10 Gaussian Process 高斯过程

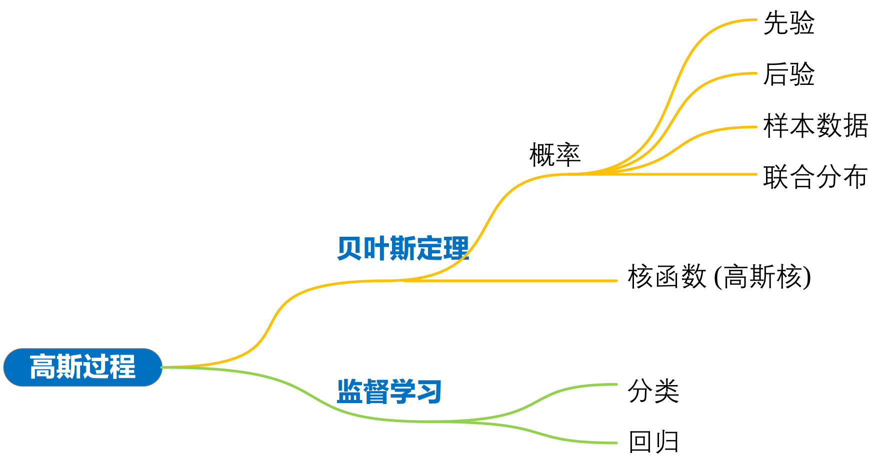
多元高斯分布的条件概率，协方差矩阵为核函数

生命就像一个永恒的春天，穿着崭新而绚丽的衣服站在我面前。

Life stands before me like an eternal spring with new and brilliant clothes.

—— 卡尔·弗里德里希·高斯 (Carl Friedrich Gauss) | 德国数学家、物理学家、天文学家 | 1777 ~ 1855

- ◀ `sklearn.gaussian_process.GaussianProcessRegressor()` 高斯过程回归函数
- ◀ `sklearn.gaussian_process.kernels.RBF()` 高斯过程高斯核函数
- ◀ `sklearn.gaussian_process.GaussianProcessClassifier()` 高斯过程分类函数



高斯过程
贝叶斯定理
概率
先验
后验
样本数据
联合分布
核函数 (高斯核)
监督学习
分类
回归

10.1 高斯过程原理

高斯过程 (Gaussian Process, GP) 既可以用来回归，又可以用来分类。高斯过程是一种概率模型，用于建模连续函数或实数值变量的概率分布。在高斯过程中，任意一组数据点都可以被视为多元高斯分布的样本，该分布的均值和协方差矩阵由先验信息和数据点间的相似度计算而得。通过高斯过程，可以对函数进行预测并对其不确定性进行量化，这使得其在机器学习、优化和贝叶斯推断等领域中被广泛应用。

在使用高斯过程进行预测时，通常使用条件高斯分布来表示先验和后验分布。通过先验分布和数据点的观测，可以计算后验分布，并通过该分布来预测新数据点的值。在高斯过程中，协方差函数或核函数起着重要的作用，它定义了数据点间的相似性，不同的核函数也适用于不同的应用场景。一些常见的核函数包括线性核、多项式核、高斯核、拉普拉斯核等。

高斯过程具有许多优点，如对噪声和异常值具有鲁棒性、能够对预测结果的不确定性进行量化、对于小样本也能够进行有效的预测等。然而，高斯过程的计算复杂度较高，通常需要通过一些技巧来提高其效率。

想要理解高斯过程，必须对多元高斯分布、条件概率、协方差矩阵、贝叶斯推断等数学工具烂熟于心。

《统计至简》第 11 章讲解多元高斯分布，第 12 章将讲解条件高斯分布，第 13 章介绍协方差矩阵，第 20、21 两章介绍贝叶斯推断，建议大家回顾。

先验

$\boldsymbol{x}_2$ 为一系列需要预测的点，$\boldsymbol{y}_2 = \mathrm{GP}(\boldsymbol{x}_2)$ 对应高斯过程预测结果。

高斯过程的先验为：

$$\boldsymbol{y}_2 \sim N\left(\boldsymbol{\mu}_2, \boldsymbol{K}_{22}\right) \tag{1}$$

其中，$\boldsymbol{\mu}_2$ 为高斯过程的均值，$\boldsymbol{K}_{22}$ 为协方差矩阵。之所以写成 $\boldsymbol{K}_{22}$ 这种形式，是因为高斯过程的协方差矩阵通过核函数定义。

本章主要利用高斯核 (Gaussian kernel)，也叫径向基核 RBF。

在 Scikit-learn 中，高斯核的定义为：

$$\kappa\left(x_i, x_j\right) = \exp\left(-\frac{\left(x_i - x_j\right)^2}{2l^2}\right) \tag{2}$$

当输入为多元的情况，上式分子为向量差的欧氏距离平方，即 $\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2^2$。图 1 所示为 $l = 1$ 时，先验协方差矩阵的热图。

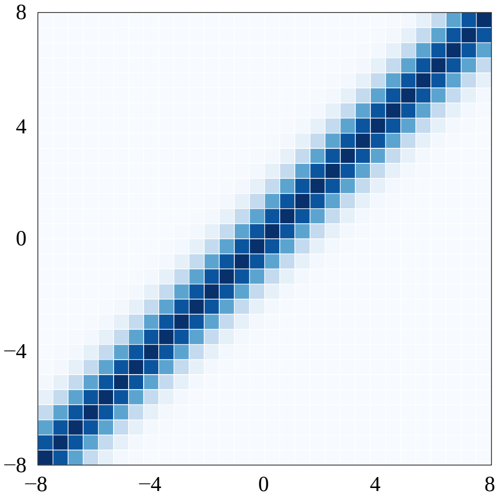


图 1. 高斯过程的先验协方差矩阵，高斯核

图 2 所示为参数 l 对先验协方差矩阵的影响。

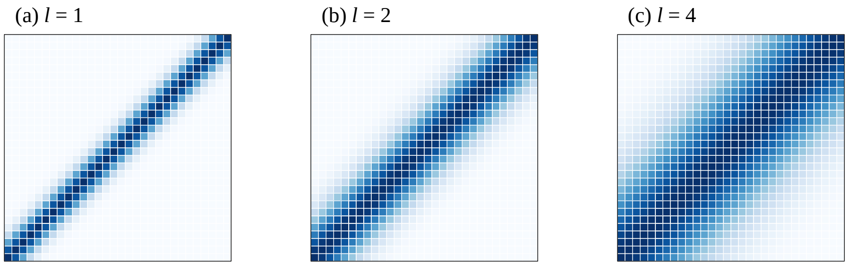


图 2. 先验协方差矩阵随着参数 l 变化，高斯核

很多其他文献上中高斯核定义为：

$$\kappa\left(x_i, x_j\right) = \sigma^2 \exp\left(-\frac{\left(x_i - x_j\right)^2}{2l^2}\right) \tag{3}$$

本章采用 Scikit-learn 中高斯核的定义。

本书前文介绍过，核函数本身实际上可以看做一个格拉姆矩阵。协方差矩阵也是格拉姆矩阵的一种特例。

⚠ 注意，高斯过程可以选择的核函数有很多。此外，不同核函数还可以叠加组合。

图 3 所示每一条代表一个根据当前均值、协方差的函数采样。打个比方，在没有引入数据之前，图 3 的曲线可以看成是一捆没有扎紧的丝带，随着微风飘动。

图 3 中的红线为高斯过程的均值，本章假设均值为 0。

《统计至简》第 9 章介绍过“68-95-99.7 法则”，图 3 中±2σ 对应约 95%。即约 95%样本位于距平均值 1 正负 2 个标准差之内。

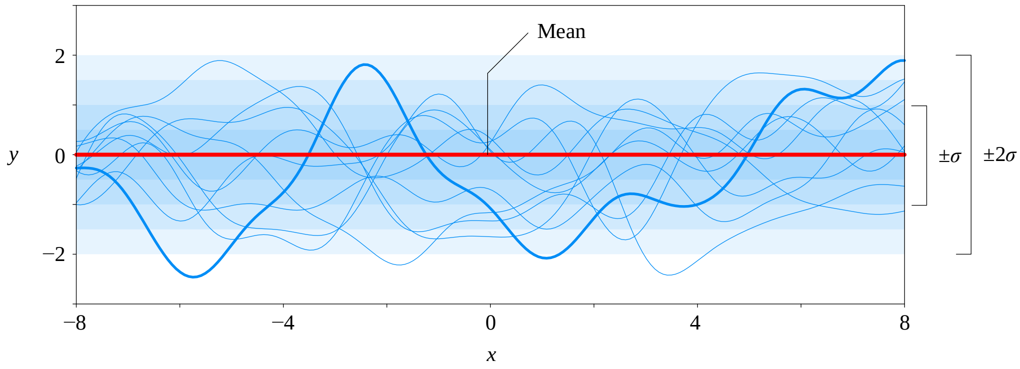


图 3. 高斯过程的采样函数，高斯核

样本

观测到的样本数据为 (x_1, y_1)。图 4 给出两个样本点，它们相当于扎紧丝带的两个节点。

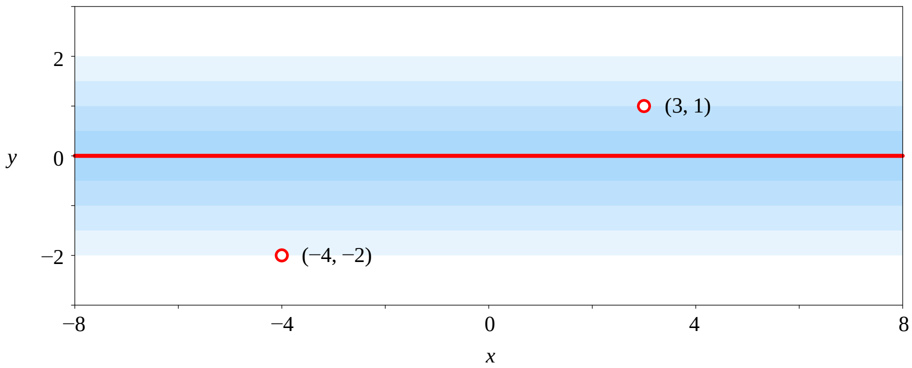


图 4. 给定样本数据 {(−4, −2), (3, 1)}

联合分布

假设样本数据 y_1 和预测值 y_2 服从联合高斯分布：

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}_{11} & \boldsymbol{K}_{12} \\ \boldsymbol{K}_{21} & \boldsymbol{K}_{22} \end{bmatrix} \right) \tag{4}$$

简单来说，高斯过程对应的分布可以看成是无限多个随机变量的联合分布。图 5 中的协方差矩阵来自 $[x_1, x_2]$ 的核函数。

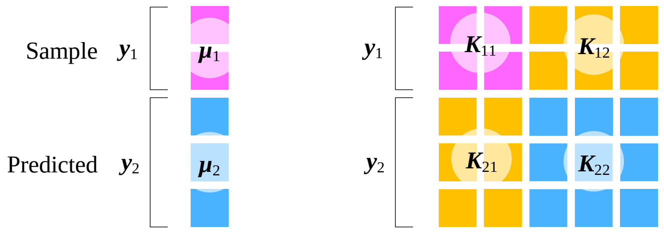


图 5. 样本数据 $\boldsymbol{y}_1$ 和预测值 $\boldsymbol{y}_2$ 服从联合高斯分布

后验

后验分布为：

$$f\left(\boldsymbol{y}_2 \mid \boldsymbol{y}_1\right) \sim N\left(\underbrace{\boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1}\left(\boldsymbol{y}_1-\boldsymbol{\mu}_1\right)+\boldsymbol{\mu}_2}_{\text{Expectation}}, \underbrace{\boldsymbol{K}_{22}-\boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1}\boldsymbol{K}_{12}}_{\text{Covariance matrix}}\right) \tag{5}$$

图 6 所示为引入样本数据 {(−4, −2), (3, 1)} 后，后验协方差矩阵的热图。

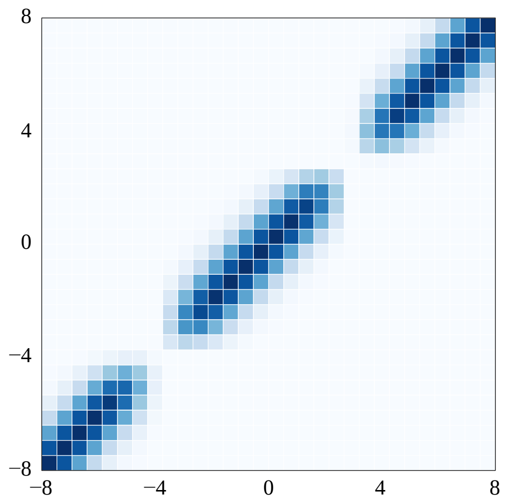


图 6. 高斯过程的后验协方差矩阵，高斯核

如图 7 所示，样本点位置丝带被锁紧，而其余部分丝带仍然舞动。

图 7 中红色曲线对应后验分布的均值：

$$\boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1}\left(\boldsymbol{y}_1-\boldsymbol{\mu}_1\right)+\boldsymbol{\mu}_2 \tag{6}$$

图 7 中带宽对应一系列标准差，它们的方差为：

$$\operatorname{diag}\left(\boldsymbol{K}_{22}-\boldsymbol{K}_{21}\boldsymbol{K}_{11}^{-1}\boldsymbol{K}_{12}\right) \tag{7}$$

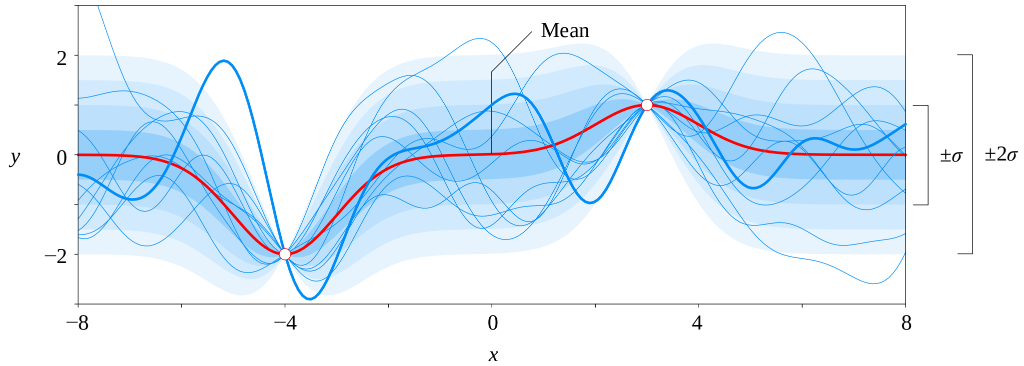


图 7. 高斯过程后验分布的采样函数，高斯核

其他几组情况

图 8 所示为随着样本不断增加，后验概率分布的协方差矩阵热图、高斯过程后验分布采样函数变化情况。

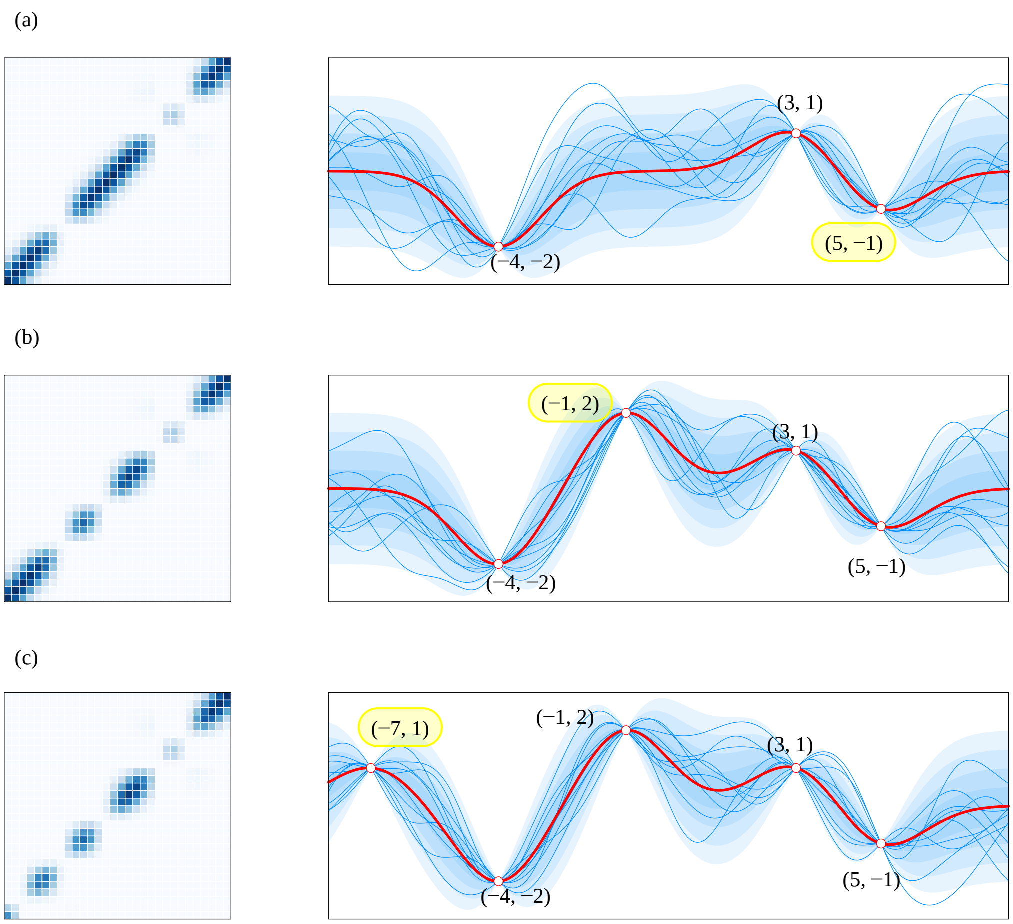


图 8. 样本数据不断增加

10.2 解决回归问题

Scikit-learn 解决回归问题的函数为 sklearn.gaussian_process.GaussianProcessRegressor()。

图 9 (a) 中蓝色曲线为真实曲线，对应函数为 $f(x) = x\sin(x)$。图 9 (a) 中红色点为样本点，蓝色曲线为高斯过程回归曲线，浅蓝色宽带为 95%置信区间。

图 9 (b) 所示为在样本点上加上噪音后的回归结果。

这个例子中使用的是高斯核函数，对应 sklearn.gaussian_process.kernels.RBF()。请大家条件参数，观察对回归结果的影响。此外，请大家试着使用其他核函数，并比较回归曲线。

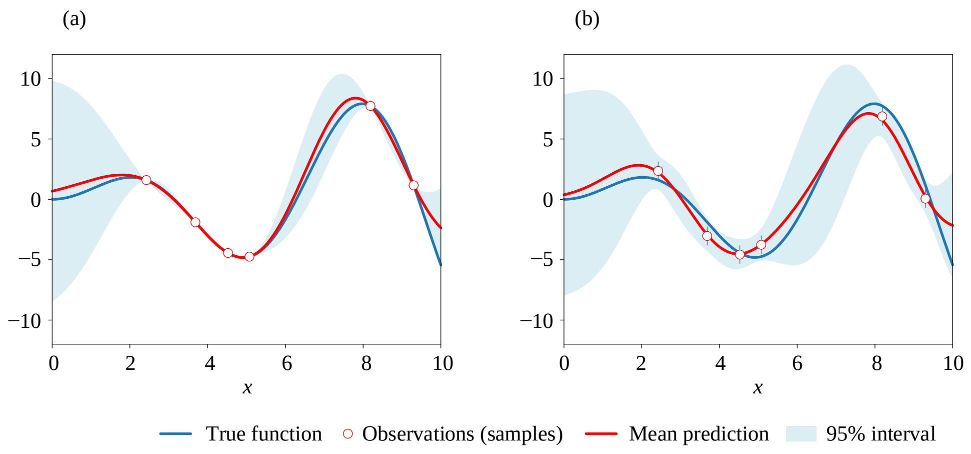


图 9. 使用高斯过程完成回归

本节高斯过程回归参考如下 Scikit-learn 示例，请大家自行学习：

https://scikit-learn.org/stable/auto_examples/gaussian_process/plot_gpr_noisy_targets.html

10.3 解决分类问题

sklearn.gaussian_process.GaussianProcessClassifier() 是 Scikit-learn 中专门用来解决高斯过程分类的函数。本例利用这函数根据花萼长度、花萼宽度分类鸢尾花。图 10 所示为采用高斯过程分类得到的决策边界。图 11 所示为三个后验曲面三维曲面和平面等高线。

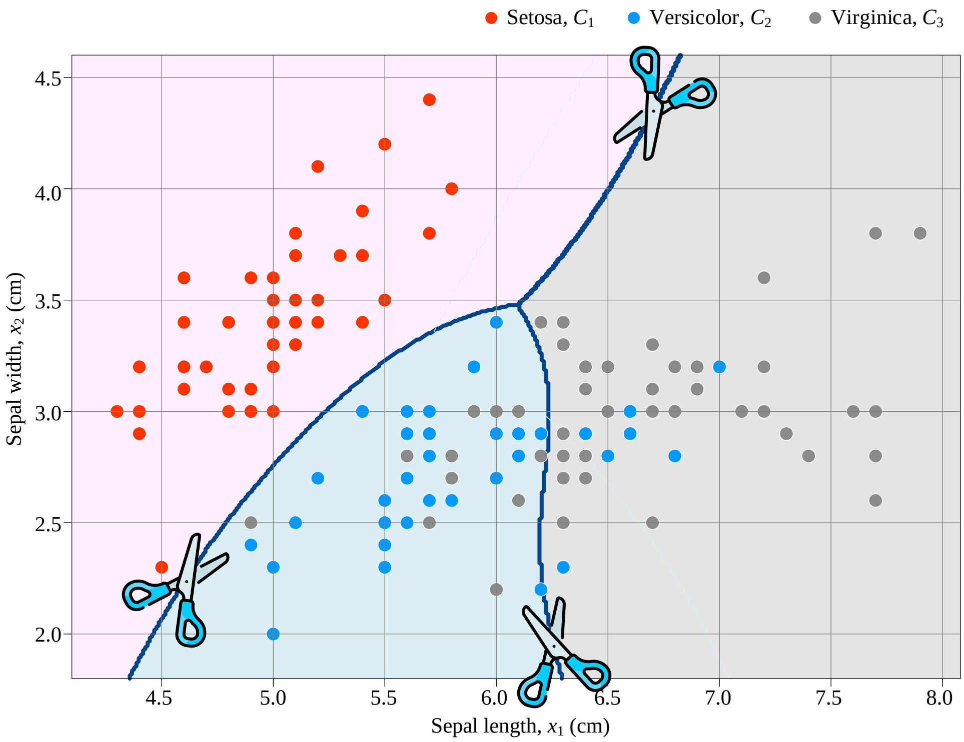


图 10. 使用高斯过程完成鸢尾花分类

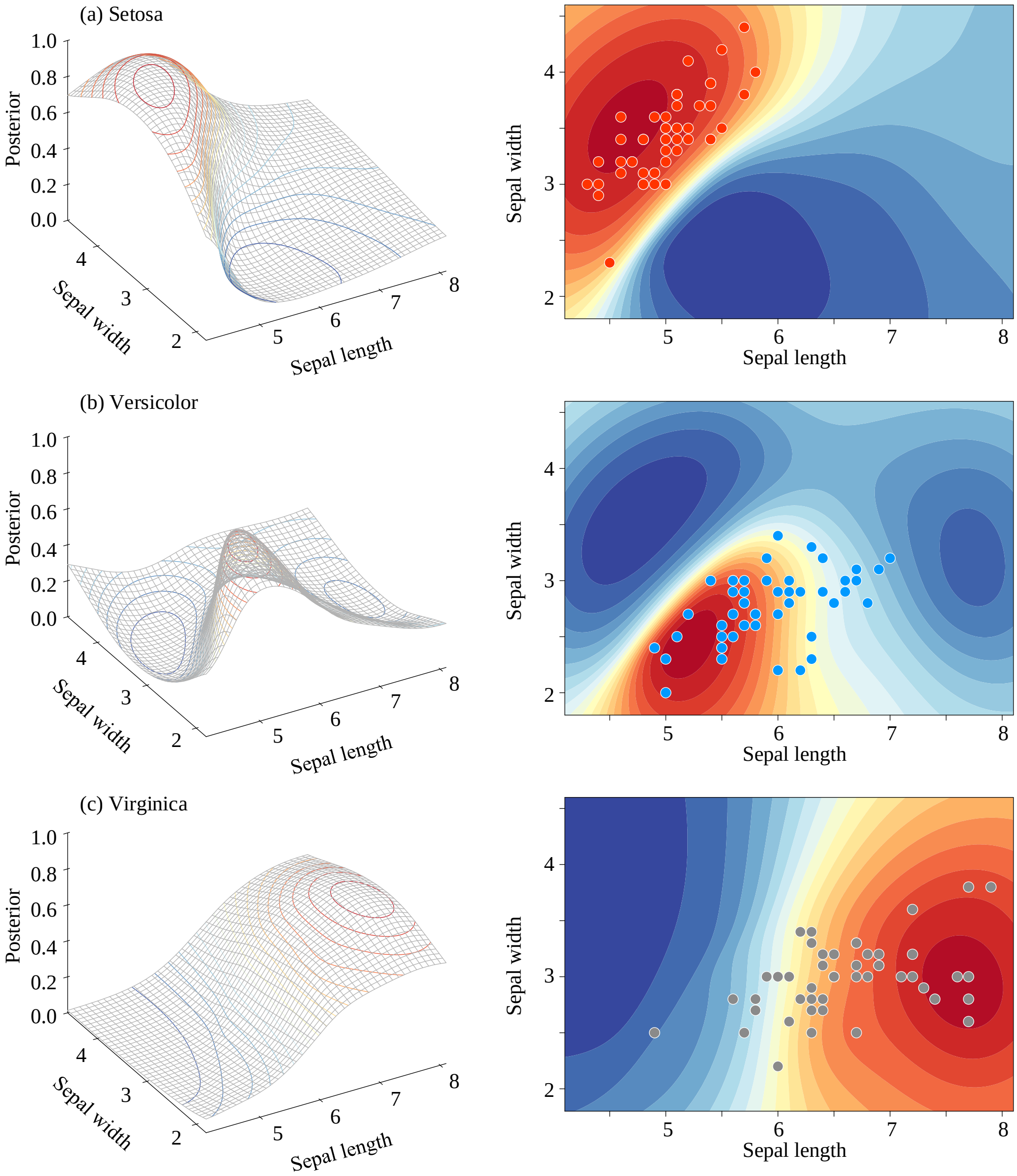


图 11. 后验概率曲面

Bk7_Ch10_01.py 利用高斯过程完成鸢尾花分类，并绘制图 10 和图 11。

高斯过程是一种基于概率论的非参数模型，可以用于建模连续函数或实数值变量的概率分布。在高斯过程中，先验分布通过核函数和一些超参数来定义，数据点的观测可以通过似然函数与先验分布相结合，计算后验分布。高斯过程中的核函数通常定义了数据点之间的相似性，超参数可以通过最大化似然或最大化边缘似然来优化。

在回归问题中，高斯过程可以用于预测连续变量的值，并估计预测值的不确定性。在分类问题中，高斯过程分类器可以将先验分布定义为高斯分布，通过后验分布进行分类预测。高斯过程的优点包括模型具有灵活性、能够对预测结果的不确定性进行量化、对噪声和异常值具有鲁棒性等。但高斯过程的计算复杂度较高，需要进行计算优化和近似方法。

大家想要深入学习高斯过程，请参考如下开源图书 *Gaussian Processes for Machine Learning*：

https://gaussianprocess.org/gpml/

这篇博士论文中专门介绍了不同核函数的叠加：

https://www.cs.toronto.edu/~duvenaud/thesis.pdf

作者认为下面这篇文章解释高斯过程做的交互设计最佳，这篇文章给了作者很多可视化方面的启发：

https://distill.pub/2019/visual-exploration-gaussian-processes/