Seminar for Linear Regression Part 2

注意事项：

本次研讨会极大概率出现：

- 逆天讲解
- 抽象图例
- 可爱小猫
- ~~伪~~随机提问

接下来就让我们一起研究贝叶斯线性回归罢！😋

3.3 贝叶斯线性回归 (Bayesian Linear Regression)

贝叶斯线性回归（Bayesian Linear Regression）是一种基于贝叶斯统计方法的回归分析技术，一般用于建立和预测变量之间的线性关系。它结合了贝叶斯概率理论和线性回归模型，允许在建模过程中引入不确定性，并提供了对参数和预测的不确定性估计。

STAT2000小课堂

~~(STAT1021小课堂)（绝望）~~

贝叶斯定理（Bayes' theorem）：

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

其中：

- $P(A|B)$ 是给定观测到事件 B 后，事件 A 发生的后验概率。
- $P(B|A)$ 是在事件 A 发生的情况下，事件 B 发生的条件概率。
- $P(A)$ 是事件 A 的先验概率，即在考虑任何新信息之前，事件 A 发生的概率。
- $P(B)$ 是事件 B 的边际概率，即在考虑事件 A 和其他因素的情况下，事件 B 发生的概率。

传统的线性回归方法通常仅给出点估计的参数值，而贝叶斯线性回归考虑了参数的概率分布。

先验概率（Prior Probability）： 先验概率是在考虑任何新信息之前，我们对一个事件或假设的概率。它基于我们的经验、知识或主观判断。先验概率可以被看作是"初始"概率，因为它是在考虑新数据之前的概率。

后验概率（Posterior Probability）： 后验概率是在考虑了新信息或数据后，更新过的事件或假设的概率。通过应用贝叶斯定理，我们可以将先验概率和新数据结合起来，得出后验概率。后验概率提供了一个更为准确和实际的概率度量，考虑了我们在观测数据后对事件的新认识。

我们从之前的小章节可以看出，极大似然方法需要根据数据集大小来确定基函数的数量,从而控制模型复杂度。加入正则项后,可以通过调节正则化系数来控制模型复杂度,但基函数的数量和形式选择也很重要。

但是，极大似然法容易过拟合现象,我们可以使用贝叶斯方法来规避这一问题，同时我们也可以使用贝叶斯方法来解决特定问题的数据集大小的模型复杂度问题。

~~那么古尔丹，代价是什么呢？~~

3.3.1 参数分布(Parameter distribution)

1. 我们以介绍模型参数 w 的先验概率分布来开始我们的线性回归的贝叶斯方法讨论。在这个阶段，我们把噪声精度参数 β 当做已知常数。首先，注意到，公式（3.10）定义的似然函数 $p(t|w)$ 是 w 的二次函数的指数形式。于是对应的共轭先验是形式：

$$p(w) = \mathcal{N}(w|m_0, S_0) \quad （3.48）$$

均值和方差分别为 m_0 和 S_0 的高斯分布。最大似然函数，形式为：

$$p(\mathbf{t}|\mathbf{X}, w, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|w^T\phi(x_n), \beta^{-1}) \quad （3.10）$$

计算后验分布，类比直接写。由于在推导公式（2.116）时，已经进行了必要的工作，所以我们能够直接写出后验概率分布的形式：

$$p(w|\mathbf{t}) = \mathcal{N}(w|m_N, S_N) \quad （3.49）$$

其中：

$$m_N = S_N(S_0^{-1}m_0 + \beta\Phi^T t) \quad （3.50）$$

$$S_N^{-1} = S_0^{-1} + \beta\Phi^T\Phi \quad （3.51）$$

注意，由于后验分布是高斯的，它的众数（概率分布最高点）与均值正好相同。因此最大后验权向量由 $w_{\text{MAP}} = m_N$ 给出。如果考虑一个无限宽的先验 $S_0 = \alpha^{-1}I$，其中 $\alpha \to 0$，那么后验分布的均值 m_N 就退化成了由式（3.15）给出的最大似然值 w_{ML}（无信息先验）。类似地，如果 $N = 0$，那么后验分布就被还原成先验分布。此外，如果数据点是顺序到达的，那么任何一个阶段的后验分布都可以看成后续数据点的先验分布。此时新的后验分布再次由式（3.49）给出。

简化，有一个特定的形式高斯先验：一个只由一个精度参数 α 控制的零均值各向同性高斯分布：

$$p(w|\alpha) = \mathcal{N}(w|0, \alpha^{-1}I) \quad （3.52）$$

且对应的关于 w 的后验分布由式（3.49）给出，其中：

$$m_N = \beta S_N\Phi^T t \quad （3.53）$$

$$S_N^{-1} = \alpha I + \beta\Phi^T\Phi \quad （3.54）$$

后验分布的对数是由对数似然与先验的对数的和给出的一个关于 (w) 的函数，形式为：

$$\ln p(w|\mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - w^T\phi(x_n)\}^2 - \frac{\alpha}{2}w^T w + \text{const} \quad （3.55）$$

小例子：简化模型方便理解：考虑单输入变量 x ，单目标变量 t 和线性模型形式$y(x, w) = w_0 + w_1 x$。

我们从带有参数$a_0 = -0.3, a_1 = 0.5$的函数$f(x, a) = a_0 + a_1 x$ 中生成数据。生成方法是：首先从均匀分布 $U(x|-1, 1)$ 中选择 x_n 的值，再计算 $f(x_n, a)$ ，最后加上一个标准差为0.2的高斯噪声，得到目标变量 t_n。我们的目标是从这样的数据中恢复a_0, a_1的值,并探索模型对数据集规模的依赖关系。

这里我们假设噪声方差是已知的，所以我们把精度参数设置为它的真实值$\beta = (1/0.2)^2 = 25$。同样的，我们把α固定为2.0。然后就有图3.7可以看：

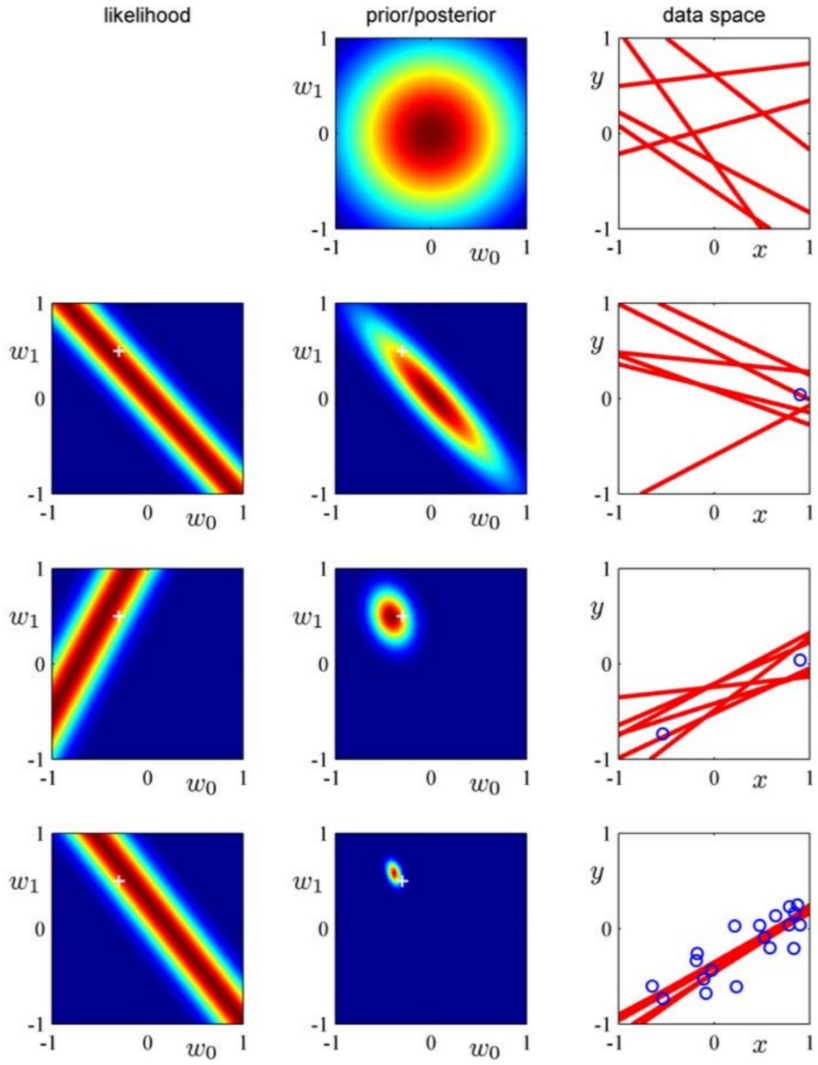


Figure 3.7 Illustration of sequential Bayesian learning for a simple linear model of the form $y(x, \mathbf{w}) = w_0 + w_1 x$. A detailed description of this figure is given in the text.

图3.7展示了当数据集的规模增加时贝叶斯学习的结果，还展示了贝叶斯学习的顺序本质，即当新数据点被观测到的时候，当前的后验分布变成了先验分布。

后日谈：

总结：参数分布是贝叶斯方法中的关键概念，用主要于描述模型参数的不确定性，并通过将先验信息与观测数据相结合，更新参数的后验分布。

小小的问题：

还记得我们之前的小猫年龄和小猫猫粮消耗的例子吗？如你所

见，小猫猛吃，把猫粮吃光光了。现在小黄打算使用贝叶斯方法来预测小猫第十一个月的消耗量，请你指出小黄的行为是不是合理？（小提示：还是只有十个数据）(~~Tsai, P. (2023). 终于最后七箱..... Moments.~~)

~~答案~~我不好说：

Q：为什么小黄能够使用贝叶斯方法来预测呢？

A：在我们这种小样本情况$N = 10$下，数据量有限，很难仅仅从数据本身获得准确的模型参数估计。贝叶斯方法能够利用先验信息，将先验分布与观测数据相结合，从而提供关于参数的更准确估计。这种先验信息有助于填补数据的不足，从而减少对数据本身的依赖。

那这里的先验信息是什么？
~~小猫的营养需求、Peiyu的经验、小猫的生物学特征。~~

3.3.2 预测分布 (Predictive distribution)

在实际应用中，我们对新的 (x) 的值预测出 (t) 比 (w) 的值本身更感兴趣。这需要我们估计出定义为：

$$p(t|\mathbf{t},\alpha,\beta) = \int p(t|w,\beta)p(w|\mathbf{t},\alpha,\beta)dw \quad (3.57)$$

的预测分布。其中 t 是训练集中的目标向量，我们可以得到的预测分布的形式为：

$$p(t|x,\mathbf{t},\alpha,\beta) = \mathcal{N}(t|m_N^T\phi(x),\sigma_N^2(x)) \quad (3.58)$$

其中预测分布的方差 $\sigma_N^2(x)$ 是：

$$\sigma_N^2(x) = \frac{1}{\beta} + \phi(x)^T S_N \phi(x) \quad (3.59)$$

第一项表示数据上的噪声，第二项表示与参数 (w) 关联的不确定性。

由于噪声的处理与 w 的分布是相互独立的高斯，它们的方差是可加的。注意，当额外的数据点被观测到的时候，后验概率分布会变窄。从而可以证明 $\sigma^2_{N+1}(x) \leq \sigma^2_N(x)$（Qazaz et al., 1997）。在取极限 $N \to \infty$ 的情况下，公式（3.59）的第二项趋向于0，这时预测分布的方差只来自于由参数 β 控制的可加噪声。

为了阐释贝叶斯线性回归模型的预测分布，我们可以看看1.1节中人工生成的正弦数据集~~（就是你Oscar几百年前讲chap1的时候掏出来的正弦函数图。后面围绕这个图还讲了不少内容，记不得的小窗出去谢罪）~~。在图3.8中，我们展示了在由高斯基函数线性组合成的模型下，不同数据集大小和对应的后验分布的关系。其中，绿色曲线对应生成数据（加上高斯噪声）的函数 $\sin(2\pi x)$。大小为 $N = 1, N = 2, N = 4, N = 25$ 的数据集以蓝色的圈展示在图中。每幅图中红色曲线展示了高斯预测分布的均值，红色阴影区域是均值两侧的一个标准差范围的区域。注意，预测的不确定性依赖于x，且在数据点的邻域内最小。我们可以发现，不确定性的程度随着观测到的数据点的增多而逐渐减小。

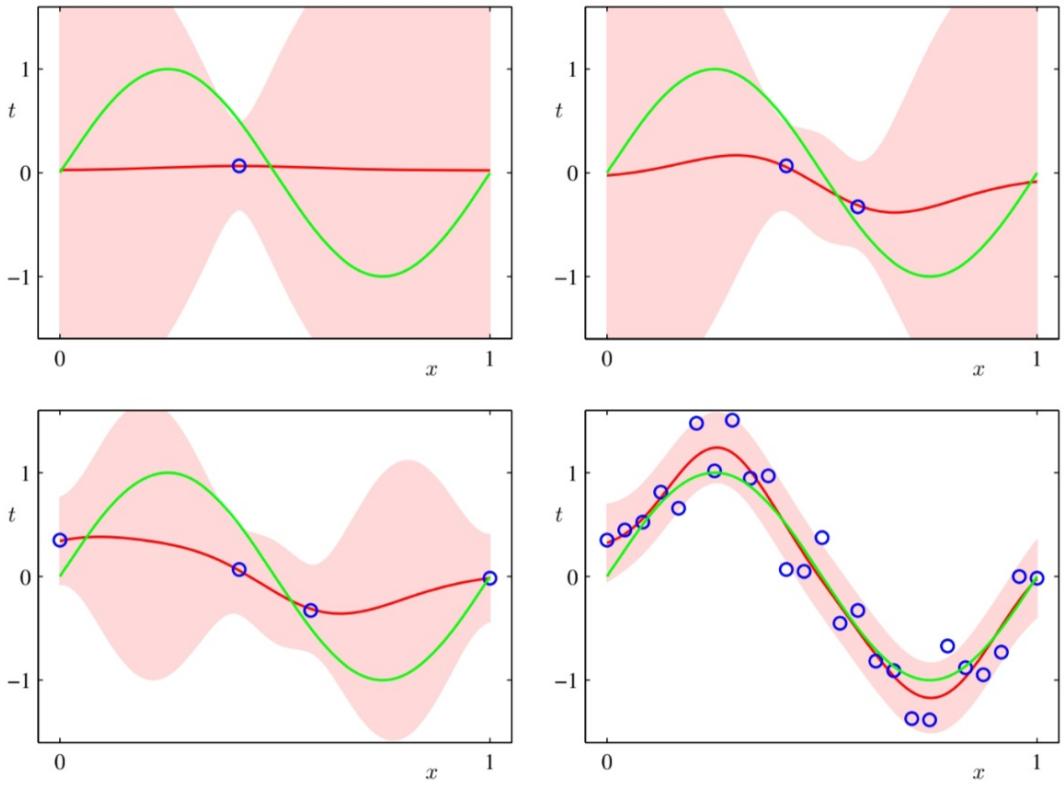


Figure 3.8 Examples of the predictive distribution (3.58) for a model consisting of 9 Gaussian basis functions of the form (3.4) using the synthetic sinusoidal data set of Section 1.1. See the text for a detailed discussion.

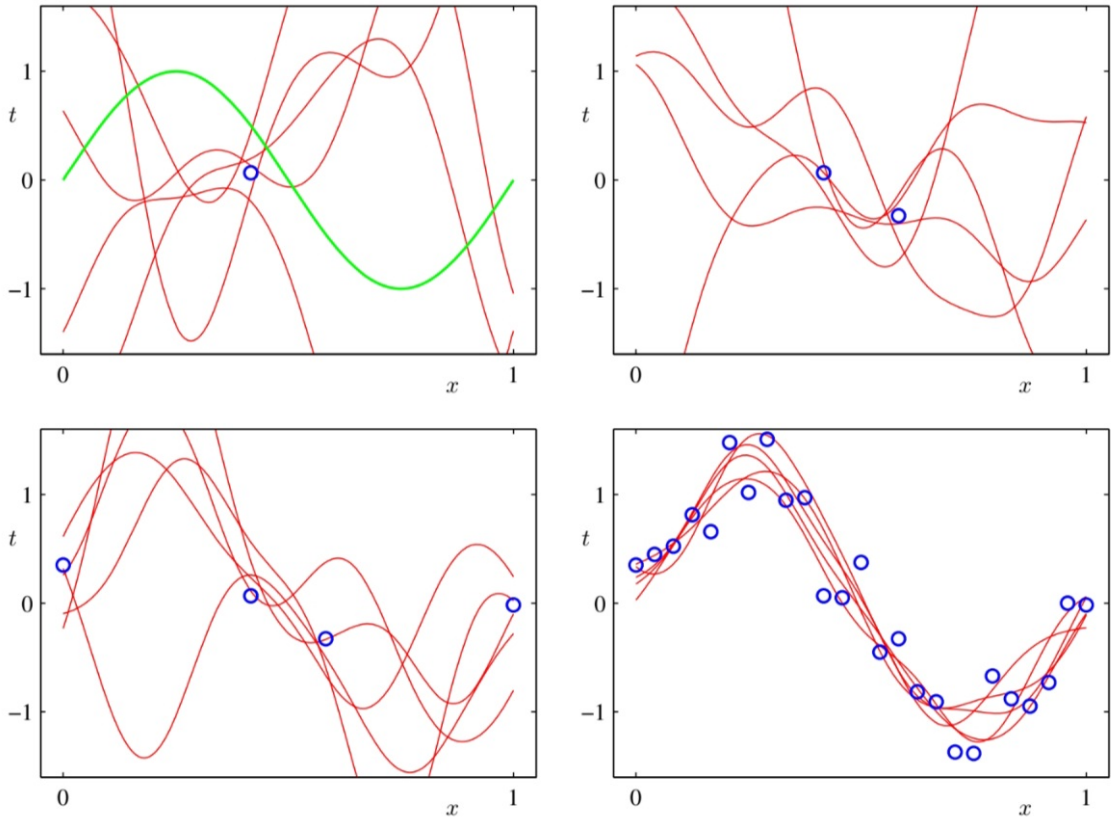


Figure 3.9 Plots of the function $y(x, \mathbf{w})$ using samples from the posterior distributions over $\mathbf{w}$ corresponding to the plots in Figure 3.8.

图3.9 函数$y(x, w)$的图像，使用了服从w上的后验概率分布的样本，对应图3.8

如果我们使用局部化的基函数（如高斯基函数），那么在距离基函数中心比较远的区域，预测方差（3.59）的第二项的贡献将会趋向于0，只剩下噪声贡献 β^{-1} 。

因此，模型会被认为对于处于基函数所在的区域之外的区域所做出的预测十分准确，但是这通常不是我们想要的结果（我们要预测基函数所在区域之内）。我们可以采用一种被称为高斯过程的另一种贝叶斯方法，从而避免这个问题。

~~我还没有验证过，这是《PRML》上提供的解决方案，有没有大佬讲一下？~~

注意，如果 w, β 都被当成未知的，那么根据2.3.6节的讨论，我们可以引入一个由高斯-Gamma分布给出的共轭先验分布 $p(w, \beta)$(Denison et al., 2002)。在这种情况下，预测是一个t分布。

后日谈：

总结：预测分布是贝叶斯线性回归的核心内容，它能够让我们量化新数据预测的不确定性（看图3.8），使得我们能够评估模型是否有很好的预测能力。

小小的问题：
预测分布的形式是否会随着数据集大小而发生变化？~~（非常简单问题，大概）~~

答案我不好说：
我们可以观察图3.8发现：
随着数据集增大，我们的预测分布实际上是收敛的，预测的准确性提高（逼近真实函数）预测分布变窄从而产生更稳定的预测结果

3.3.3 等效核 (Equivalent kernel)

我们可以考虑先验概率 $p(w \mid \alpha)=\mathcal{N}(w \mid 0, \alpha^{-1} I)$，其中 α 是一个常数，I 是单位矩阵。对应的后验概率为 $p(w \mid t)=\mathcal{N}(w \mid m_N, S_N)$。其中：

$$m_N = \beta S_N \Phi^\top t$$

$$\$S_N^{-1} = \alpha I + \beta \Phi^\top \Phi \$$$

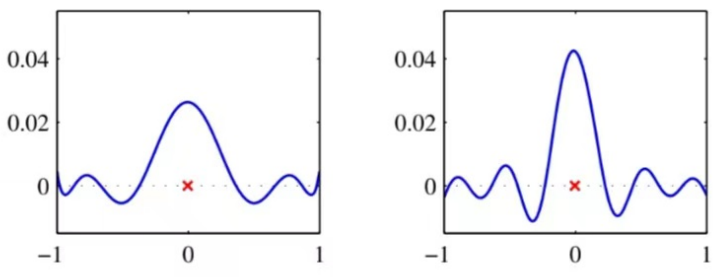
这里的 t 是一组观测数据，Φ 是基函数矩阵，β、α 是超参数。

我们可以使用使用等效核表示，可以将预测函数 $y(x, m_N)$ 表示为：

$$y(x, m_N)=\sum_{n=1}^{N} k(x, x_n) t_n \quad (3.61)$$

其中等效核 $k(x, x')=\beta \phi(x)^\top S_N \phi(x')$，其中$\phi(x)$ 是基函数的映射。

Figure 3.11 Examples of equivalent kernels $k(x, x')$ for $x = 0$ plotted as a function of x', corresponding (left) to the polynomial basis functions and (right) to the sigmoidal basis functions shown in Figure 3.1. Note that these are localized functions of x' even though the corresponding basis functions are nonlocal.



左为多项式基，右为sigmoid级

我们可以发现，在不同基函数下，等效核呈现单峰状，无论基函数如何变化，当 x 固定时，$k(x,x')$ 的函数图像呈现单峰。

两个预测值 $y(x)$ 和 $y(x')$ 的协方差为：

$$\mathrm{cov}[y(x),y(x')]=\beta^{-1}k(x,x') \quad (3.63)$$

这表明在预测中，x 的值与其附近点的预测值有较高的相关性。

使用高斯过程

通过使用等效核的方式，可以将数据预测问题转化为高斯过程，其中 $k(x,x_n)$ 可视为每个数据点的权重，满足

$$\sum_{n=1}^{N}k(x,x_n)=1 \quad (3.64)$$

等效核也可以的定义为以下形式：

$$k(x,z)=\psi(x)^{\top}\psi(z) \quad (3.65)$$

$$\psi(x)=\beta^{1/2}S_N^{1/2}\phi(x)$$

后日谈：

总结： 个人感觉等效核在高斯过程回归中扮演着所谓桥梁的角色，将输入数据、基函数、协方差矩阵和预测值之间的关系联系在一起。

总结

贝叶斯线性回归通过**引入先验分布**和**贝叶斯方法**，处理模型参数的不确定性，避免过拟合问题。

参数分布通过后验概率进行更新，预测分布考虑了新数据预测的不确定性。

等效核将预测问题转化为高斯过程，关联基函数、协方差矩阵和预测值。