(x^T A x)] \\ &= \mathrm{tr}(E(x^T A x)) \\ &= \mathrm{tr}(E[x x^T] A) \\ &= \mathrm{tr}((\Sigma + E[x]E[x]^T) A) \\ &= \mathrm{tr}(\Sigma A) + \mathrm{tr}(E[x]E[x]^T A) \\ &= \mathrm{tr}(A\Sigma) + \mathrm{tr}(E[x]^T A E[x]) \\ &= \mathrm{tr}(A\Sigma) + E[x]^T A E[x] \end{aligned} \tag{1.30}$$

上式式子主要用到期望和迹的线性性质和迹的交换性：

$$\mathrm{tr}(AB) = \mathrm{tr}(BA) \tag{1.31}$$

$$\mathrm{tr}(E(x)) = E[\mathrm{tr}(x)] \tag{1.32}$$

从这里我们可以看出期望 EPE 是 p 的线性递增函数，斜率为 σ^2/N。如果 N 很大或者 σ^2 很小，这种差异增长可以忽略。通过对拟合的模型类别施加一些严格的限制，我们避免了维数灾难。

对于公式1.25和公式1.29的一些解释：方差是基于所有的训练集 $\mathcal{T}$ 和所有 y_0 的值，同时 x_0 已经确定了。注意 x_0 和 y_0 的选择是与 $\mathcal{T}$ 独立的，所以我们有:$E_{y_0|x_0} E_{\mathcal{T}} = E_{\mathcal{T}} E_{y_0|x_0}$。另外还有 $E_{\mathcal{T}} = E_{\mathcal{X}} E_{\mathcal{Y}|\mathcal{X}}$

我们将 $y_0 - \hat{y}_0$ 写为三项之和

$$(y_0 - x_0^T\beta) - (\hat{y}_0 - E_{\mathcal{T}}(\hat{y}_0)) - (E_{\mathcal{T}}(\hat{y}_0) - x_0^T\beta) = U_1 - U_2 - U_3$$

首先对于 U_1，我们有 $E_{y_0|x_0}U_1 = 0, E_{\mathcal{T}}U_1 = U_1E_{\mathcal{T}}$。对于 U_2，我们有 $E_{y_0|x_0}U_2 = U_2E_{y_0|x_0}, E_{\mathcal{T}}U_2 = 0$ 另外根据上文提到的公式 $\hat{y}_0 = x_0^T\hat{\beta} = x_0^T\beta + x_0^T(X^TX)^{-1}X^T\epsilon$ 我们有

$$U_3 = E_{\mathcal{T}}(\hat{y}_0) - x_0^T\beta = x_0^TE_{\mathcal{X}}((X^TX)^{-1}X^TE_{\mathcal{Y}|\mathcal{X}}\epsilon) = 0$$

因为 $E_{\mathcal{Y}|\mathcal{X}}(\epsilon) = 0$，所以 U_3 等于 0。

现在我们将剩余的两部分平方并应用 $E_{y_0|x_0}E_{\mathcal{T}}$。$U_1$ 和 U_2 交叉项为零，因为 $E_{y_0|x_0}(U_1U_2) = U_2E_{y_0|x_0}(U_1) = 0$

现在我们还剩下两项平方项，根据方差的定义我们有 $E_{y_0|x_0}E_{\mathcal{T}}U_1^2 = \mathrm{Var}(y_0|x_0) = \sigma^2$。现在只剩下 $E_{\mathcal{T}}(\hat{y}_0 - E_{\mathcal{T}}\hat{y}_0)^2 = \mathrm{Var}(\hat{y}_0)$ 这一项。因为 $U_3 = 0$，我们有 $E_{\mathcal{T}}\hat{y}_0 = x_0^T\beta$。

如果 m 是一个以 μ 为元素的 1×1 矩阵，则 mm^T 是以 μ^2 为元素的 1×1 矩阵。还是根据公式 $\hat{y}_0 = x_0^T\hat{\beta} = x_0^T\beta + x_0^T(X^TX)^{-1}X^T\epsilon$，最后剩余的一项方差就等同于：

$$E_{\mathcal{T}}(x_0^T(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}x_0)$$

因为 $E_{\mathcal{T}} = E_{\mathcal{X}}E_{\mathcal{Y}|\mathcal{X}}$，并且 $\epsilon\epsilon^T$ 的期望是 σ^2I_N，这就等同于

$$\sigma^2x_0^TE_{\mathcal{T}}((X^TX)^{-1})x_0 = \sigma^2x_0^TE_{\mathcal{X}}((X^TX/N)^{-1})x_0/N$$

我们假设 X 和 x_0 的均值为 0。对于较大的 N，X^TX/N 近似于 $\mathrm{Cov}(X) = \mathrm{Cov}(x_0)$，对于期望 $E_{\mathcal{X}}$ 来说，$(X^TX/N)^{-1}$ 是一个常数。因此根据 x_0 对上式求期望，得：

$$\begin{aligned}
\sigma^2\mathrm{E}_{x_0}\left(x_0^T\,\mathrm{Cov}(X)^{-1}x_0\right)/N &= \sigma^2\mathrm{E}_{x_0}\left(\mathrm{trace}\left(x_0^T\,\mathrm{Cov}(X)^{-1}x_0\right)\right)/N \\
&= \sigma^2\mathrm{E}_{x_0}\left(\mathrm{trace}\left(\mathrm{Cov}(X)^{-1}x_0x_0^T\right)\right)/N \\
&= \sigma^2\,\mathrm{trace}\left(\mathrm{Cov}(X)^{-1}\,\mathrm{Cov}\left(x_0\right)\right)/N \\
&= \sigma^2\,\mathrm{trace}\left(I_p\right)/N \\
&= \sigma^2p/N
\end{aligned}$$

图1.4在两种情况下比较了 1 最近邻法和最小二乘法，两种情况下 $Y = f(X) + \epsilon$，X 均匀分布并且 $\epsilon \sim N(0,1)$。样本大小 $N = 500$。对于橙色的曲线，$f(x)$ 是线性的，而对于蓝色曲线则是非线性的。图1.4展示的最小二乘法和 1 最近邻方法的 EPE 比值。在线性模型中比值大致是从 2 开始的，并且最小二乘法在此情况下是无偏的。1− 最近邻的 EPE 总是大于 2，因为在这种情况下 $\hat{f}(x_0)$ 的方差至少是 σ^2，并且随着最近邻偏离目标点，该比率随着维数增加。对于非线性情况，最小二乘是有偏差的，这调节了比率。很明显，我们可以制造这样的例子: 最小二乘法的偏差将大于方差，1− 最近邻将成为赢家。

对于图1.4中线性条件下我们再讨论一下：

对于 1NN

$$\mathrm{EPE}\left(x_0\right) = \sigma^2 + \mathrm{Var}\,\mathcal{T}\left(\hat{y}_0\right) + \mathrm{Bias}^2\left(\hat{y}_0\right) \gtrsim \sigma^2 + \sigma^2 + c \geq 2\sigma^2$$

而对于 OLS:

$$\mathrm{E}_{x_0}\,\mathrm{EPE}\left(x_0\right) = \sigma^2(p/N) + \sigma^2 \approx \sigma^2$$

则 EPE 比率大致为 2，又因为 1−NN 中的方差项（第二项）随 p 增大的速率大于 OLS 中方差项（第一项），则 EPE 比率也会缓慢随 p 增长。

1.4 统计模型和函数逼近

我们的目标是找到一个函数 $\hat{f}(x)$ 来近似真实函数 $f(x)$。

联合概率密度函数 $\Pr(X, Y)$ 的统计模型

假设我们的数据从统计模型

$$Y = f(X) + \epsilon \tag{1.33}$$

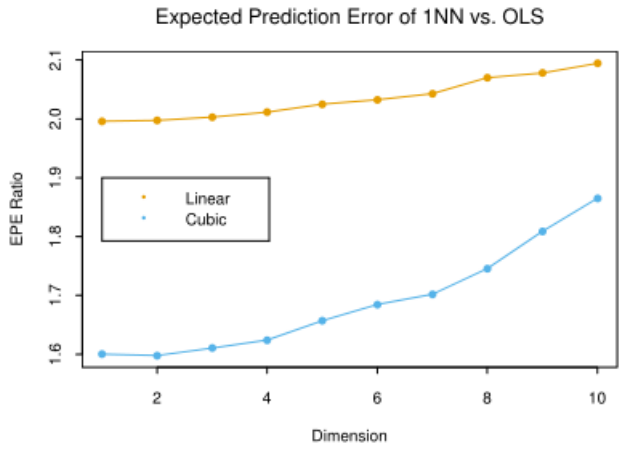

图 1.4: 1− 最近邻和最小二乘法比较，模型都是 $Y = f(X) + \epsilon$，对于橙色的曲线，$f(x) = x_1$，而蓝色的曲线 $f(x) = \frac{1}{2}(x_1 + 1)^3$

中产生，随机误差 ϵ，其 $E(\epsilon) = 0$ 并且与 X 相互独立。对于这个模型，$f(x) = E(Y|X = x)$，实际上条件分布 $\Pr(Y|X)$ 仅通过条件均值 $f(x)$ 依赖于 X。

实际上公式1.4所说的 ϵ 独立于 X 的假设并不是必须的。例如，我们可以假设 $\mathrm{Var}(Y|X = x) = \sigma(x)$，这样均值和方差就都依赖于 X。

到现在为止我们讨论的都是响应变量为定量变量的例子。误差可加模型一般不用于定性的输出变量 G，这种情况下目标函数是条件密度 $\Pr(G|X)$，这是直接建模的。举个例子，对于二分类问题，一种情况的输出概率为 $p(X)$，另一种为 $1-p(X)$。因此，如果 Y 是 $0-1$ 编码的，然后 $E(Y|X = x) = p(x)$，但是方差同样依赖 x:$\mathrm{Var}(Y|X = x) = p(x)(1 - p(x))$。

函数逼近

我们遇到的许多近似值斗鱼一组参数 θ 有关，这些参数可以修改来拟

合数据。一种有用的近似可以表示为线性基展开

$$f_\theta(x) = \sum_{k=1}^{K} h_k(x)\theta_k \tag{1.34}$$

h_k 是输入向量的一系列合适的函数或者变换。传统的例子是多项式或三角函数，如 $x_1^2, x_1x_2^2, \cos(x)$ 或者其他的。我们也可能遇到非线性的表达式，如神经网络的激活函数：

$$h_k(x) = \frac{1}{1+\exp(-x^T\beta_k)} \tag{1.35}$$

我们可以用最小二乘法按照我们处理线性模型的方法来估计 f_θ 的参数 θ，通过最小化平方误差和：

$$\mathrm{RSS}(\theta) = \sum_{i=1}^{N}(y_i - f_\theta(x_i))^2 \tag{1.36}$$

为 θ 的函数。

尽管最小二乘法应用起来非常方便，但是在某些情况下并不能得到很好的结果。一个应用更加广泛的方法是极大似然估计。假设我们有从由一些参数 θ 决定的概率密度函数 $\mathrm{Pr}_\theta(y)$ 随机样本 $y_i, i = 1, \cdots, N$。观测样本的对数概率为：

$$L(\theta) = \sum_{i=1}^{N} \log \mathrm{Pr}_\theta(y_i) \tag{1.37}$$

一个更有趣的例子是对于定性的输出 G 的函数 $\mathrm{Pr}(G|X)$ 的概率。假设我们有一个模型，给定 X，每一类的概率为 $operatornamePr(G = \mathcal{G}|X = x) = p_{k,\theta}(x), k = 1, \cdots, K$。然后对数概率（也被称作交叉熵）为：

$$L(\theta) = \sum_{i=1}^{N} \log p_{g_i,\theta}(x_i) \tag{1.38}$$

，求解使 L 最大化的参数 θ。

1.5 结构化回归模型

对任意函数 f，考虑 RSS 准则

$$\mathrm{RSS}(f) = \sum_{i=1}^{N}(y_i - f(x_i))^2 \tag{1.39}$$

最小化公式1.4可以有无穷多个解，任何通过训练点 (x_i, y_i) 的 $\hat{f}$ 都可以看作是一个解。如果在每一个点 x_i 有很多个 $y_{il}, l = 1, \cdots, N_i$ 与其对应，在这种情况下，得到的 $\hat{f}$ 会穿过 y_{il} 的中点。

一般地，大多数学习方法施加的约束条件都可以描述为对复杂度的限制。这也意味着在输入空间的小邻域的一些规则的行为。这就是，对于在某种度量下充分互相接近的所有输入点 x, $\hat{f}$ 展现了一些特殊结构比如说接近常值，线性或者低次的多项式。然后在邻域内平均或者进行多项式拟合得到估计量。

现在必须澄清一个事实。任何在各向同性的邻域中试图产生局部变化的函数会在高维中遇到问题——还是维数灾难。相反地，所有克服维数问题的方法在衡量邻域时有一个对应的度量，该度量的基本要求是不允许邻域在各个方向都同时小。

1.6 限制性估计的种类

非参回归技巧的多样性或学习方法的类型根据其限制条件的本质可以分成不同的种类。这些种类不是完全不同的，而且确实一些方法可以归为好几种不同的类别。每个类都有与之对应的一个或多个参数，有时恰当地称之为光滑化参数，这些参数控制着局部邻域的有效大小。这里我们描述三个大的类别。

粗糙度惩罚和贝叶斯方法

这是由显式的惩罚 $\mathrm{RSS}(f)$ 以及粗糙度惩罚控制的函数类别：

$$\mathrm{PRSS}(f;\lambda) = \mathrm{f} + \lambda J(f) \tag{1.40}$$

对于在输入空间的小邻域变换太快的函数 f，用户选择的函数 $J(f)$ 会变大。举个例子，著名的用于一维输入的三次光滑样条的是带惩罚的最小二乘的准则的解

$$\mathrm{PRSS}(f;\lambda) = \sum_{i=1}^{N}(y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx \tag{1.41}$$

核方法与局部回归

这些方法可以认为是通过确定局部邻域的本质来显式给出回归函数的估计或条件期望，并且属于局部拟合得很好的规则函数类。局部邻域由核函数 $K_\lambda(x_0, x)$ 确定，它确定了 x_0 邻域内的点 x 的权重。例如，高斯核函数有一个基于高斯密度函数的权重函数：

$$K_\lambda(x_0, x) = \frac{1}{\lambda} \exp[-\frac{||x - x_0||^2}{2\lambda}] \tag{1.42}$$

参数 λ 为高斯密度函数的方差，同时也控制住邻域的大小。最简单的核估计方法为 Nadaraya-Watson 系数平均：

$$\mathrm{RSS}(f_\theta, x_0) = \frac{\sum_{i=1}^N K_\lambda(x_0, x_i) y_i}{\sum_{i=1}^N K_\lambda(x_0, x_i)} \tag{1.43}$$

一般地，我们可以定义 $f(x_0)$ 的局部回归估计为 $f_{\hat{\theta}}(x_0)$，其中 $\hat{\theta}$ 使下式最小化：

$$\mathrm{RSS}(f_\theta, x_0) = \sum_{i=1}^N K_\lambda(x_0, x_i)(y_i - f_\theta(x_i))^2 \tag{1.44}$$

最近邻方法可以看作是某个更加依赖数据的度量的核方法。k 最近邻的度量方法为：

$$K_k(x, x_0) = I(||x - x_0|| \leq ||x_{(k)} - x_0||) \tag{1.45}$$

其中 $x_{(k)}$ 是训练观测值离 x_0 的距离第 k 近的观测，而且 $I(S)$ 是集合 S 的指标函数。

基函数和字典方法

这个方法的类别包括熟悉的线性和多项式展开式，但是最重要的有多种多样的更灵活的模型．这些关于 f 的模型是基本函数的线性展开:

$$f_\theta(x) = \sum_{m=1}^M \theta_m h_m(x) \tag{1.46}$$

其中每个输入 h_m 都是输入 x 的函数，并且其中的线性项与参数 θ 有关。

对于一维 x，阶为 K 的多项式样条可以通过 M 个样条基函数合适的序列来表示，从而由 $M-K-1$ 个结点来确定。在节点中间产生阶为 K 的分段多项式函数，并且在每个结点处由 $K-1$ 阶连续函数来连接。

径向基函数是在质心处对称的 p 维核，

$$f_\theta(x)=\sum_{m=1}^{M}K_{\lambda_m}(\mu_m,x)\theta_m \tag{1.47}$$

单层的向前反馈的带有线性输出权重的神经网络模型可以认为是一种自适应的基函数方法．模型有如下形式:

$$f_\theta(x)=\sum_{m=1}^{M}\beta_m\sigma(\alpha_m^Tx+b_m) \tag{1.48}$$

其中 $\sigma(x)$ 为激活函数。

这些自适应选择基函数的方法也被称为字典方法，其中有一个可用的候选基函数的可能无限集或字典 $\mathcal{D}$ 可供选择。

1.7 模型的选择和偏差-方差权衡

我们以 k 最近邻回归的拟合值 $\hat{f}_k(x_0)$ 作为例子。假设数据来自模型 $Y=f(X)+\epsilon, E(\epsilon)=0, \mathrm{Var}(\epsilon)=\sigma^2$。我们假设 x_0 提前确定。在 x_0 处的期望预测误差，也被称为测试或泛化误差：

$$\begin{aligned}\mathrm{EPE}_k(x_0)&=\mathrm{E}\left[\left(Y-\hat{f}_k(x_0)\right)^2\mid X=x_0\right]\\&=\sigma^2+\left[\mathrm{Bias}^2\left(\hat{f}_k(x_0)\right)+\mathrm{Var}_{\mathcal{T}}\left(\hat{f}_k(x_0)\right)\right]\\&=\sigma^2+\left[f(x_0)-\frac{1}{k}\sum_{\ell=1}^{k}f\left(x_{(\ell)}\right)\right]^2+\frac{\sigma^2}{k}\end{aligned} \tag{1.49}$$

带括号的下标 (ℓ) 表示 x_0 的最近邻的顺序。

在展开式中有三项．第一项 σ^2 是不可约减的。误差是新测试目标点的方差，而且我们不能够控制。第二项和第三项在我们的控制范围内，并且构成了估计 $f(x_0)$ 时 $\hat{f}_k(x_0)$ 的均方误差，均方误差经常被分解成偏差部分和方差部分。如果真实的函数相当地光滑，这一项很可能随着 k 的增加而增

加。对于较小的 k 值和较少的近邻点会导致值 $f(x_{(\ell)})$ 与 $f(x_0)$ 很接近，所以它们的平均应该距离 $f(x_0)$ 很近。当 k 值增加，邻域远离，然后任何事情都可能发生。这里的方差项是方差的简单平均，因为 k 的倒数关系，随 k 变大而变小．所以当 k 变化，会有偏差——误差的权衡。

更一般地，随着我们过程的模型复杂度增加，方差趋于上升，偏差趋于下降。当模型复杂度下降时会发生相反的行为。对于 k-最近邻，模型复杂度由 k 来控制。

1.8 习题

习题 1.1 假设我们有一个 K 分类问题，每个类表示为向量 t_k。t_k 中除了第 k 个位置的元素为 1 以外，其余元素均为 0。证明如果 $\hat{y}$ 的和为 1，那么找出 $\hat{y}$ 中最大元素的问题等同于解：

$$\min_k ||t_k - \hat{y}||$$

解 1.1

我们假设 $||\cdot||$ 为欧几里得范式

$$\arg\min_k ||t_k - \hat{y}|| = \arg\min_k ||t_k - \hat{y}||_2^2$$

因此，

$$\begin{aligned} ||t_k - \hat{y}||_2^2 &= t_k^T t_k - 2t_k^T \hat{y} + \hat{y}^T \hat{y} \\ &= 1 + \hat{y}^T \hat{y} - 2\hat{y}_k \end{aligned}$$

因为 $1 + \hat{y}^T \hat{y}$ 是常数，所以

$$\begin{aligned} \arg\min_k ||t_k - \hat{y}||_2^2 &= \arg\min_k -2\hat{y}_k \\ &= \arg\max_k \hat{y}_k \end{aligned}$$

所以原问题等价于求 $\hat{y}$ 中最大元素。

习题 1.2 首先我们先从高斯分布 $N((0,1)^T, I)$ 中产生了 10 个均值 m_k，标记为蓝色；之后从高斯分布 $N((1,0)^T, I)$ 中产生了 10 个均值标记为橙色。

然后对于每个类，我们按照如下方法每类产生 100 个数据：首先我们以 1/10 的概率随机选取一个均值 m_k，然后以正态分布 $N(m_k, I/5)$ 产生数据。如何求解该数据的贝叶斯决策边界。

解 1.2

因为两类点的数量相同，所以两类点的先验概率为：

$$P(\mathcal{G}_{\text{BLUE}}) = P(\mathcal{G}_{\text{ORANGE}})$$

而各自十个被选中的概率

$$P(m = p_i|\mathcal{G}_{\text{BLUE}}) = \frac{1}{10}, i = 1, \cdots, 10$$
$$P(m = q_i|\mathcal{G}_{\text{ORANGE}}) = \frac{1}{10}, i = 1, \cdots, 10$$

而似然概率为：

$$\begin{aligned} P\left(X = x \mid \mathcal{G}_{BLUE}\right) &= \sum_{i=1}^{10} P\left(X = x \mid \mathcal{G}_{BLUE}, m = p_i\right) P\left(m = p_i \mid \mathcal{G}_{BLUE}\right) \\ &= \sum_{i=1}^{10} (2\pi)^{-k/2} \left|\frac{I}{5}\right|^{-1/2} \exp\left(-\frac{1}{2}\left(x - p_i\right)^T \left(\frac{I}{5}\right)^{-1} \left(x - p_i\right)\right) \frac{1}{10} \\ &= (2\pi)^{-k/2} \left|\frac{I}{5}\right|^{-1/2} \frac{1}{10} \sum_{i=1}^{10} \exp\left(-\frac{5}{2}\left\|x - p_i\right\|_2^2\right) \end{aligned}$$

类似的，

$$P(X = x|\mathcal{G}_{\text{ORANGE}}) = (2\pi)^{-k/2}|\frac{I}{5}|^{-1/2}\frac{1}{10}\sum_{i=1}^{10}\exp(-\frac{5}{2}||x - q_i||_2^2)$$

根据贝叶斯公式，可知后验概率为：

$$P\left(\mathcal{G}_{BLUE} \mid X = x\right) = \frac{P\left(X = x \mid \mathcal{G}_{BLUE}\right) P\left(\mathcal{G}_{BLUE}\right)}{P(X = x)}$$

$$P\left(\mathcal{G}_{ORANGE} \mid X = x\right) = \frac{P\left(X = x \mid \mathcal{G}_{ORANGE}\right) P\left(\mathcal{G}_{ORANGE}\right)}{P(X = x)}$$

在决策边界处后验概率相等，因此

$$P(\mathcal{G}_{\text{BLUE}}|X = x) = P(\mathcal{G}_{\text{ORANGE}}|X = x)$$

带入式子化简得到：

$$\sum_{i=1}^{10}\exp(\frac{5}{2}||x-p_i||_2^2)=\sum_{i=1}^{10}\exp(\frac{5}{2}||x-q_i||_2^2)$$

习题 1.3 证明式子*1.22*

解 1.3

解法 1

对于本练习，我们将导出从原点到 n 个点 x_i 中最近的点的欧几里得距离 *(*用 y 表示*)* 的分布函数，其中每个点 x_i 在以原点为中心的 p 维单位球体内均匀分布。

对于任意给定的向量 x_i，$y=||x_i||$ 的分布函数是半径为 y 的球的体积与半径为 1 的球的体积之比。这个比值为 y^p，所以 $F(y)=y^p$。所以 y 的概率密度函数 $f(y)=py^{p-1}$。

给定 N 个 $\{x_i\}_{i=1}^N$ 这样的向量，则最小半径 Y_1 的分布函数为

$$F_{Y_1}(y)=1-(1-F(y))^N=1-(1-y^p)^N$$

关于顺序统计量

首先我们考虑事件 $\{X_r\le x\le X_{r+1}\}$ 的概率，这等价于事件 $(X_1,X_2,\cdots,X_n)$ 中有 r 个小于等于 x，$n-r$ 个大于 x。故其概率为：

$$\left(\begin{array}{c}n\\r\end{array}\right)F^r(x)[1-F(x)]^{n-r}$$

记事件 $\{X_r\le x\le X_{r+1}\}$ 为 A_r，为了符号统一记时间 $\{X_n\le x\}$ 为 A_n。则事件 $\{X_r\le x\}$ 事实上等于 $\cup_{i=r}^n A_i$，且容易验证，这些事件两两不相容。故：

$$F(X_r)=P(X_r\le x)=P(\cup_{i=r}^n A_i)=\sum_{i=r}^n\left(\begin{array}{c}n\\i\end{array}\right)F^i(x)[1-F(x)]^{n-i}$$

回到原问题，因为我们考虑的是最小的距离，故 $r=1$，故

$$\begin{aligned}F_{Y_1}(y) &= \sum_{i=1}^{n}\left(\begin{array}{c} n \\ i \end{array}\right) F^i(x)[1-F(x)]^{n-i} \\ &= \sum_{i=0}^{n}\left(\begin{array}{c} n \\ i \end{array}\right) F^i(x)[1-F(x)]^{n-i} - (1-(1-F(x))^N) \\ &= 1-(1-F(x))^N\end{aligned}$$

所以最短距离的中位数为

$$F_{Y_1}(y) = \frac{1}{2}$$

即

$$y = (1-(\frac{1}{2})^{1/N})^{1/p} \equiv d_{\text{median}}(p,N)$$

然后我们也可以求出这 N 个点最近距离的平均数,$F_{Y_1}(y)$ 的概率密度函数为：

$$f_{Y_1}(y) = N(1-y^p)^{N-1}(py^{p-1}) = pN(1-y^p)^{N-1}y^{p-1}$$

$$d_{\text{mean}}(p,N) \equiv \int_0^1 y f_{Y_1}(y)dy = pN\int_0^1 (1-y^p)^{N-1}y^p dy$$

我们无法精确地求解但是我们可以将其与 β 分布联系在一起

$$B(a,b) \equiv \int_0^1 t^{a-1}(1-t)^{b-a}dt$$

如果我们令 $t=y^p$，我们得到

$$d_{\text{mean}}(p,N) = N\int_0^1 (1-t)^{N-1}t^{\frac{1}{p}}dt = NB(1+\frac{1}{p}, N+\frac{1}{p})$$

解法 2[1]

我们用 $(x_1,\cdots,x_N)$ 来定义数据的 N 元组。令 $r_i=||x_i||$。令 $U(A)$ 是所有满足 $A<r_1<\cdots<r_N<1$ 的 N 元组的集合。所有的 N 元组是 $N!$ 个不相交集合 $U(0)$ 的并集（通过置换 $1\cdots N$ 的索引）。同样的 A 将适用于我们的每个 $N!$ 不相交的子集，因此将给出到原点的最小距离 x_i 的中位数。

[1]没看懂，可能需要多元微积分知识，等学习了再来看

我们想要找到这样的 A

$$\int_{U(A)} dx_1 \cdots dx_N = \frac{1}{2} \int_{U(0)} dx_1 \cdots dx_N$$

我们把它转换为球面坐标。我们有：

$$\int_{A<r_1<\cdots<r_N<1} r_1^{p-1} \cdots r_N^{p-1} dr_1 \cdots dr_n = \frac{1}{2} \int_{0<r_1\cdots<r_N<1} r_1^{p-1} \cdots r_N^{p-1} dr_1 \cdots dr_n$$

我们令 $s_i = r_i^p$，得到

$$\int_{A^p<s_1<\cdots<s_N<1} ds_1 \cdots ds_n = \frac{1}{2} \int_{0<s_1<\cdots<s_N<1} ds_1 \cdots ds_n$$

对于左边的积分，我们做如下变化：

$$t_0 = s_1 - A^p, t_1 = s_2 - s_1, \cdots, t_{N-1} = s_N - s_{N-1}, t_N = 1 - s_N$$

雅可比矩阵（忽略冗余项 t_0）是一个对角线元素为 -1 的三角形矩阵。它的行列式的绝对值为 1，用于改变积分的变量公式。

我们积分的区域是

$$\sum_{i=0}^{N} t_i = 1 - A^p, t_i > 0$$

是一个按照因子 $(1 - A^p)$ 缩小的 N 维单纯形 *(which is an N-dimensional simplex scaled by a factor$(1 - A^p)$)*。右边的积分通过同样的方式处理，令 $A = 0$。因为积分区域是一个 N 维的，所以要乘以 $(1 - A^p)^N$。我们通过求解 $(1 - A^p)^N = 1/2$ 来求解 A。

习题 1.4 边缘效应问题不止存在于有界区域的均匀抽样。考虑输入是从一个球面多元正态分布 $(X \sim N(0, I_p))$ 产生的。从采样点到原点的平方距离服从均值为 p 的 χ_p^2 分布[2]。考虑一个从该分布产生的预测点 x_0，并且令 $a = x_0/||x_0||$ 作为单位向量。令 $z_i = a^T x_i$ 作为所有训练点在此点上的投影。

证明 z_i 服从 $N(0, 1)$ 分布，并且与原点的期望平方距离为 1，而目标点与原点的期望平方距离为 p。

[2]卡方分布相当于 n 个标准正态分布的平方相加，其均值为 n。因此 p 维距离平方相当于 p 个标准正态分布平凡相加，故均值为 p。

当 $p = 10$ 时，随机抽取的测试点距离原点大约 3.1 个标准差，而所有训练点沿 a 的方向平均为一个标准差。所以大多数预测点认为自己位于训练集的边缘。

解 1.4

因为 $X \sim N(0, I_p)$，所以 $x_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ 各元素之间是相互独立的，且都服从于 $N(0,1)$。假设 $x_0/||x_0|| = (a_1, a_2, \cdots, a_p), \sum_{i=1}^{p} a_i^2 = 1$。所以

$$\frac{x_0 x_i}{||x_0||} = a_1 x_{i1} + a_2 x_{i2} + \cdots + a_p x_{ip}$$

多个正态分布相加，其均值等于多个正态分布均值相加，故 z_i 的均值为 $a_1 \times 0 + \cdots + a_p \times 0 = 0$。其方差等于方差相加,故其方差为 $a_1^2 \times 1 + \cdots + a_p^2 \times 1 = \sum_{i=1}^{p} a_i^2 = 1$。所以 $z_i \sim N(0,1)$。

习题 1.5 考虑一个以 x_i 为输入以 y_i 为输出的回归问题，通过最小二乘法来估计参数模型 $f_\theta(x)$。证明如果相同的 x 值存在不同的 y 值，则可以通过加权最小二乘法得到。

解 1.5 对于最小二乘法，当 x 存在重复时，最小二乘法就变成了一个加权最小二乘法。这种讨论的动机通常是希望在实验上得到模型的准确误差 ϵ

$$y = f(x) + \epsilon$$

要做到这一点，一种方法是执行许多实验，观察当每次实验产生相同的 x 值时，数据生成过程产生的不同 y 值。如果 $\epsilon = 0$ 我们期望每次实验得到的结果是相同的。让 N_u 为 x 不重复的值的个数。假设第 i 个不重复的 x 的值产生 n_i 个不同的 y 值。通过这些符号我们可以得到：

$$\mathrm{RSS}(\theta) = \sum_{i=1}^{N_u} \sum_{j=1}^{n_i} (y_{ij} - f_\theta(x_i))^2$$

将上式进一步展开我们得到：

$$\begin{aligned}
\mathrm{RSS}(\theta) &= \sum_{i=1}^{N_u} \sum_{j=1}^{n_i} (y_{ij}^2 - 2 f_\theta(x_i) y_{ij} + f_\theta(x_i)^2) \\
&= \sum_{i=1}^{N_u} n_i \left(\frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{2}{n_i} f_\theta(x_i) \left(\sum_{j=1}^{n_i} y_{ij}\right) + f_\theta(x_i)^2\right)
\end{aligned}$$

令 $\overline{y}_i \equiv \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$，这样我们得到：

$$\mathrm{RSS}(\theta) = \sum_{i=1}^{N_u} n_i(\overline{y}_i - f_\theta(x_i))^2 + \sum_{i=1}^{N_u}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{N_u} n_i\overline{y}_i^2$$

一旦接受了测量结果，采样点 y 就被固定了并且不会改变。因此最小化上述公式就变成了最小化

$$\mathrm{RSS}(\theta) = \sum_{i=1}^{N_u}(\overline{y}_i - f_\theta(x_i))^2$$

这个问题就变成了加权最小二乘法，因为输入向量 x_i 被用来拟合平均值 $\overline{y}_i$ 并且每一项的残差通过 x_i 出现多少次来进行加权。这是一个简化的问题因为 $N_u < N$。

习题 1.6 假设我们有 N 个数据对 x_i, y_i 从下列分布中独立同分布产生的

1. $x_i \sim h(x)$，概率密度
2. $y_i = f(x_i) + \epsilon_i$, f 是回归函数
3. $\epsilon_i \sim (0, \sigma^2)$

我们构建了 y_i 中线性函数 f 的一个估计量

$$\hat{f}(x_0) = \sum_{i=1}^{N} \ell_i(x_0;\chi)y_i$$

$\ell_i(x_0;\chi)$ 独立于 y_i，但是依赖于整个训练数据集 x_i, 即 χ

a. 证明线性回归和 k 最近邻法都是这种估计方法的一种。明确描述这两种情况下的权重 $\ell_i(x_0;\chi)$

b. 分解下列条件概率 $E_{\mathcal{Y}|\mathcal{X}}(f(x_0) - \hat{f}(x_0))^2$ 为条件平方偏差个条件方差两部分。

c. 分解非条件概率均方误差 $E_{\mathcal{Y},\mathcal{X}}(f(x_0) - \hat{f}(x_0))^2$ 为平方偏差和方差

d. 建立上述两种情况下的平方偏差和方差之间的关系

解 1.6

为了简化这个问题我们认为只有一个响应变量 y 和一个预测变量 x。因此 y 和 X 的标准定义为：

$$y^T = (y_1, \cdots, y_n)$$

$$X^T = \begin{bmatrix} 1 \cdots 1 \\ x_1 \cdots x_n \end{bmatrix}$$

问题 *1:*

让我们先考虑线性回归。根据以前的知识，我们有 $\hat{\beta} = (X^T X)^{-1} X^T y$，然后

$$\hat{f}(x_0) = [x_0 \quad 1]\hat{\beta} = [x_0 \quad 1](X^T X)^{-1} X^T y$$

因此从上式可以得出：

$$\ell_i(x_0; \chi) = [x_0 \quad 1](X^T X)^{-1} \begin{bmatrix} 1 \\ x_i \end{bmatrix}, \forall 1 \leq i \leq n$$

对于 k 最近邻，$\ell_i(x_0; \chi) = 1/k$，如果 x_i 是 k 个最近的点之一，否则为 0。

问题 *2*

在这里 $\mathcal{X}$ 是固定的，$\mathcal{Y}$ 是变量。另外 x_0 和 $f(x_0)$ 也是固定的，所以

$$\mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\left(f(x_0) - \hat{f}(x_0)\right)^2\right) = f(x_0)^2 - 2.f(x_0) \cdot \mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\hat{f}(x_0)\right) + \mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\left(\hat{f}(x_0)\right)^2\right)$$

$$= \left(f(x_0) - \mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\hat{f}(x_0)\right)\right)^2 + \mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\left(\hat{f}(x_0)\right)^2\right) - \left(\mathrm{E}_{\mathcal{Y}|\mathcal{X}}\left(\hat{f}(x_0)\right)\right)^2$$

$$= (\mathrm{bias})^2 + \mathrm{Var}\left(\hat{f}(x_0)\right)$$

问题 *3*

计算方法和问题 *2* 大致相同，除了 $\mathcal{X}$ 和 $\mathcal{Y}$ 均为变量，而 x_0 和 $f(x_0)$ 为常量

$$\mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\left(f(x_0) - \hat{f}(x_0)\right)^2\right) = f(x_0)^2 - 2.f(x_0) \cdot \mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\hat{f}(x_0)\right) + \mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\left(\hat{f}(x_0)\right)^2\right)$$

$$= \left(f(x_0) - \mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\hat{f}(x_0)\right)\right)^2 + \mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\left(\hat{f}(x_0)\right)^2\right) - \left(\mathrm{E}_{\mathcal{X},\mathcal{Y}}\left(\hat{f}(x_0)\right)\right)^2$$

$$= (\mathrm{bias})^2 + \mathrm{Var}\left(\hat{f}(x_0)\right)$$

习题 1.7 考虑一个有 p 个参数的线性回归模型,通过训练数据 $(x_1, y_1), \cdots, (x_N, y_N)$ 进行拟合。令 $\hat{\beta}$ 为最小二乘法估计的结果。假设我们有一些测试数据 $(\tilde{x}_1, \tilde{y}_1), \cdots, (\tilde{x}_M, \tilde{y}_N)$。如果 $R_{tr}(\beta) = \frac{1}{N} = \sum_1^N (y_i - \beta^T x_i)^2, R_{te}(\beta) = \frac{1}{M} \sum_1^M (\tilde{y}_i - \beta^T \tilde{x}_i)^2$，证明

$$E[R_{tr}(\hat{\beta})] \leq E[R_{te}(\hat{\beta})]$$

解 1.7

首先我们要先证明 $E[R_{te}(\hat{\beta})]$ 与 M 无关。令 $\mathrm{E}[\cdot]$ 表示任何随机事件的期望。我们有

$$\begin{aligned}
\mathbb{E}\left[R_{te}(\hat{\beta})\right] &= \mathbb{E}\left[\frac{1}{M}\sum_{i=1}^{M}\left(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i\right)^2\right] \\
&= \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[\left(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i\right)^2\right] \\
&= \frac{1}{M}\sum_{i=1}^{M}\left(\mathbb{E}\left[\tilde{y}_i^2\right] + \mathbb{E}\left[\left(\hat{\beta}^T\tilde{x}_i\right)^2\right] - 2\mathbb{E}\left[\tilde{y}_i\tilde{x}_i^T\hat{\beta}\right]\right)
\end{aligned}$$

因为所有测试数据都是独立同分布的。所以我们可以用 $E[y^2]$ 代替 $E[\tilde{y}_i^2]$。类似的，我们也可以用 $E[x^2]$ 和 $E[yx^T]$ 代替 $E[\tilde{x}_i^2]$ 和 $E[\tilde{y}_i\tilde{x}_i^T]$。此外，由于测试集和训练集是随机独立抽取的，因此我将使用塔式分解规则将样本的期望值 *(涉及训练集和测试集的术语)* 分解为期望值的乘积。我的解释如下：

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{\beta}^T\tilde{x}_i\right)^2\right] &= \mathbb{E}_{Train}\,\mathbb{E}_{Test\ |\ Train}\,\left(\hat{\beta}^T\tilde{x}_i\right)^2 \\
&= \mathbb{E}_{Train}\,\hat{\beta}^T\mathbb{E}_{Test}\,\left[x^2\right] \\
&= \mathbb{E}[\hat{\beta}]^T\mathbb{E}\left[x^2\right]
\end{aligned}$$

同样的，我们也可以得到

$$\mathbb{E}[R_{te}(\beta)] = \mathbb{E}[y^2] + \mathbb{E}[\hat{\beta}]^T\mathbb{E}[x^2] + \mathbb{E}[xy]^T\mathbb{E}[\hat{\beta}]$$

就像上面所证明的，期望测试误差与样本量 M 无关，这意味着我们可以取 N 个测试样本，即

$$\mathbb{E}[R_{te}(\hat{\beta})] = \mathbb{E}[\frac{1}{N}\sum_{i=1}^{N}(\tilde{y}_i - \hat{\beta}^T\tilde{x}_i)^2]$$

现在我们知道最小化测试集残差平方和的线性回归系数的估计不是 $\hat{\beta}$。相反，它将是通过在测试集上而不是在训练集上使用最小二乘进行拟合而得到的某个估计 $\hat{\beta}_{Test}$。

$$\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{y}_i-\hat{\beta}^T\tilde{x}_i\right)^2 \geq \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{y}_i-\hat{\beta}_{Test}^T\tilde{x}_i\right)^2$$

$$\Longrightarrow \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{y}_i-\hat{\beta}^T\tilde{x}_i\right)^2\right] \geq \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{y}_i-\hat{\beta}_{Test}^T\tilde{x}_i\right)^2\right]$$

$$\Longrightarrow \mathbb{E}\left[R_{te}(\hat{\beta})\right] \geq \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{y}_i-\hat{\beta}_{Test}^T\tilde{x}_i\right)^2\right]$$

现在数据 $(\tilde{x}_1,\tilde{y}_1,\cdots,(\tilde{x}_N,\tilde{y}_N))$ 而 $\hat{\beta}^{Test}$ 是通过这随机选择的 N 个数据估计出的一个线性回归参数，因此可以说下面的两个随机变量具有相同的分布。

$$\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{y}_i-\hat{\beta}_{Test}^T\tilde{x}_i\right)^2 \sim \frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{\beta}^T x_i\right)^2$$

$$\Longrightarrow \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{y}_i-\hat{\beta}_{Test}^T\tilde{x}_i\right)^2\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{\beta}^T x_i\right)^2\right]$$

$$\Longrightarrow \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left(\tilde{y}_i-\hat{\beta}_{Test}^T\tilde{x}_i\right)^2\right] = \mathbb{E}\left[R_{tr}(\hat{\beta})\right]$$

所以 $\mathbb{E}[R_{tr}(\hat{\beta})] \leq \mathbb{E}[R_{te}(\hat{\beta})]$