# COMP9417 - 机器学习教程。回归II

**问题1.最大似然估计(MLE)**

在这个问题中，我们将首先回顾，然后通过几个使用MLE技术进行参数估计的例子。如果你对MLE的概念已经很熟悉的话，下面的介绍可以跳过。

设置如下：我们对$n$个观测值（数据）进行抽样，我们用$X_1, X_2, \ldots \ldots$表示。$, X_n$，我们假设数据是从某个概率分布$P$中独立抽取的。

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P,$$

where i.i.d. stands for independent and identically distributed. In practice, we never have access to $P$, we are just able to observe samples from $P$ (namely $X_1, \ldots, X_n$), which we will use to learn something about $P$. In the simplest case, we assume that $P$ belongs to a parametric family. For example, if we assume that $P$ belongs to the family of normal distributions, then we are assuming that $P$ has a probability density function (pdf) of the form

$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad \theta = (\mu, \sigma^2)。\ \mu \in \mathbb{R}, \sigma > 0。$$

其中，我们通常将$\mu$称为均值，$\sigma^2$称为方差，我们将所有未知的参数合并为一个单一的参数向量$\theta$，它存在于某个参数空间$\Theta$中。在这个特定的例子中，$\Theta = \mathbb{R}$ $[0, )$。在这个假设下，如果我们知道$\theta$，那么我们就会知道$P$，因此，学习问题简化为学习可能的最佳参数$\theta^*$，因此被称为*参数化*。

继续这个例子，我们需要一种方法来量化一个特定的$\theta$的选择是多么好。为了做到这一点，我们首先回顾这样一个事实：对于独立的集合$A$、$B$、$C$来说，$P$（$A$和$B$和$C$）$=P$（$A$）$P$（$B$）$P$（$C$）。因此，我们有

$$\begin{aligned} \text{观察到} X_1, \ldots\ldots \text{的可能性。}, X_n &= \text{观察} X \text{的概率}_1 \times - - \times \text{观察} X \text{的概率}_n \\ &= p_\theta(X_1) \times - - \times p_\theta(X_n) \\ &= \prod_{i=1}^{n} p_\theta(X_i) \\ &=: L(\theta)。 \end{aligned}$$

我们把$L(\theta)$称为*可能性*，它是参数向量$\theta$的函数。我们把这个量解释为使用特定的参数选择时观察数据的概率。很明显，我们希望

选择能给我们带来最大可能的参数θ，即我们希望找到*最大可能的估计值*

$$\hat{\theta}_{MLE} := \arg\max_{\theta \in \Theta} L(\theta)。$$

由于这只是一个优化问题，我们可以依靠我们对微积分的了解来解决MLE估计器。

(a) 假设$X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, 1)$，也就是说，我们已经知道基础分布是正态，人口方差为1，但人口平均数不详。计算$\hat{\mu}_{MLE}$。
**提示：**用对数可能性来工作通常要容易得多，即解决优化问题。

$$\hat{\theta}_{MLE} := \arg\max_{\theta \in \Theta} \log L(\theta)。$$

这与解决原始问题的答案完全相同（为什么？）

**解决方案。**
这里的对数可能性是

$$\log L(m) = \log \left( \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(X_i - m)^2\right) \right)$$
$$= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i - m)^2$$

对$m$进行微分并设为零，可得到。

$$\frac{\partial}{\partial m}\log L(m) = \sum_{i=1}^{n}(X_i - m) = 0 \Rightarrow \hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$$

为了看到这确实是一个最大值，我们应该进行二次导数检验，从而得到。

$$\frac{\partial^2}{\partial m^2}\log L(m) = -n < 0。$$

(b) 假设$X_1, \ldots, X_n \overset{i.i.d.}{\sim} \text{Bernoulli}(p)$，计算$\hat{p}_{MLE}$。回顾一下，伯努利分布是离散的，有概率质量函数。

$$P(X = k) = p^k (1-p)^{1-k}, \quad k=0, 1 \; p \in [0, 1]。$$

**解决方案。**
注意这里的$\theta = p$，参数空间为$\Theta = [0, 1]$。我们构建的对数似然在

通常的方式。

$$\log L(p) = \log \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$$
$$= n\bar{X} \log p + n(1 - \bar{X}) \log(1 - p)$$

然后，进行微分并设为零，得到。

$$\frac{\partial}{\partial p} \log L(p) = 0 \Rightarrow \hat{p}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{X}$$

(c) **可选的。**假设$X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$。计算$(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE})$。

**解决方案。**

这里的对数可能性是

$$\log L(m, s^2) = \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi s^2}} \exp - \frac{(X_i - m)^2}{2s^2}$$
$$= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(s^2) - \frac{1}{2s^2} \sum_{i=1}^{n} (X_i - m)^2$$

为了同时求解两个MLE估计值，我们需要对两个参数的对数可能性进行微分，并将每个参数设为零，这将产生两个方程（即一个方程组）。同时求解这些方程就可以得到正确的解决方案。首先对$m$进行微分，并将其设为零，就可以得到。

$$\frac{\partial}{\partial m} \log L(m, s^2) = \frac{1}{s^2} \sum_{i=1}^{n} (X_i - m) = 0 \Rightarrow \hat{\mu}_{MLE} = \bar{X}.$$

请注意，在这种情况下，第一个方程不取决于第二个参数，所以我们可以直接解决它（这并不总是这样）。接下来，对$s^2$进行微分

$$\frac{\partial}{\partial s^2} \log L(m, s^2) = -\frac{n}{2s^2} + \frac{1}{2s^4} \sum_{i=1}^{n} (X_i - m)^2 = 0 \Rightarrow \hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (X_i - m)^2 = 0.$$

要解决这个问题，我们需要参考两个方程中的第一个，它告诉我们$m=\bar{X}$是最佳的，所以

$$(\hat{\mu}_{MLE}, \hat{\sigma}^2_{MLE}) = \left(\bar{X}, \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2\right) \quad 。$$

注意，为了在寻找一个函数的局部最大值（$t\hat{}1$，$t\hat{}2$）时做到完全严谨$F$（$t1$，$t2$），我们需要检查以下三个条件。

1. 在（$t\hat{}1, t\hat{}2$）的一阶偏导数为零。
2. 至少有一个二阶偏导数是负的
3. Hessian矩阵的行列式是正的。

我们不会在这里验证这些条件，这已经超出了课程的范围，但必须注意到我们遗漏了一些细节。

**问题2.估算器的偏差和方差**

在上一个问题中，我们讨论了MLE作为一种估计参数的方法。但是，估计一个参数的方法有无数种之多。例如，我们可以选择使用样本中位数而不是MLE。有一个框架是很有用的，在这个框架中我们可以系统地比较估计者，这就把我们带到了机器学习的两个核心概念：偏差和方差。假设真实参数是θ，我们有一个估计值θ^。请注意，一个估计值只是观察到的（随机）数据的一个函数（即我们总是可以写成θ^ = $\theta\hat{}(X)$），所以它本身就是一个随机变量因此，我们可以定义。

$$\text{bias}(\theta\hat{}) = E(\theta\hat{}) - \theta。$$
$$\text{var}(\theta\hat{}) = E(\theta\hat{} - E(\theta\hat{}))^2。$$

本周的实验也探讨了这些概念，我们鼓励你在完成本题的同时做实验练习，以获得全面的了解。用文字对实验进行简短的总结。

- 偏见：告诉我们估计者的预期值离事实有多远。回顾一下，估计值是我们观察到的数据样本的一个函数。估算器的期望值可以用以下方式来思考：想象一下，我们没有单一的数据样本，而是有无限多的数据样本。我们在每个样本上计算同一个估计值，然后取一个平均值。这就是估计器的期望值。
- 变异性：我们的估计器有多大的可变性。同样，如果我们有无限多的数据样本，我们将能够无限次地计算估计值，并检查所有样本中估计值的变化。

一个好的估计器应该具有低偏差和低方差。

(a) 找到$\mu$的偏差和方差 MLE 其中，$X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, 1)$。

**解决方案。**

我们已经发现，$\hat{\mu}_{MLE} = \bar{X}$. 因此

$$\begin{aligned}\text{bias}(\hat{\mu}_{MLE}) &= \text{bias}(\bar{X}) \\ &= \text{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] - \mu \\ &= \frac{1}{n}\sum_{i=1}^{n} \text{E}(X_i) - \mu \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu - \mu \\ &= \mu - \mu \\ &= 0,\end{aligned}$$

我们说$\bar{X}$是$\mu$的无偏估计值。 接下来，我们有

$$\begin{aligned}\text{var}(\hat{\mu}_{MLE}) &= \text{var}(\bar{X}) \\ &= \text{var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(X_i) \\ &= \frac{1}{n},\end{aligned}$$

在第三个等式中，我们使用了$X_i$'s的独立性，在最后一个等式中，我们使用了$X_i \sim N(\mu, 1)$ 的事实。

(b) 求$\hat{p}_{MLE}$的偏差和方差　　其中$X_1, X_2, ...., X_n \overset{\text{i.i.d.}}{\sim}$ Bernoulli（$p$）。

**解决方案**

我们已经发现，$\hat{p}_{MLE} = \bar{X}$，而且可以很容易地证明$\text{var}(X_i) = p(1-p)$。我们因此得到的是

$$\text{bias}(\hat{p}_{MLE}) = 0。 \quad \text{var}(\hat{p}_{MLE}) = \frac{p(1-p)}{n}$$

。

(c) 平均平方误差（MSE）是一个在统计学和机器学习中被广泛使用的指标。对于真实参数θ的估计器$\hat{\theta}$，我们通过以下方式定义其MSE。

$$\text{MSE}(\hat{\theta}) := \text{E}(\hat{\theta} - \theta)^2。$$

表明MSE服从偏差-方差分解，即我们可以写出

$$\text{MSE}(\hat{\theta}) := \text{bias}(\hat{\theta})^2 + \text{var}(\hat{\theta})。$$

**解决方案。**

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}(\hat{\theta} - \theta)^2 \\
&= \mathrm{E}[\hat{\theta} - \mathrm{E}(\hat{\theta}) + \mathrm{E}(\hat{\theta}) - \theta]^2 \\
&= \mathrm{E}[(\hat{\theta} - \mathrm{E}(\hat{\theta}))^2 + 2(\hat{\theta} - \mathrm{E}(\hat{\theta}))(\mathrm{E}(\hat{\theta}) - \theta) + (\mathrm{E}(\hat{\theta}) - \theta)^2] \\
&= \mathrm{E}(\hat{\theta} - \mathrm{E}(\hat{\theta}))^2 + 2\mathrm{E}[(\hat{\theta} - \mathrm{E}(\hat{\theta}))(\mathrm{E}(\hat{\theta}) - \theta)] + \mathrm{E}[\mathrm{E}(\hat{\theta}) - \theta]^2 \text{。} \\
&= \mathrm{E}(\hat{\theta} - \mathrm{E}(\hat{\theta}))^2 + 2\left(\mathrm{E}(\hat{\theta}) - \mathrm{E}(\hat{\theta})\right)\left(\mathrm{E}(\hat{\theta}) - \theta\right)\text{]。} + [\mathrm{E}(\hat{\theta}) - \theta]^2 \text{。} \\
&= \mathrm{E}(\hat{\theta} - \mathrm{E}(\hat{\theta}))^2 + 0 + [\mathrm{E}(\hat{\theta}) - \theta]^2 \text{。} \\
&= \mathrm{var}(\hat{\theta}) + \mathrm{bias}(\hat{\theta})^2 \text{。}
\end{aligned}$$

**问题3.最小二乘法回归的概率论观点**

在上周的教程中，我们纯粹从优化的角度看待最小二乘法问题。我们指定了我们想要拟合的模型，即。

$$\hat{y} = w^T x$$

以及损失函数（MSE），并简单地找到使损失最小的权重向量$w$。我们证明，当使用MSE时，可能的最佳权重向量是由以下公式给出的

$$\hat{w} = (X^T X)^{-1} X^T y \text{。}$$

在这个问题上，我们将探讨一个不同的观点，我们可以称之为统计学观点。统计学观点的核心是数据生成过程（DGP），它假设有一些真正的基础函数生成数据，我们称之为$f$，但我们只能获得$f$的噪声观测。也就是说，我们观察到

$$y = f(x) + E, \qquad E\text{是一些随机噪声。}$$

例如，假设你的$y$代表肯辛顿的每日温度。任何温度计--即使是最昂贵的--都容易产生测量误差，因此我们实际观察到的是真实的温度（$f(x)$）加上一些随机噪声$E$。最常见的是，我们将假设噪声为正态分布，均值为零，方差为$\sigma^2$。现在，考虑$f(x)$是线性的（强）假设。

这意味着有一些真$\beta^*$，使$f(x) = x\beta^{T*}$。因此，我们有：

$$y = x\beta^{T*} + E \text{。} \qquad E \sim N(0, \sigma^2) \text{。}$$

因此。

$$y|x \sim N(x\beta^{T*}, \sigma^2) \text{。}$$

这说明我们的反应（以知道特征值$x$为条件）遵循正态分布，均值$x\beta^{T*}$，方差$\sigma^2$。因此，我们可以把我们的数据看作是来自这个分布的随机观察样本，这反过来又使我们能够估计未知的参数
通过最大似然法，就像我们在前面的问题中做的那样。

(a) 给你一个数据集$D = (x_1, y_1), \ldots, (x_n, y_n)$，对于一些未知的$\beta^*$和$\varepsilon_i \sim N(0, \sigma^2)$，其中所有的$\varepsilon_i$'s是相互独立的。写下这个问题的对数似然以及最大似然估计目标，并求解MLE估计器$\hat{\beta}_{MLE}$。

**解决方案**

在这个假设下，我们有：$y_i | x_i \sim N(x_i^T \beta^*), i = 1, \ldots, n$，或者我们可以把它写成矩阵记号为。

$$y|X \sim N(X\beta^*, \sigma^2 I)。$$

我们可以计算出数据的可能性为

$$L(\beta) = P(y|X, \beta)$$

。

正确解释这个概率是很重要的：它是在我们有特征$X$的情况下看到反应$Y$的概率，并且我们假设基础矢量是$\beta$。$L(\beta)$将为我们提供不同的$\beta$选择的可能性值，我们要选择最好的$\beta$，即使$L(\beta)$最大化的那个。现在，让我们详细写出什么是$\log L(\beta)$。

$$\begin{aligned}
\log L(\beta) &= \log P(y|X, \beta) \\
&= \text{对数} \prod_{i=1}^{n} P(y_i | x_i, \beta) \\
&= \sum_{i=1}^{n} \log P(y_i | x_i, \beta) \\
&= \sum_{i=1}^{n} \text{日志} \frac{1}{\sqrt{2\pi\sigma^2}} \text{支出} - \frac{(y_i - x_i^T \beta)^2}{2\sigma^2} \\
&= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 \\
&= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|y - X\beta\|_2^2。
\end{aligned}$$

因此，我们得到$\beta$的MLE估计值是

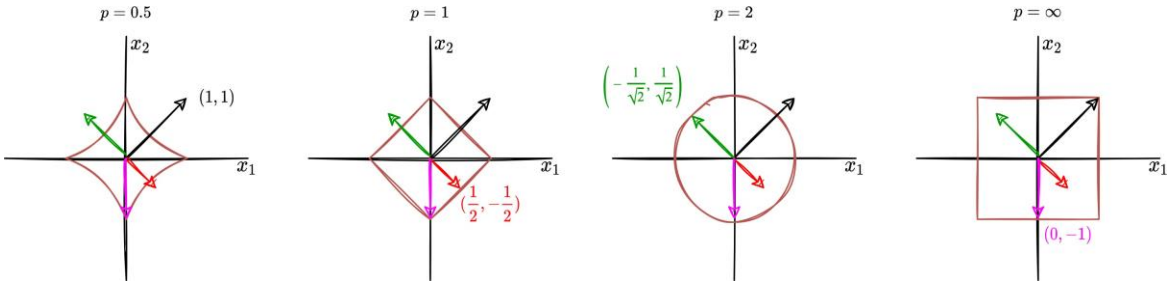
$$\begin{aligned}
\hat{\beta}_{MLE} &= \arg\max_{\beta} \log L(\beta) \\
&= \arg\max_{\beta} -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|y - X\beta\|_2^2 \\
&= \arg\min_{\beta} \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \|y - X\beta\|_2^2 \\
&= \arg\min_{\beta} \|y - X\beta\|_2^2 \\
&\overset{I}{=} (X^T X)^{-1} X^T y \\
&= \hat{\beta}_{LS}。
\end{aligned}$$

第二个等式成立，因为最大化一个目标与最小化同一目标的负数是一样的，第三个等式成立，因为最小化器不受第一个项的影响。请注意，我们已经表明，在这种情况下，MLE与最小二乘法估计器完全相同，这在一般情况下不是真的。例如，如果我们使用不同的损失函数（不是MSE），或者对噪声的不同假设，或者不同的分布（除了正态分布），那么我们就不会得到LS和MLE之间的等价关系。我们可以认为这是做最小二乘法的一个概率论的理由。

**问题4.几何学解释**

在这个问题上，我们将探讨最小二乘法（LS）、山脊和LASSO回归模型的一些几何直觉。

(a) 考虑下图，它表示在不同的*P-下*单位球的等高线图。
准则（如果你不熟悉co√ntour，请看实验0）。解释发生了什么，并评论
四个向量（（1，1），（1/2，-1/2），（-1/2，1/2），（0，-1））， 在图中表示。
此外，第一幅图（$p$=0.5）和其他图之间的区别是什么？




**解决方案**

首先，$x \in \mathbb{R}^2$ 中的一个矢量由两个坐标$x = (x_1, x_2)$表示。接下来，回顾一下，a二维向量的*p-norm*定义为

$$IxI_p = (|x_1|^p + |x_2|^p)^{1/p}, \qquad p \geq 1.$$

请注意，这排除了$p$=0.5被称为 "规范 "的可能性。规范是一个长度的概念，因此任何函数$g$都必须满足一些条件，我们才允许将其称为规范。这些条件是。

(i) 三角形不等式：$g(x+y) \leq g(x)+g(y)$，对于任何两个向量$x$，$y$，这只是说两个向量的长度之和小于长度之和。

(ii) 绝对同质性：对于任何矢量$x$和标量$c$，$g(cx)=|c|g(x)$，这说明如果你将一个向量乘以一个常数，新向量的长度就是旧向量的长度，并以$c$为尺度。

(iii) 正确定性：$g(0) = 0$。

你可以检查一下，对于上面定义的*p-norms*，这三个总是被满足的。然而，当$p$=0.5时，我们可以找到一个三角形不等式的反例。考虑$x$=（0，9）和$y$=（4，0），那么

$$IxI_{0.5} = (\sqrt{0} + \sqrt{9})^2 = 3^2 = 9$$

$$IyI_{0.5} = (\sqrt{4})^2 = 4$$

$$IxI_{0.5} + IyI_{0.5} = 13$$

$$Ix + yI_{0.5} = I\ (4,\ 9)\ \sqrt{4}\ \sqrt{9}) = 25$$

所以我们看到 $I_{0.5} = (\quad + \quad^2$

，

$$Ix + yI_{0.5} > IxI_{0.5} + IyI_{0.5}$$

所以函数$g(x)=x_{0.5}$ 不满足三角形不等式，也不是一个规范，尽管它经常被称为*伪规范*。现在，让我们试着弄清楚这些图中发生了什么。在每个图中，我们都有对应于$p$=0.5、1、2、∞的p-*norm*选择的*单位圆*，这意味着每个图中用棕色勾勒的形状代表了满足$IxI$的向量$x_p = 1$。这通常被称为*单位圆*，通常被写为

$$\{x : IxI_p = 1\}。$$

这个符号应作如下解释：所有对象$x$的集合，使$x$满足$x_p = 1$的要求。这告诉我们的是，根据准则的选择，一个向量的长度可以是不同的。让我们看一个具体的例子。在黑色中我们有一个向量（1，1），注意

$$i(1,\ 1)i_{0.5} = (\sqrt{1} + \sqrt{1})^2 = 2^2 = 4$$

$$i(1,\ 1)i_1 = |1| + |1| = 2$$

$$I(1,\ 1)I_2 = \sqrt{|1|^2 + |1|^2} = \sqrt{2}$$

$$I(1,\ 1)I_\infty = \max\{|1|,\ |1|\} = 1。$$

所以这个向量只在-norm圆上∞这意味着我们只在使用$p$=-norm时测量其长度为1。在所有其他情况下，(1, 1)的规范都大于1。这给了我们一个关于*p-规范的*一般规则。

$$IxI_p \geq IxI_q \quad \text{只要} p < q。$$

请注意，从几何学的角度来看，只要$p$为*1*，单位圆所包围的区域（称为单位球）就是一个凸集，这在图2-4中可以看到。当$p$<1时，凸性不成立，在第一个图中可以看到。关于凸集的更多信息，请参见维基页面。

(b) 我们之前看到，山脊回归的目标是通过以下方式定义的。

$$\hat{\beta}^{ridge} = \arg\min_{\beta} Iy - X\beta I_2^2 + \lambda I\beta I2。$$

另一种定义山脊目标的（等价）方式是通过约束性优化。

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \|y - X\beta\|_2^2 \text{受制于} \|\beta\|_2 \leq k。$$

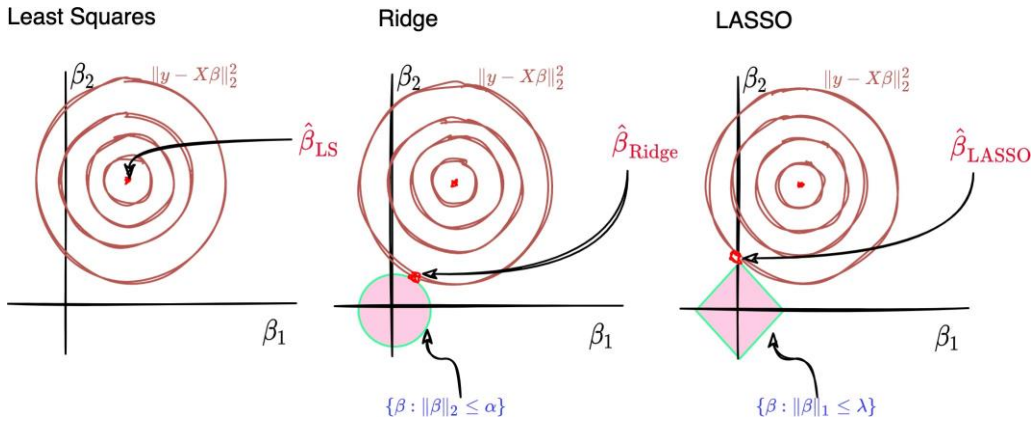
这说明我们想找到最小化平方损失的β，但该解决方案必须同时属于半径为$k$的2-norm球。受约束的优化声明给我们提供了一个很好的脊柱解决方案的几何解释，我们现在要探讨一下。在这样做之前，我们还要注意到LASSO有一个无约束的版本。

$$\hat{\beta}^{LASSO} = \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 。$$

还有一个受限的版本

$$\hat{\beta}^{LASSO} = \arg\min_{\beta} \|y - X\beta\|_2^2 \quad \text{受制于} \|\beta\|_1 \leq k。$$

其中，Ridge和LASSO的λ、$k$通常是不同的。这些目标几乎是相同的，它们只因用于惩罚/约束项的准则的选择而不同。这实际上导致了实践中的巨大和非常重要的差异。现在，考虑到这一点，解释一下下面的图。




明确讨论Ridge和LASSO解决方案的差异。

**解决方案。**

在这些图中，我们用图形描述了二维的最小二乘法、山脊法和LASSO优化，即β=（$\beta_1$，$\beta_2$）。在第一幅图中，我们用图形描述了最小二乘法的优化，记得是这样的。

$$\hat{\beta}_{\text{辽宁省}} = \arg\min_{\beta} \|y - X\beta\|_2^2$$

棕色圆圈代表平方误差函数$\|y - X\beta\|_2^2$的轮廓。同样，如果你不熟悉这个概念，请看实验0。我们知道轮廓线的小圆圈

代表函数的较小值。因此，最小二乘法的解决方案只是点
直接在中间，这给了我们$\|y - X\beta\|_2^2$的最小可能值。
在第二幅图中，我们现在正在做脊回归。从约束优化的角度来看，我们仍然希望最小化平方误差，所以我们仍然看到平方误差的轮廓。然而，我们还需要遵守一个约束条件，即我们的解决方案的规范属于半径为$\alpha$的规范球，也就是说，我们的解决方案必须在粉红色的阴影区域（这被称为可行解决方案的集合）。那么，山脊问题的解决方案就是以下这一点
是在可行的解决区域内，并使$\|y - X\beta\|_2^2$项尽可能的小。图形-
从这个意义上说，这只是用红色标出的交点。最后，一个相同的解释是
对LASSO问题来说也是如此，只是现在我们使用的是1-norm而不是2-norm。众所周知，
LASSO可以找到*稀疏的*解决方案。如果一个向量只有几个元素是非零的，那么它就是稀疏的。例如，我们会说这个向量

$$[0, 0, 0, 1, 0, 0, 0.2, 0]$$

是稀疏的，而

$$[0, 2, 2, 1, 0, 3, 0.2, 1]$$

稀疏的。原因是球在(1, 0), (0, 1), (-1, 0), (0, -1)这几个点上*有角*或*尖刺*。
是不稀疏的，请注意1-norm球的形状给我们的解决方案位于其中
并且在这些点之间是平坦的，所以平方误差很可能与这些点之一相约束集相交。这很重要，因为这些点对应于将其中一个权重设置为零。2-norm球的边是圆的，所以在山脊的情况下，这种情况不太可能发生。

(c) 据说LASSO会*诱发稀疏性*。这是什么意思？为什么希望有一个稀疏的解决方案？

**解决方案。**

我们在前面的问题中已经看到了为什么LASSO会引起稀疏性，但是我们还没有讨论为什么我们可能想要一个稀疏的解决方案。获得一个稀疏的解决方案在科学应用中往往是很重要的：如果我们试图理解某样东西是如何工作的，那么拥有一个稀疏的模型意味着我们可以将反应归因于少数特征，而不是拥有一个具有大量特征的模型，这使得我们无法做任何合理的解释。从另一个角度来看，如果收集数据是昂贵的，那么把效果缩小到几个好的特征就可以使你的模型操作起来更加便宜。