第三章

本章目的

前面提及的都是无监督学习，比如密度估计或聚类；本章开始讨论有监督学习，从回归入手。

回归的目的是，基于给定的一个D维样本点x，预测其对应的一个或多个连续目标变量 t。

对于输入变量，可以构造出一套基函数，它们只需要是关于参数的线性函数，而不一定是关于输入变量的线性函数。

对于给定的N个样本点构成的数据集，以及其对应的N个目标值，需要基于已知数据对新的样本点x预测其对应的目标值。最简单的方法是直接构造一个合适的y(x)，对新样本点，预测其目标值为t=y(x)。从概率的角度说，我们希望能够对预测分布 p(t|x) 建模，因为这个分布表示了对于每个新的x，其可能的目标值t的不确定性。基于p(t|x)，我们选择能够最小化所规定的损失函数的期望的 t。

$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$

其中， $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(x), \ldots, \phi_{M-1}(\mathbf{x}))^T$, 称为基函数。
$\mathbf{w} = (w_1, w_2, \ldots, w_{M-1})^T$

通常定义一个额外的虚"基函数"，$\phi_0(\mathbf{x}) = 1$, 则
$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$

基函数的选择:

- 最简单的形式 $y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$。由于 $y(\mathbf{x}, \mathbf{w})$是关于 $\mathbf{w}$ 的线性函数，故称之为 **线性模型**
- **幂次函数**: $\phi_j(x) = x^j$ ，是全局的global，small change in x 会影响所有的基函数

- **高斯函数**: $\phi_j(x) = exp\{-\frac{(x-\mu_j)^2}{2s^2}\}$, 是局部的 local，small change in x 只会影响附近的基函数
- **sigmoid函数**: $\phi_j(x) = \sigma(\frac{x-\mu_j}{s}), \sigma(a) = \frac{1}{1+exp(-a)}$，也是局部的，small change in x 只会影响附近

3.1 最大似然估计和最小二乘法

对于多项式拟合问题：

通过最小化均方误差得到的结果 与

【假设噪声项服从高斯分布，从频率论的角度进行最大似然估计，得到的结果】 是一致的。

下面详细探讨二者关系

1. 设目标值为t，则 $t = y(\mathbf{x}, \mathbf{w}) + \epsilon$， 假设 ϵ 服从高斯分布， $\epsilon \sim \mathcal{N}(0, \beta^{-1})$。

   则 $p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1})$
2. 考虑一个数据集，一共有N个样本，每个样本对应一个目标值

   $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, 对应目标值 $\mathbf{t} = (t_1, \ldots, t_N)^T$

   假设这些数据点是独立地从分布中抽取的，则

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

   对于有监督学习的回归和分类问题，并不是关注于对输入变量x的概率分布建模，x已经给定，为了保持简洁，省略掉x，对上式求对数：

   $\ln p(\mathbf{t}|\mathbf{w}, \beta) = \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^T\phi(\mathbf{x}_n), \beta^{-1}) = \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})$

   其中，平方误差函数 $E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2$ **相当于最大化似然函数等价于最小化平方误差损失**

   将对数函数 对于w 求导，令梯度为零，

   $\bigtriangledown \ln p(\mathbf{t}|\mathbf{w}, \beta) = \beta\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}\phi(\mathbf{x}_n)^T = 0$

   得 $\mathbf{w}_{ML} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{t}$

   这是最小二乘问题的规范方程， Φ是一个N*M的矩阵，称为设计矩阵(design matrix)

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \cdots & \phi_{M-1}(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \cdots & \phi_{M-1}(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \cdots & \phi_{M-1}(\boldsymbol{x}_N) \end{pmatrix}$$

$$\boldsymbol{\Phi}^{\dagger} \equiv (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T$$

$\Phi^{\dagger}$ 为 Φ 的伪逆矩阵，可以被看作逆矩阵对于非方阵的推广。如果 Φ 为方阵，显然 $\Phi^{\dagger} = \Phi^{-1}$

对于最大似然函数，也可以求关于噪声参数的ML

求得 $\frac{1}{\beta_{ML}} = \frac{1}{N}\sum_{n=1} N\{t_n - \mathbf{w}_{ML}^T\phi(\mathbf{x}_n)\}^2$， 即$\beta_{ML}$倒数为目标值在回归函数周围的残留方

- 认识偏置参数 w_0：

  将平方误差函数 $E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w_0 - \sum_{j=1}^{M-1} w_j\phi_j(\mathbf{x}_n)\}^2$

  对 w_0 求导，令导数为0，则

  $w_0 = \bar{t} - \sum_{j=1}^{M-1} w_j\bar{\phi}_j$，其中 $\bar{t} = \frac{1}{N}\sum_{n=1}^{N} t_n, \bar{\phi}_j = \frac{1}{N}\sum_{n=1}^{N}\phi_j(\mathbf{x}_n)$

  则 w_0 的意义在于其补偿了目标值的平均值(在训练集上的) 与 基函数的值的平均值(每个基函数在训练集上的平均值)的加权求和 之差。

3. 对最小二乘解的**几何解释**：

   假设数据集为N维，则N个样本对应的N个目标值 $\mathbf{t}$也是在N维空间上的一点，坐标为 $(t_1, t_2, \ldots, t_N)$， 即$\mathbf{t} = (t_1, t_2, \ldots, t_N)^T$是这个空间中的一个向量。
   而，同一个基函数对于每个数据点($\mathbf{x}_n$是列向量，有M维)也有N个值，每个值是 $\phi_j(\mathbf{x}_n)$， 也是在N维空间中的一个向量，记作 φ_j。(注意 φ_j 是 Φ 矩阵中的第 j 列，而 $\phi(\mathbf{x}_i)$ 是 Φ 矩阵的第 i 行)

   如果基函数的数量M(每个数据点本身的维度)小于数据集的大小N，则M个向量 φ_j 会张成 M维的子空间S。

定义 $\mathbf{y}$ 是一个N维向量， 它的第n个元素为 $y(\mathbf{x}_n, \mathbf{w})$，则它是M个 φ_j 的任意线性组合[φ_j也在N维空间中，且公式的角度上看也确实是如此]。故 $\mathbf{y}$一定位于由M维子空间S上的任意位置[也位于N维空间]。

因此，平方误差函数为 $\mathbf{y}$ 和 $\mathbf{t}$ 的欧氏距离，$\mathbf{w}$的最小二乘解对应于位于子空间S的与 $\mathbf{t}$最近的$\mathbf{y}$的选择，直觉上可以理解为这个解对应于$\mathbf{t}$在子空间S上的正交投影。

序列化学习

对于批处理技术，如最大似然估计，一次需要计算整个数据集，耗费较大。当数据集较大时，可考虑使用序列化学习，即在线学习，一次只考虑一个数据点，模型参数随之不断更新。

常见的序列化学习方法，有随机梯度算法:

$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \bigtriangledown E_n$

代入上面的平方误差函数

$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta(t^n - \mathbf{w}^{(\tau)T}\phi_n)\phi_n$, 其中 $\phi_n = \phi(\mathbf{x}_n)$ 。该公式为最小均方算法。

正则项

为了防止过拟合引入正则项。一个最简单的形式是
$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) = E_D(\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$

更一般化的形式是 $E_D(\mathbf{w}) + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$, q=1时为Lasso(L1正则项)， q=2时为L2正则项

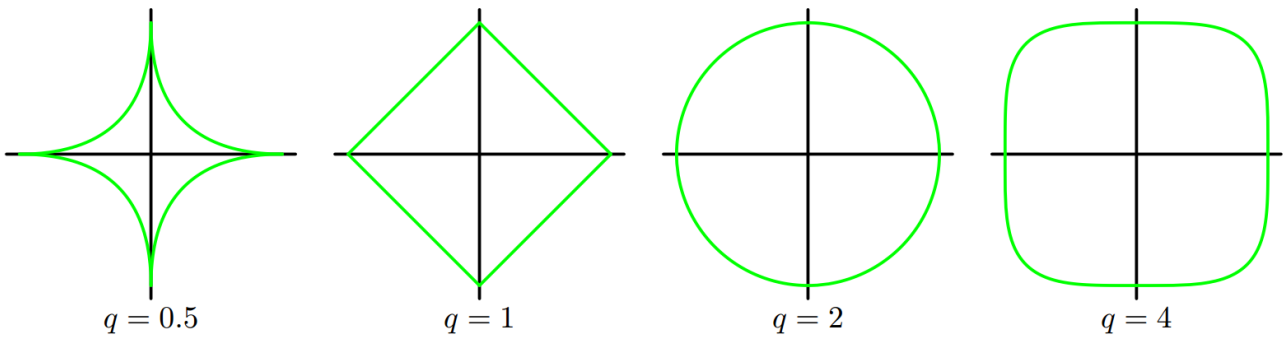

图 3.3: 对于不同的参数q，公式（3.29）中的正则化项的轮廓线。

q=1时，如果λ 充分大，那么某些系数w_j 会变为零，从而产生一个稀疏的解(即待求参数个数变少)

多个输出 multiple output

多个输出，即目标变量有K维，若一个数据点有M维，则权重W为M*K矩阵，基函数为M维向量。对每一维目标变量分析，类似一维目标变量时的情况。

两种方法:

1. 将输出看作互相独立，每个输出对应一套基；
2. 所有的输出都用相同的基，只是使用不同的w来计算(更常用)

3.2 偏置-方差分解

最大似然估计或最小二乘法会存在过拟合的问题，对此的一些解决方案：

- 限制基函数的数量，然而这样的模型难以捕捉到数据中的一些有趣且重要的趋势
- 加入正则项，然而正则项的参数λ 难以确定合适的值

区分决策论中的平方损失函数以及最大似然估计中的平方误差和：

若损失函数为平方损失函数，则模型最优估计$h(\mathbf{x})$ 通过t的条件期望求出，即
$h(\mathbf{x}) = \mathbb{E}(t|\mathbf{x}) = \int tp(t|\mathbf{x})dt$

期望的平方损失可以写成：
$\mathbb{E}(L) = \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x})d\mathbf{x} + \int\int \{h(\mathbf{x}) - t\}^2 p(\mathbf{x}, t)d\mathbf{x}dt$

式子的第二项是由数据的噪声造成的， 是**期望的平方损失能达到的最小值**（即当第一项为零时)

- $y(\mathbf{x})$是模型对于x的估计结果，t是真实值，$h(\mathbf{x})$是模型对于x能得到的最优估计值
- 平方损失衡量的是 $\{y(\mathbf{x}) - t\}^2$，当平方损失最小的时候，即 $y(\mathbf{x}) = h(\mathbf{x})$，此时损失为$\{h(\mathbf{x}) - t\}^2$

式子的第一项依赖于$y(\mathbf{x})$的选择，需要选择一个最优的解使得整个式子最小化，则最优解应使得第一项为零。

- 若有无穷的数据集，则原则上能够以任意精度找到回归函数$h(\mathbf{x})$，从而给出$y(\mathbf{x})$的最优解，然而实际上数据集有限，不能够精确知道回归函数$h(\mathbf{x})$
- 假设用$y(\mathbf{x}, \mathbf{w})$对$h(\mathbf{x})$建模:
  - *从贝叶斯的观点看*，模型的不确定性是通过$\mathbf{w}$的后验概率分布体现的，即$\mathbf{w}$不是固定值，而是随机变量
  - *从频率学角度看*，模型基于数据集D对参数$\mathbf{w}$进行点估计，假设数据集无限大，然而每次只能选择其中一个样本来处理，因此一个样本导致一个$\mathbf{w}$(但对数据集而言$\mathbf{w}$本身是未知的常量)
  - 假设一共有很多个数据集，每个数据集的大小是N，这些数据集都是独立地从同一个分布$p(t, \mathbf{x})$中抽取的。对于某个给定的数据集，通过学习算法可以得到模型$y(\mathbf{x}; \mathcal{D})$。根据在不同数据集上的模型表现，取平均，可以用于评估这个学习算法。模型的表现使用平方损失衡量，即

$$
\begin{aligned}
&\{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2 \\
&= \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] + \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\
&= \{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})\}^2 + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2 \\
&+ 2\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})\}\{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}
\end{aligned}
$$

    将该式子对于$\mathcal{D}$ 求期望，**不知道为什么上式的最后一项等于零**，

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\{y(\mathbf{x}; \mathcal{D}) - h(\mathbf{x})\}^2] \\
&= \mathbb{E}_{\mathcal{D}}[\{y(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})\}^2] + \{\mathbb{E}_{\mathcal{D}}[y(\mathbf{x}; \mathcal{D})] - h(\mathbf{x})\}^2
\end{aligned}
$$

    第一项是方差，衡量的是对于单独的数据集，模型给出的解在平均值附近波动的情况，即学习到的模型和它自己的期望的偏离程度。

    第二项是偏置的平方，衡量的是所有数据集的平均预测与 预期的回归函数之间的差异，即学习到的模型和真实模型的偏离程度。

由此，期望平方损失的式子可以改写成 期望损失 = 偏置^2 + 方差 + 噪声

偏置-方差折中

对于灵活的模型(比如多项式拟合中使用高阶多项式拟合)，总是有较低的偏置(容易过拟合)和较高的方差(波动性大)；对于相对刚性的模型，有较高的偏置和较低的方差。最优的模型能够在偏置和方差中取得最好的平衡。

偏置-方差分解依赖于对所有的数据集求平均，然而实际上可能只有一个观测数据集。如果有多个独立的数据集，最好的方法是将它们组合成一个更大的数据集，来降低过拟合程度。

偏置-方差，从频率学角度为模型复杂度提供了一些认识。

模型的复杂度可通过设置基函数的数量或引入正则项来调整

3.3 贝叶斯线性回归

贝叶斯将参数w 看作随机变量，参数求解则是通过计算w的后验分布，即$p(w|t)$

而在回归问题中，我们更关心的是预测分布 $p(t|D,\alpha,\beta)$，其中D 是已有数据集，α,β 为控制参数w的分布的参数。即根据已有数据集以及参数，对新样本点的预测分布

1.参数分布

假设参数的先验分布$p(w)=N(w|m_0,S_0)$，则后验$p(w|t)=N(w|m_N,S_N)$

$m_N=S_N(S_0^{-1}m_0+\beta\Phi^Tt),\ S_N^{-1}=S_0^{-1}+\beta\Phi^T\Phi$

简单起见，假设参数的先验$p(w|\alpha)=N(w|0,\alpha^{-1}I)$

对应的后验 由上面公式给出，其中 $m_N=\beta S_N\Phi^Tt\quad S_N^{-1}=\alpha I+\beta\Phi^T\Phi$

后验分布的对数由对数似然与先验的对数求和得到，它们都是关于w的函数

$\ln p(w|t)=-\frac{\beta}{2}\sum_{n=1}^N\{t_n-w^T\phi(x_n)\}^2-\frac{\alpha}{2}w^Tw\ +$ 常数

因此对后验分布最大化等价于最小化(平方误差函数+二次正则项)

参数更新过程：每处理一个新的数据点，当前的参数先验分布为上一次的参数后验分布，将当前的先验分布与似然函数相乘，得到当前的后验分布

2.预测分布

$\mathbf{t},\mathbf{x}$ 为已有的数据集，α,β 为控制着模型参数$\mathbf{w}$ 的分布的参数。预测分布即如何根据已有的数据集，对新样本点的预测分布。

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

其中预测分布的方差$\sigma_N^2(x)=\frac{1}{\beta}+\phi(x)^T S_N\phi(x)$

公式的第一项是数据中的噪声，第二项则反映与参数w关联的不确定性[S_N是 w 的后验分布的协方差]

由于噪声和参数w是相互独立的高斯分布，故可以相加。当额外的数据点被观测到，w的后验概率分布会变窄，从而$\sigma_{N+1}^2(x)\leq\sigma_N^2(x)$，当$N\to\infty$，第二项趋于零。

3. 等价核

由决策论知，若损失函数定义为平方损失，则最优预测由条件期望给出，即$E_t[t|x]$

而由预测分布一节知，$p(t|x)$ 服从高斯分布，则均值$E[t|x]=mean=m_N^T\phi(x)$

即最优预测
$y(\mathbf{x},m_N)=m_N^T\phi(x)=\beta\phi(x)^T S_N\Phi^T t=\sum_{n=1}^N\beta\phi(x)^T S_N\phi(x_n)t_n=\sum_{n=1}^N k(x,x_n)t_n$

其中$k(x,x_n)$ 称为等价核

非局部(non-local)的基函数具有局部的等价核，因此这种情况下可以用等价核代替基函数

3.4 贝叶斯模型比较

与最大似然估计相关联的过拟合问题，可以通过对模型的参数进行求和或积分的方式(而不是进行点估计)来避免。

1.模型证据 model evidence

假设一共有L个模型$\{\mathcal{M}_i\}$，则对于每个模型，$P(\mathcal{M}_i|\mathcal{D})\propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i)$

式子的左边为后验概率，右边第一项为先验，第二项为模型证据model evidence 也称作边缘似然marginal likelihood。 贝叶斯因子 $\frac{p(\mathcal{D}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_j)}$

2.模型平均 model averaging

预测分布可以写成 $p(t|X,D)=\sum_{i=1}^{L}p(t|X,M_i,D)p(M_i|D)$

- 此即混合分布，即整体的预测分布是将每个模型的预测值加权平均得到的。

3.模型选择 model selection

对于模型平均的近似情况是只选择一个模型来做出预测，称为模型选择

对于由参数w 控制的模型，$p(D|M_i)=\int p(D|w,M_i)p(w|M_i)dw$。 从取样的角度来看，边缘似然函数可以看成从一个模型中生成数据集D的概率，这个模型的参数是从先验分布中随机取样的。

通过对参数w的积分进行简单的近似，来更加深刻地认识模型证据：

- 首先考虑模型由一个参数w的情形，这个参数的后验概率正比于 $p(D|w)p(w)$，其中为了简化记号，我们省略它对模型M_i 的依赖。
- 如果我们假定后验分布在最大可能值w_{MAP} 附近是一个尖峰，宽度为 $\Delta w_{posterior}$ ，那么我们可以用被积函数的值乘以尖峰的宽度来近似这个积分。进一步假设先验分布是平的(均匀分布)，宽度为 ，即$\Delta w_{prior},p(w)=\frac{1}{\Delta w_{prior}}$ ，那么我们有：

  $p(D)=\int p(D|w)p(w)dw\approx p(D|w_{MAP})\frac{\Delta w_{posterior}}{\Delta w_{prior}}$

  取对数，$\ln p(D)\approx \ln p(D|w_{MAP})+\ln\frac{\Delta w_{postperior}}{\Delta w_{prior}}$
- 第一项表示由最可能的模型参数对数据的拟合效果。对于平的先验分布来说，对应于对数似然。
- 第二项根据模型的复杂度来惩罚模型。
  - 由于 $\Delta w_{posterior}<\Delta w_{prior}$因此这一项为负，并且随着 $\frac{\Delta w_{posterior}}{\Delta w_{prior}}$减小，它的绝对值会增加。
  - 因此如果参数精确地调整为后验分布的数据，那么惩罚项会很大。

对于一个有M个参数的模型，我们可以对每个参数进行类似的近似。假设所有参数的 $\frac{\Delta w_{posterior}}{\Delta w_{prior}}$都相同，我们有：

$$\ln p(D)\approx p(D|w_{MAP})+M\ln\frac{\Delta w_{posterior}}{\Delta w_{prior}}$$

模型复杂度的讨论:

因此，在这种非常简单的近似下，复杂度的惩罚项的大小随着模型中参数的数量M线性增加。

随着我们增加模型的复杂度，第一项通常会decrease，因为一个更复杂的模型能更好地拟合数据。而第二项会减小，因为它依赖于参数个数M的大小。

由最大证据确定的最优模型复杂度需要在这两个相互竞争的项之间进行折中。

贝叶斯模型比较框架中，隐含的一个假设是，生成数据的真实的概率分布包含在考虑的模型集合当中，平均来看(即对贝叶斯因子求期望)，贝叶斯模型比较会倾向于选择正确的模型。

3.5 证据近似

完整的贝叶斯回归模型，还会引入对于超参数α, β 的先验分布，然而这在求积分之后通常是十分复杂的形式。一种解决方法是将超参数设置为一些特定的值。[a.k.a. empirical Bayes, type II or generalized ML or evidence approximation]

贝叶斯模型的预测分布，完整写法
$p(t|\mathbf{t}) = \iiint p(t|\mathbf{w},\beta)p(\mathbf{w}|\mathbf{t},\alpha,\beta)p(\alpha,\beta|\mathbf{t})d\mathbf{w}d\alpha d\beta$

近似写作 $p(t|\mathbf{t}) \approx p(t|\mathbf{t},\hat{\alpha},\hat{\beta}) = \int p(t|\mathbf{w},\hat{\beta})p(\mathbf{w}|\mathbf{t},\hat{\alpha},\hat{\beta})d\mathbf{w}$