PRML

Chowwy

2015-12-11

Notation

x:一个多维的样本输入$\{x_1, x_2..., x_D\}$
x:多个一维样本组成的矩阵
X:有多个多维样本组成的矩阵,第n行为x_n^T

1 Bayes' Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{1}$$

基于频率方法，做多次iid的实验,次数趋近无限则频率等于概率。认为概率分布参数是一个定值,去求那个定值.(客观性) 贝叶斯方法(主观性,用不断得到的数据和现象修正自己的主观性,学习的过程,频率方法认为只有样本可以用概率描述,而参数是无法用概率描述的,因为他客观存在,而Bayes方法认为由于我们当前无法认识到参数的本质,所以对我们而言可以用概率分布来描述这个参数)，引入先验概率，再用观察到的数据修正先验概率，得到后验概率,认为概率分布参数也满足一个概率分布,试图描述在观察数据下参数的分布。连续投掷一枚硬币(不知道均匀与否),5次正面朝上(可做无限次实验，但有些事件无法做无限次实验).

$$posterior \propto likelihood \times prior$$

$p(X|Y)$:likelihood,$p(Y)$:prior(计算复杂)(先验分布的选择对于后验概率影响很大)

2 Gauss Distribution

单变量:

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

多变量:

$$\mathcal{N}(\mathrm{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathrm{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathrm{x}-\boldsymbol{\mu})\right\}$$

3 Linear Regression

3.1 参数选择(频率方法)

房子的例子

$$y(\mathrm{x},\mathrm{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathrm{x}) \tag{2}$$

令$\phi_0(\mathrm{x0}) = 1$

$$y(\mathrm{x},\mathrm{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathrm{x}) = \mathrm{w}^T \phi(\mathrm{x})$$

给出一组样本输入x和输出t,我们假定t是由一个确定的函数y(x,w)加上一个噪声组成,(同一个输入值投1000次硬币,正面朝上的次数不是固定值,而是一个分布)

$$t = y(\mathrm{x},\mathrm{w}) + \epsilon$$

ϵ是一个平均值为0,方差为β^-1(我们称β为准确度precision) t的概率分布为

$$p(t|\mathrm{x},\mathrm{w},\beta) = \mathcal{N}(t|y(\mathrm{x},\mathrm{w}),\beta^{-1})$$

最优的预测应该等于t的条件期望

$$\mathbb{E}[t|\mathrm{x}] = \int tp(t|\mathrm{x})dt = y(\mathrm{x},\mathrm{w})$$

最大似然(当前出现即概率最大,结果应该是怎样)

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

将Gauss分布带入,再取对数ln

$$\begin{aligned} p(\boldsymbol{t}|\mathrm{X},\mathrm{w},\beta) &= \prod_{n=1}^{N} ln\mathcal{N}(t_n|\mathrm{w}^T\phi(\mathrm{x}_n),\beta^{-1}) \\ &= \frac{N}{2}ln\beta - \frac{N}{2}ln(2\pi) - \beta E_D(\mathrm{w}) \end{aligned} \tag{3}$$

这里我们把$E_D(\mathrm{w})$称为误差平方和,最小二乘就是Gauss分布的最大似然

$$E_D(\mathrm{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathrm{w}^T\phi(\mathrm{x}_n)\}^2$$

为了求w,对w^T求偏导(即对最大似然的求解)

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

令导数等于0

$$0 = \sum_{n=1}^{N} t_n \phi(\mathrm{x}_n)^T - \mathrm{w}^T \left(\sum_{n=1}^{N} \phi(\mathrm{x}_n) \phi(\mathrm{x}_n)^T \right)$$

得到

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

其中

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

几何理解, $(\Phi^T \Phi)$取逆的条件
单独看看w_0

$$E_D(\mathrm{w}) = \frac{1}{2} \sum_{n=1}^{N} \{t_n - w_0 - \sum_{j=1}^{M-1} w_j \phi_j(\mathrm{x}_n)\}^2$$

对w_0求导

$$w_0 = \bar{t} - \sum_{j=1}^{M-1} w_j \bar{\phi_j}$$

w_0是用来补偿基函数的不足的(例子),顺便可以把β求出来

3.1.1 正则项(属于选择参数一部分)

为了弥补最大似然方法过拟合的情况,我们加入一个正则项

$$E_W(\mathrm{w}) = \frac{1}{2} \mathrm{w}^T \mathrm{w}$$

误差函数(损失函数)变成了,λ 控制一个正则化的度,对于E_D和E_W的侧重问题(模型复杂度和误差)

$$\frac{1}{2} \sum_{n=1}^{N} \{t_n - \mathrm{w}^T \phi(\mathrm{x}_n)\}^2 + \frac{\lambda}{2} \mathrm{w}^T \mathrm{w}$$

l1范数, λ足够大,可以达到稀疏效果

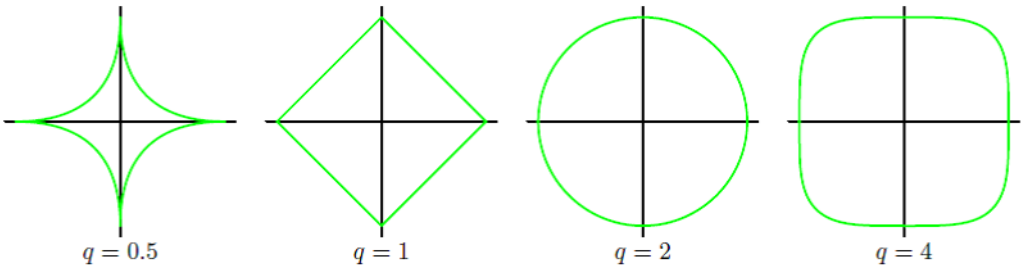

Figure 3.3 Contours of the regularization term in (3.29) for various values of the parameter q.

图 1: Regularization Term

Figure 3.4 Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector $\mathbf{w}$ is denoted by $\mathbf{w}^\star$. The lasso gives a sparse solution in which $w_1^\star = 0$.

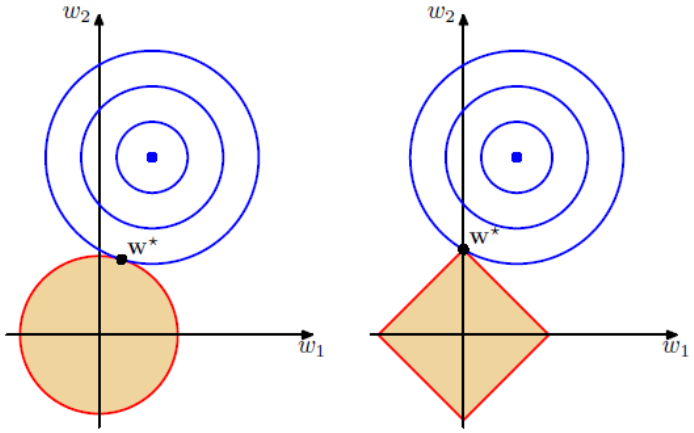

图 2: L1L2Regularization Term

3.1.2 多维目标值t

W每一列对映一个输出值

$$\mathrm{W}_{ML} = (\Phi^T\Phi)^{-1}\Phi^T\mathrm{T}$$

T的每一行对映一个多维目标值

3.2 模型选择

上面我们谈论的是固定模型下,参数的选择,现在我们来考虑模型的选择(基函数的选择),通常我们把数据集分成三份,一份用来训练(训练集training),一份用来选择(验证集validation)参数λ和模型,一份用来评估泛化能力(测试集testing)。

3.2.1 交叉验证

数据很少时,如果验证集中数据很少容易出现噪声干扰的情况,这时我们采用交叉验证来避免偶然性,即把数据集分成S份,验证集每次去其中一份,用剩下的S-1份来训练,将S次结果做平均就是我们最后选取的值。

3.3 bias-variance分解

再探讨bias-variance问题,y(x)是我们预测的t值,我们定义预测值和t有偏差则损失为L(t,y(x)),损失函数的期望为(即对泛化能力的评估,拿来预测真实数据的能力),我们真正要优化的是什么

$$\mathbb{E}[L] = \int\int L(t, y(\mathrm{x}))p(\mathrm{x}, t)d\mathrm{x}dt$$

我们通常将损失函数设为$L(t, y(\mathrm{x})) = \{y(\mathrm{x}) - t\}^2$

$$\mathbb{E}[L] = \int\int \{y(\mathrm{x}) - t\}^2 p(\mathrm{x}, t)d\mathrm{x}dt$$

我们定义

$$h(\mathrm{x}) = \mathbb{E}[t|\mathrm{x}] = \int tp(t|\mathrm{x})dt$$

h(x)是客观存在的t的期望, 我们发现每次都要考虑t的分布，然而我们并不关心t的分布,最好的情况就是y(x)=h(x),所以我们令

$$\begin{aligned}\{y(\mathrm{x}) - t\}^2 &= \{y(\mathrm{x}) - h(\mathrm{x}) + h(\mathrm{x}) - t\}^2 \\ &= \{y(\mathrm{x}) - h(x)\}^2 + 2\{y(\mathrm{x}) - h(\mathrm{x})\}\{h(\mathrm{x}) - t\} + \{h(\mathrm{x}) - t\}^2\end{aligned} \tag{4}$$

积分后,$\int tp(t|\mathrm{x})dt = h(x)$,所以第二项为0,只剩下第一项和第三项

$$\mathbb{E}[L] = \int \{y(\mathrm{x}) - h(\mathrm{x})\}^2 p(\mathrm{x})d\mathrm{x} + \int\int \{h(\mathrm{x}) - t\}^2 p(\mathrm{x}, t)d\mathrm{x}dt$$

第二项与我们的预测无关,属于固有误差,我们只看第一项,假设我们有无穷多的样本,$y(\mathrm{x};\mathcal{D})$为从中抽取的一部分样本$\mathcal{D}$所训练出的预测函数,$\mathbb{E}_{\mathcal{D}}[y(\mathrm{x};\mathcal{D})]$为$y(\mathrm{x};\mathcal{D})$在所有$\mathcal{D}$下的期望,其实就相当于运用了无穷多个样本。令第一项为

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

对$\{y(\mathrm{x}) - h(\mathrm{x})\}^2$取期望看看平均情况下的损失

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

第一项bias：表明了预测值的期望与h(x)之间的误差,可以理解为用了无穷多个样本后预测值与h(x)之间的误差,bias小说明了模型足够复杂,在样本无限时可以与h(x)相近,bias大表示模型太简单,即使有无穷多样本也无法和h(x)很相近(注意样本数),光看第一项很容易得出模型越复杂越好的结论

第二项variance：表明了在数据集$\mathcal{D}$下的预测值,与期望预测值之间的差距大不大,variance大也就是说模型很复杂,在样本无限的情况下期望预测可能很好,但是在数据集$\mathcal{D}$下与期望之间的波动很大,可能偏离期望预测很远,variance小说明模型相对简单,任意数据集$\mathcal{D}$都和期望预测很接近。

因此bias和variance是矛盾的,模型简单时bias大,variance小,模型复杂时bias小,variance大,所以在数据集有限的情况下,问题的关键就是在bias和variance之间权衡。

如果我们有无穷多个样本那么无疑越复杂的样本越好,因为$y(\mathrm{x};\mathcal{D})$将等于$\mathbb{E}_{\mathcal{D}}[y(\mathrm{x};\mathcal{D})]$,第二项将会为0。第一项也会为0.但通常我们的样本是有限的,所以期望损失为

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}$$

Where

$$\begin{aligned} (\text{bias})^2 &= \int \{\mathbb{E}_{\mathcal{D}}[y(\mathrm{x};\mathcal{D})] - h(\mathrm{x})\}^2 p(\mathrm{x})d\mathrm{x} \\ \text{variance} &= \int \mathbb{E}_{\mathcal{D}}[\{y(\mathrm{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathrm{x};\mathcal{D})]\}^2] p(\mathrm{x})d\mathrm{x} \\ \text{noise} &= \int\int \{h(\mathrm{x}) - t\}^2 p(\mathrm{x}, t)d\mathrm{x}dt \end{aligned}$$

回过头来看λ,讨论λ大小问题
underfitting：模型复杂度不够描述这个事务
overfitting:当前数据不够支撑这么复杂的模型

3.4 从Bayes角度看待参数选择

我们前面讲过Bayes概率由一个先验概率和一个样本最大似然组成后的得到一个后验概率,现在我们就假设参数w满足Gauss分布,因为$p(t|w)$为Gauss分布,所以这里我们为了计算方便,取他的共轭分布作为先验

$$p(\mathrm{w}) = \mathcal{N}(\mathrm{w}|\mathrm{m}_0, \mathrm{S}_0)$$

m_0:平均数,S_0:协方差
利用公式我们计算$p(w|t)$的概率

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$

Where

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{t})$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi^T\Phi}$$

这里我们考虑一个均值为0,协方差为$\alpha^{-1}\mathbf{I}$prior distribution

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

后验分布中的参数计算出来为

$$\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{\Phi}^T\mathbf{t}$$
$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi^T\Phi}$$

对$p(w|t)$取对数

$$\begin{aligned}\ln p(\mathbf{w}|\mathbf{t}) =& \ln p(\mathbf{t}|\mathbf{w}) + \ln p(\mathbf{w}|\alpha)\\ =& \frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^T\phi(\mathbf{x}_n)\}^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + const\end{aligned}$$

正则项就相当于引入了一个先验Gauss分布(这个先验分布表明w在0附近的概率最大,w很大时概率很小),再求最大概率,$\lambda = \alpha/\beta$
图第一列为$p(t|x, \mathbf{w})$,第二列为$p(w|t)$,第三列为从后验分布中取一些w_0, w_1所画的线

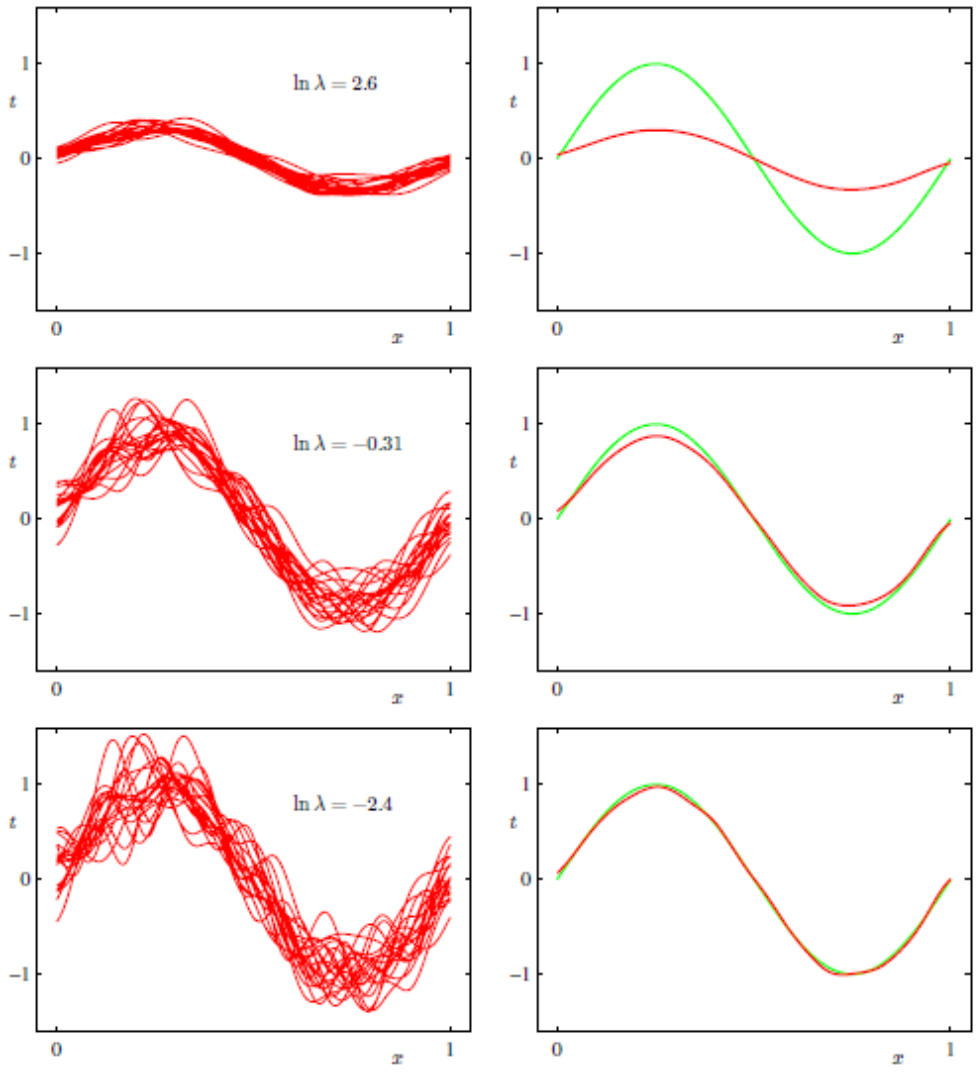

图 3: This is a vector's projection

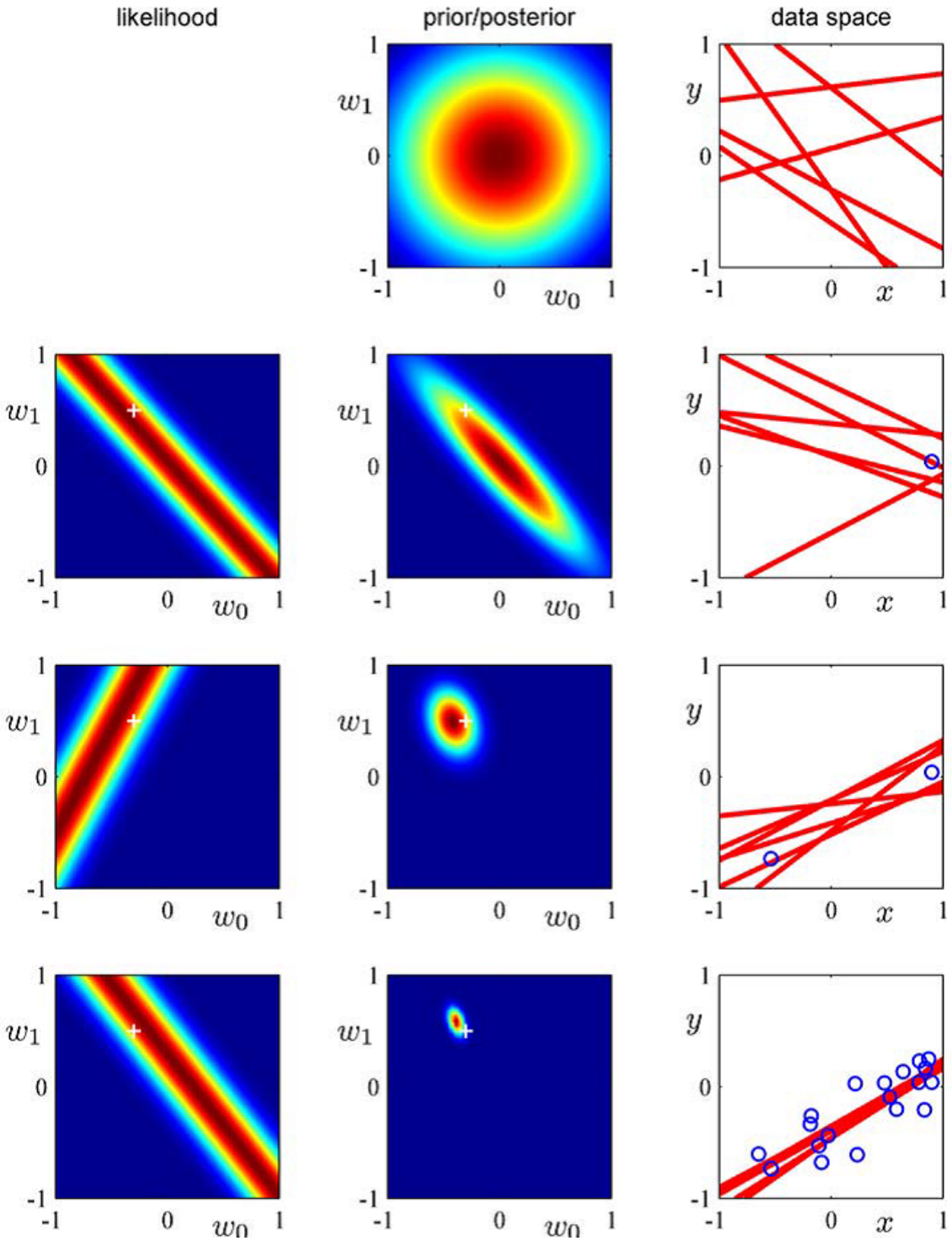

图 4: This is a vector's projection

3.4.1 预测分布

从Bayes的角度我们得到了w的分布,而我们真正关心的是对t的预测,我们可以写出t的预测分布

参考文献

[1] Michael Jordan, Jon Kleinberg, Bernhard Scholkopf ,PRML chap.1,2,3 ,2006