4.5 最大似然估计

前面的内容已经定义了一些估计并分析了其性质，但是这些估计是怎么来的，性能如何？希望有一些准则可以指导从不同模型中得到特定函数并作为好的估计，而不是猜测某些函数可能是好的估计，然后分析其偏差和方差。

最常用的准则是最大似然估计。

考虑一组含有 m 个样本的数据集 $\mathbb{X}=\left\{\boldsymbol{x}^{(1)},\cdots \boldsymbol{x}^{(m)}\right\}$，独立地由未知的真实数据生成分布 $p_{\text{data}}(\mathbf{x})$ 生成。

令 $p_{\text{model}}(\mathbf{x};\boldsymbol{\theta})$ 是一族由 $\boldsymbol{\theta}$ 确定在相同空间上的概率分布。换言之，$p_{\text{model}}(\boldsymbol{x};\boldsymbol{\theta})$ 将任意输入 $\boldsymbol{x}$ 映射到实数来估计真实概率 $p_{\text{data}}(\mathbf{x})$。

对 $\boldsymbol{\theta}$ 的最大似然估计被定义为：

$$\begin{aligned}\boldsymbol{\theta}_{\text{ML}} &= \arg\max_{\boldsymbol{\theta}} p_{\text{model}}(\mathbb{X};\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\text{model}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})\end{aligned}$$

概率的乘积会导致很多计算不方便，比如很有可能出现数值计算下溢，所以很自然地想到通过取对数可以将乘积变成求和操作：

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{\text{model}}(\boldsymbol{x}^{(i)};\boldsymbol{\theta})$$

因为当我们重新缩放代价函数时 arg max 不会改变，我们可以除以 m 得到和训练数据经验分布 $\hat{p}_{\text{data}}$ 相关的期望作为准则：

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\text{data}}} \log p_{\text{model}}(\boldsymbol{x};\boldsymbol{\theta})$$

一种解释最大似然估计的观点是将它看作最小化训练集上的经验分布 $\hat{p}_{\text{data}}$ 和模型分布之间的差异，两者之间的差异程度可以通过 KL 散度度量。KL 散度被定义为

$$D_{\text{KL}}(\hat{p}_{\text{data}} \,\|\, p_{\text{model}}) = \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\text{data}}}\left[\log \hat{p}_{\text{data}}(\boldsymbol{x}) - \log p_{\text{model}}(\boldsymbol{x})\right]$$

左边一项仅涉及到数据生成过程，和模型无关。这意味着当我们训练模型最小化

KL 散度时，我们只需要最小化

$$-\mathbb{E}_{\mathbf{x}\sim\hat{p}_{\text{data}}}[\log p_{\text{model}}(\boldsymbol{x})]$$

我们可以将最大似然看作是使模型分布尽可能地和经验分布 $\hat{p}_{\text{data}}$ 相匹配的尝试。理想情况下，我们希望匹配真实的数据生成分布 $p_{\text{data}}(\mathbf{x})$，但我们没法直接知道这个分布。

4.5.1 条件对数似然和均方误差

最大似然估计很容易扩展到估计条件概率 $P(\mathbf{y}\mid\mathbf{x};\boldsymbol{\theta})$，从而给定 $\mathbf{x}$ 预测 $\mathbf{y}$。实际上这是最常见的情况，因为这构成了大多数监督学习的基础。如果 $\boldsymbol{X}$ 表示所有的输入，$\boldsymbol{Y}$ 表示我们观测到的目标，那么条件最大似然估计是

$$\boldsymbol{\theta}_{\text{ML}}=\arg\max_{\boldsymbol{\theta}} P(\boldsymbol{Y}\mid\boldsymbol{X};\boldsymbol{\theta})$$

如果假设样本是独立同分布的，那么这可以分解成

$$\boldsymbol{\theta}_{\text{ML}}=\arg\max_{\boldsymbol{\theta}}\sum_{i=1}^{m}\log P(\boldsymbol{y}^{(i)}\mid\boldsymbol{x}^{(i)};\boldsymbol{\theta})$$

示例：线性回归作为最大似然

在 2.15.1 中通过最小化均方误差准则进行线性回归，现在，我们以最大似然估计的角度重新审视线性回归。我们现在希望模型能够得到**条件概率 $p(y\mid\boldsymbol{x})$**，而不只是得到一个单独的预测 $\hat{y}$。想象有一个无限大的训练集，我们可能会观测到几个训练样本有相同的输入 $\boldsymbol{x}$ 但是不同的 y。现在学习算法的目标是拟合分布 $p(y\mid\boldsymbol{x})$ 到和 $\boldsymbol{x}$ 相匹配的不同的 y。为了得到我们之前推导出的相同的线性回归算法，我们定义 $p(y\mid\boldsymbol{x})=N(y;\hat{y}(\boldsymbol{x};\boldsymbol{\omega}),\sigma^2)$。函数 $\hat{y}(\boldsymbol{x};\boldsymbol{\omega})$ 预测高斯的均值。在这个例子中，我们假设方差是用户固定的某个常量 σ^2。这种函数形式 $p(y\mid\boldsymbol{x})$ 会使得最大似然估计得出和之前相同的学习算法。由于假设样本是独立同分布的，条件对数似然如下

$$\begin{aligned}
\boldsymbol{\theta}_{\text{ML}}&=\arg\max_{\boldsymbol{\theta}}\sum_{i=1}^{m}\log P(\boldsymbol{y}^{(i)}\mid\boldsymbol{x}^{(i)};\boldsymbol{\theta})\\
&=\arg\max_{\boldsymbol{\theta}}\sum_{i=1}^{m}\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp^{-\frac{(\hat{y}^{(i)}-y^{(i)})^2}{2\sigma^2}}\right)\\
&=\arg\max_{\boldsymbol{\theta}}\left(-m\log\sigma-\frac{m}{2}\log(2\pi)-\sum_{i=1}^{m}\frac{\left\|\hat{y}^{(i)}-y^{(i)}\right\|^2}{2\sigma^2}\right)
\end{aligned}$$

其中 $\hat{y}^{(i)}$ 是线性回归在第 i 个输入 $\boldsymbol{x}^{(i)}$ 上的输出，m 是训练样本的数目。对比均方误差和对数似然，

$$\mathrm{MSE}_{\mathrm{train}} = \frac{1}{m}\sum_{i=1}^{m}\left\|\hat{y}^{(i)} - y^{(i)}\right\|^2$$

我们立刻可以看出最大化关于 $\boldsymbol{\omega}$ 的对数似然和最小化均方误差会得到相同的参数估计 $\boldsymbol{\omega}$。但是对于相同的最优 $\boldsymbol{\omega}$，这两个准则有着不同的值。这验证了 MSE 可以用于最大似然估计。

4.5.2 最大似然的性质

最大似然估计最吸引人的地方在于，它被证明当样本数目 $m \to \infty$ 时，就收敛率而言是最好的渐近估计。

在合适的条件下，最大似然估计具有一致性，意味着训练样本数目趋向于无穷大时，参数的最大似然估计会收敛到参数的真实值。这些条件是：

- 真实分布 p_{data} 必须在模型族 $p_{\mathrm{model}}(\cdot;\boldsymbol{\theta})$ 中。否则，没有估计可以还原 p_{data}。
- 真实分布 p_{data} 必须刚好对应一个 $\boldsymbol{\theta}$ 值。否则，最大似然估计恢复出真实分布 p_{data} 后，也不能决定数据生成过程使用哪个 $\boldsymbol{\theta}$。

4.6 贝叶斯统计

至此我们已经讨论了频率派统计（frequentist statistics）方法和基于估计单一值 $\boldsymbol{\theta}$ 的方法，然后基于该估计作所有的预测。另一种方法是在做预测时会考虑所有可能的 $\boldsymbol{\theta}$。后者属于贝叶斯统计（Bayesian statistics）的范畴。

频率派的视角是真实参数 $\boldsymbol{\theta}$ 是未知的定值，而点估计 $\hat{\boldsymbol{\theta}}$ 是考虑数据集上函数（可以看作是随机的）的随机变量。

贝叶斯统计的视角完全不同。贝叶斯用概率反映知识状态的确定性程度。数据集能够被直接观测到，因此不是随机的。另一方面，真实参数 $\boldsymbol{\theta}$ 是未知或不确定的，因此可以表示成随机变量。

在观察到数据前，我们将 $\boldsymbol{\theta}$ 的已知知识表示成先验概率分布（prior probability distribution），$p(\boldsymbol{\theta})$（有时简单地称为“先验”）。一般而言，机器学习实践者会选择一个相当宽泛的（即，高熵的）先验分布，反映在观测到任何数据前参数 $\boldsymbol{\theta}$ 的高度不确定性。例如，我们可能会假设先验 $\boldsymbol{\theta}$ 在有限区间中均匀分布。许多先验偏好于“更简单”的解

（如小幅度的系数，或是接近常数的函数）。

现在假设我们有一组数据样本 $\{x^{(1)},\cdots,x^{(m)}\}$。通过贝叶斯规则结合数据似然 $p(x^{(1)},\cdots,x^{(m)}\mid\boldsymbol{\theta})$ 和先验，我们可以恢复数据对我们关于 $\boldsymbol{\theta}$ 信念的影响：

$$p(\boldsymbol{\theta}\mid x^{(1)},\cdots,x^{(m)})=\frac{p(x^{(1)},\cdots,x^{(m)}\mid\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(x^{(1)},\cdots,x^{(m)})}$$

在贝叶斯估计常用的情景下，先验开始是相对均匀的分布或高熵的高斯分布，观测数据通常会使后验的熵下降，并集中在参数的几个可能性很高的值。

相对于最大似然估计，贝叶斯估计有两个重要区别。第一，不像最大似然方法预测时使用 $\boldsymbol{\theta}$ 的点估计，贝叶斯方法使用 $\boldsymbol{\theta}$ 的全分布。例如，在观测到 m 个样本后，下一个数据样本 $x^{(m+1)}$ 的预测分布如下：

$$p(x^{(m+1)}\mid x^{(1)},\cdots,x^{(m)})=\int p\left(x^{(m+1)}\mid\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\mid x^{(1)},\cdots,x^{(m)}\right)d\boldsymbol{\theta}$$

这里，每个具有正概率密度的 $\boldsymbol{\theta}$ 的值有助于下一个样本的预测，其中贡献由后验密度本身加权。在观测到数据集 $\{x^{(1)},\cdots,x^{(m)}\}$ 之后，如果我们仍然非常不确定 $\boldsymbol{\theta}$ 的值，那么这个不确定性会直接包含在我们所做的任何预测中。

贝叶斯方法和最大似然方法的第二个最大区别是由贝叶斯先验分布造成的。先验能够影响概率质量密度朝参数空间中偏好先验的区域偏移。实践中，先验通常表现为偏好更简单或更光滑的模型。对贝叶斯方法的批判认为先验是人为主观判断影响预测的来源。

当训练数据很有限时，贝叶斯方法通常泛化得更好，但是当训练样本数目很大时，通常会有很大的计算代价。

示例：贝叶斯线性回归

在线性回归中，我们学习从输入向量 $\boldsymbol{x}\in\mathbb{R}^n$ 预测标量 $y\in\mathbb{R}$ 的线性映射。该预测由向量 $\boldsymbol{\omega}\in\mathbb{R}^n$ 参数化：

$$\hat{y}=\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{x}$$

给定一组 m 个训练样本 $\left(\boldsymbol{X}^{(\text{train})},\boldsymbol{y}^{(\text{train})}\right)$，我们可以表示整个训练集对 y 的预测：

$$\hat{\boldsymbol{y}}^{(\text{train})}=\boldsymbol{X}^{(\text{train})}\boldsymbol{\omega}$$

表示为 $\boldsymbol{y}^{(\text{train})}$ 上的高斯条件分布，我们得到

$$p\left(\boldsymbol{y}^{(\text{train})} \mid \boldsymbol{X}^{(\text{train})}, \boldsymbol{\omega}\right) = N\left(\boldsymbol{y}^{(\text{train})}; \boldsymbol{X}^{(\text{train})}\boldsymbol{\omega}, \boldsymbol{I}\right)$$
$$\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{y}^{(\text{train})} - \boldsymbol{X}^{(\text{train})}\boldsymbol{\omega}\right)^{\mathrm{T}}\left(\boldsymbol{y}^{(\text{train})} - \boldsymbol{X}^{(\text{train})}\boldsymbol{\omega}\right)\right)$$

其中，我们根据标准的 MSE 公式假设 y 上的高斯方差为 1。在下文中，为减少符号负担，我们将 $\left(\boldsymbol{X}^{(\text{train})}, \boldsymbol{y}^{(\text{train})}\right)$ 简单表示为 $(\boldsymbol{X}, \boldsymbol{y})$。

为确定模型参数向量 $\boldsymbol{\omega}$ 的后验分布，我们首先需要指定一个先验分布。先验应该反映我们对这些参数取值的信念。虽然有时将我们的先验信念表示为模型的参数很难或很不自然，但在实践中我们通常假设一个相当广泛的分布来表示 $\boldsymbol{\theta}$ 的高度不确定性。实数值参数通常使用高斯作为先验分布：

$$p(\boldsymbol{\omega}) = N(\boldsymbol{\omega}; \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu}_0)^{\mathrm{T}} \boldsymbol{\Lambda}_0^{-1} (\boldsymbol{\omega} - \boldsymbol{\mu}_0)\right)$$

其中，$\boldsymbol{\mu}_0$ 和 $\boldsymbol{\Lambda}_0$ 分别是先验分布的均值向量和协方差矩阵。[3]

确定好先验后，我们现在可以继续确定模型参数的后验分布。

$$p(\boldsymbol{\omega} \mid \boldsymbol{X}, \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\omega})\, p(\boldsymbol{\omega})$$
$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega})^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\omega})\right)\exp\left(-\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu}_0)^{\mathrm{T}} \boldsymbol{\Lambda}_0^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}_0)\right)$$
$$\propto \exp\left(-\frac{1}{2}\left(-2\boldsymbol{y}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\omega} + \boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\omega} + \boldsymbol{\omega}^{\mathrm{T}}\Lambda_0^{-1}\boldsymbol{\omega} - 2\boldsymbol{\mu}_0^{\mathrm{T}}\Lambda_0^{-1}\boldsymbol{\omega}\right)\right)$$

现在我们定义 $\boldsymbol{\Lambda}_m = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{\Lambda}_0^{-1}\right)^{-1}$ 和 $\boldsymbol{\mu}_m = \boldsymbol{\Lambda}_m\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + \boldsymbol{\Lambda}_0^{-1}\boldsymbol{\mu}_0\right)$。使用这些新的变量，我们发现后验可改写为高斯分布：

[3] 除非有理由使用协方差矩阵的特定结构，我们通常假设其为对角协方差矩阵 $\boldsymbol{\Lambda}_0 = \mathrm{diag}(\lambda_0)$。

$$p(\boldsymbol{\omega} \mid \boldsymbol{X}, \boldsymbol{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\omega}-\boldsymbol{\mu}_m)^{\mathrm{T}} \boldsymbol{\Lambda}_m^{-1}(\boldsymbol{\omega}-\boldsymbol{\mu}_m)+\frac{1}{2}\boldsymbol{\mu}_m{}^{\mathrm{T}}\boldsymbol{\Lambda}_m{}^{-1}\boldsymbol{\mu}_m\right)$$
$$\propto \exp\left(-\frac{1}{2}(\boldsymbol{\omega}-\boldsymbol{\mu}_m)^{\mathrm{T}} \boldsymbol{\Lambda}_m^{-1}(\boldsymbol{\omega}-\boldsymbol{\mu}_m)\right)$$

4.6.1 最大后验(MAP)估计

原则上，我们应该使用参数$\boldsymbol{\theta}$的完整贝叶斯后验分布进行预测，但单点估计常常也是需要的。希望使用点估计的一个常见原因是，对于大多数有意义的模型而言，大多数涉及到贝叶斯后验的计算是非常棘手的，点估计提供了一个可行的近似解。我们仍然可以让先验影响点估计的选择来利用贝叶斯方法的优点，而不是简单地回到最大似然估计。一种能够做到这一点的合理方式是选择最大后验（Maximum A Posteriori, MAP）点估计。MAP 估计选择后验概率最大的点（或在$\boldsymbol{\theta}$是连续值的更常见情况下，概率密度最大的点）：

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \arg\max_{\boldsymbol{\theta}} \log p(\boldsymbol{x} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

我们可以认出上式右边的$\log p(\boldsymbol{x} \mid \boldsymbol{\theta})$对应着标准的对数似然项， $\log p(\boldsymbol{\theta})$对应着先验分布。

4.7 监督学习算法

监督学习算法是给定一组输入$\boldsymbol{x}$和输出$\boldsymbol{y}$的训练集，学习如何关联输入和输出。在许多情况下，输出$\boldsymbol{y}$很难自动收集，必须由人来提供“监督”，不过该术语仍然适用于训练集目标可以被自动收集的情况。

4.7.1 概率监督学习

本书的大部分监督学习算法都是基于估计概率分布$p(y \mid \boldsymbol{x})$的。我们可以使用最大似然估计找到对于有参分布族$p(y \mid \boldsymbol{x};\boldsymbol{\theta})$最好的参数向量$\boldsymbol{\theta}$。

我们已经看到，线性回归对应于分布族

$$p(y \mid \boldsymbol{x};\boldsymbol{\theta}) = N(y;\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{I})$$

通过定义一族不同的概率分布，我们可以将线性回归扩展到分类情况中。如果我们有两个类，类 0 和类 1，那么我们只需要指定这两类之一的概率。类 1 的概率决定了类 0 的概率，因为这两个值加起来必须等于 1。

我们用于线性回归的实数正态分布是用均值参数化的。我们提供这个均值的任何值

都是有效的。二元变量上的分布稍微复杂些，因为它的均值必须始终在 0 和 1 之间。解决这个问题的一种方法是使用 logistic sigmoid 函数将线性函数的输出压缩进区间(0,1)。该值可以解释为概率：

$$p(y=1\mid \boldsymbol{x};\boldsymbol{\theta})=\sigma(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x})$$

这个方法被称为逻辑回归（logistic regression）。可以看成是线性回归的拓展，同样是输入 $\boldsymbol{x}$，找到参数 $\boldsymbol{\theta}$，再去计算两者之间的点积，拓展的地方在于使用 sigmoid 函数对线性回归的输出进行了压缩。

由 sigmoid 函数定义可知

$$\delta(x)=\frac{1}{1+\exp(-x)}$$

所以

$$p(y=1\mid \boldsymbol{x};\boldsymbol{\theta})=\frac{1}{1+\exp(-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x})}=\frac{\exp(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x})}{1+\exp(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x})}=\frac{\exp(\boldsymbol{\theta}\cdot\boldsymbol{x}+b)}{1+\exp(\boldsymbol{\theta}\cdot\boldsymbol{x}+b)}$$

$$p(y=0\mid \boldsymbol{x};\boldsymbol{\theta})=1-p(y=1\mid \boldsymbol{x};\boldsymbol{\theta})=\frac{1}{1+\exp(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{x})}=\frac{1}{1+\exp(\boldsymbol{\theta}\cdot\boldsymbol{x}+b)}$$

为了表示方便，将权值向量和输入向量进行扩充，$\boldsymbol{\theta}=\left(\boldsymbol{\theta}^{(1)},\cdots,\boldsymbol{\theta}^{(n)},b\right)^T$，$\boldsymbol{x}=\left(\boldsymbol{x}^{(1)},\cdots,\boldsymbol{x}^{(n)},1\right)^T$，则有

$$p(y=1\mid \boldsymbol{x};\boldsymbol{\theta})=\frac{\exp(\boldsymbol{\theta}\cdot\boldsymbol{x})}{1+\exp(\boldsymbol{\theta}\cdot\boldsymbol{x})}=\pi(\boldsymbol{x})$$

$$p(y=0\mid \boldsymbol{x};\boldsymbol{\theta})=\frac{1}{1+\exp(\boldsymbol{\theta}\cdot\boldsymbol{x})}=1-\pi(\boldsymbol{x})$$

可以应用极大似然估计法估计模型参数，从而得到逻辑回归模型。

似然函数为

$$\prod_{i=1}^{n}\left[\pi(x_i)\right]^{y_i}\left[1-\pi(x_i)\right]^{1-y_i}$$

对数似然函数为

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\left[y_i \log \pi(x_i)+(1-y_i)\log(1-\pi(x_i))\right] \\
&= \sum_{i=1}^{n}\left[y_i \log \frac{\pi(x_i)}{1-\pi(x_i)}+\log(1-\pi(x_i))\right] \\
&= \sum_{i=1}^{n}\left[y_i(\boldsymbol{\theta}\cdot x_i)-\log(1+\exp(\boldsymbol{\theta}\cdot x_i))\right]
\end{aligned}
$$

对其进行求极大值，得到$\boldsymbol{\theta}$的估计值。

$$
\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n}\left[y_i x_i - \frac{\exp(\boldsymbol{\theta} x_i)}{1+\exp(\boldsymbol{\theta} x_i)} x_i\right] = 0
$$

线性回归中，我们能够通过求解正规方程以找到最佳权重。相比而言，逻辑回归会更困难些。其最佳权重没有闭解。反之，我们必须最大化对数似然来搜索最优解。我们可以通过梯度下降算法最小化负对数似然来搜索。

作业:

a. 最大似然估计与贝叶斯估计的区别有哪些?

b. 线性回归中，假设标签$y \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \beta)$，用最大似然估计来优化参数 w。（验证：最大化花书（5.64）式，将得到与（5.12）式相同结果。）

c. 逻辑回归中，若标签$y \in \{0,1\}$，权重$\beta = (-\ln(4), \ln(2), -\ln(3))$，当输入特征向量$x = (1,1,1)$时，求对应标签 y=1 的概率。

d. 我们通过梯度下降算法最小化负对数似然求解逻辑回归，试给出更新公式。

图 32

a、一方面，最大似然估计认为真实参数是未知的固定值，而贝叶斯估计认为是未知不确定的，是符合一定分布的随机变量；另一方面，前者在预测时使用的是参数的点估计，后者则是参数的全分布，并且贝叶斯引入了先验，先验能够影响概率质量密度朝参数空间中偏好先验的区域偏移。

b、通过最大化条件对数似然，有

$$\begin{aligned}
\boldsymbol{\omega}_{\mathrm{ML}} &= \arg\max_{\boldsymbol{\omega}} \sum_{i=1}^{m} \log P(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\omega}) \\
&= \arg\max_{\boldsymbol{\omega}} \sum_{i=1}^{m} \log\left(\frac{1}{\sqrt{2\pi}\beta} \exp^{-\frac{(y^{(i)}-\boldsymbol{\omega}^{\mathrm{T}}x^{(i)})^2}{2\beta^2}} \right) \\
&= \arg\max_{\boldsymbol{\omega}} \left(-m\log\beta - \frac{m}{2}\log(2\pi) - \sum_{i=1}^{m} \frac{\left\| y^{(i)} - \boldsymbol{\omega}^{\mathrm{T}} x^{(i)} \right\|^2}{2\beta^2} \right) \\
&\Rightarrow \arg\min_{\boldsymbol{\omega}} \left(\sum_{i=1}^{m} \left\| y^{(i)} - \boldsymbol{\omega}^{\mathrm{T}} x^{(i)} \right\|^2 \right)
\end{aligned}$$

所以最大化条件对话似然等价于最小化均方误差，对$\boldsymbol{\omega}$求导，可得

$$\boldsymbol{w} = \left(\mathbf{x}^{\mathrm{T}}\mathbf{x}\right)^{-1} \mathbf{x}^{\mathrm{T}} \boldsymbol{y}$$

c、

$$\begin{aligned}
p(y=1 \mid \boldsymbol{x}; \boldsymbol{\beta}) &= \frac{\exp(\boldsymbol{\beta}\cdot\boldsymbol{x})}{1+\exp(\boldsymbol{\beta}\cdot\boldsymbol{x})} \\
&= \frac{\exp\left(\begin{pmatrix} -\ln(4) & \ln(2) & -\ln(3) \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right)}{1+\exp\left(\begin{pmatrix} -\ln(4) & \ln(2) & -\ln(3) \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right)} \\
&= \frac{\frac{1}{6}}{1+\frac{1}{6}} = \frac{1}{7} \approx 0.143
\end{aligned}$$

所以 y=1 的概率为 0.143。

d、由$\dfrac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n}\left[y_i x_i - \dfrac{\exp(\boldsymbol{\theta}x_i)}{1+\exp(\boldsymbol{\theta}x_i)} x_i \right]$得负对数

$$-\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -\sum_{i=1}^{n}\left[y_i x_i - \frac{\exp(\boldsymbol{\theta}x_i)}{1+\exp(\boldsymbol{\theta}x_i)} x_i \right] = \sum_{i=1}^{n}\left[\left(\frac{\exp(\boldsymbol{\theta}x_i)}{1+\exp(\boldsymbol{\theta}x_i)} - y_i \right) x_i \right] = \sum_{i=1}^{n}\left[\left(\pi\left(x_i\right) - y_i \right) x_i \right]$$

则

$$-\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{(j)}}=\sum_{i=1}^{n}\left[\left(\pi\left(x_i\right)-y_i\right)x_i^{(j)}\right]$$

所以有如下梯度下降法的迭代公式：

$$\boldsymbol{\theta}^{(j)}=\boldsymbol{\theta}^{(j)}-\alpha\sum_{i=1}^{n}\left[\left(\pi\left(x_i\right)-y_i\right)x_i^{(j)}\right]$$