深入理解线性回归算法（二）：正则项的详细分析

原创 石头 机器学习算法那些事 2018-10-27

前言

当模型的复杂度达到一定程度时，则模型处于过拟合状态，类似这种意思相信大家看到个很多次了，本文首先讨论了怎么去理解复杂度这一概念，然后回顾贝叶斯思想（原谅我有点啰嗦），并从贝叶斯的角度去理解正则项的含义以及正则项降低模型复杂度的方法，最后总结全文。

目录

1、怎么去理解复杂度

怎么去理解复杂度，可能有人认为模型的参数越多，模型越复杂。笔者认为最好是通过结果去理解复杂度，比如当模型训练误差很小且测试误差很大时，则模型的复杂度较高，降低复杂度的方法包括减少模型参数的个数和降低模型参数值的大小等。笔者引用《Pattern Recognition and Machine Learning》的相关内容去阐述复杂度，希望能够加深大家对复杂度的理解。

1、方差理解复杂度

当模型的复杂度较高时，模型对训练数据集非常敏感，符合同一分布的不同训练数据集构建的模型相差很大，即**方差越高，模型的复杂度越大。**

方差定义：

$$\text{variance} = \int \mathbb{E}_{\mathcal{D}}\left[\{y(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[y(\mathbf{x};\mathcal{D})]\}^2\right] p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

其中，x表示抽样的测试数据，D为抽样的训练数据集，y(x;D)表示输入变量x在特定训练数据集D构建的模型的输出，不同的训练数据集D有不同的输出变量。

用样本的统计量来表示方差，如下图：

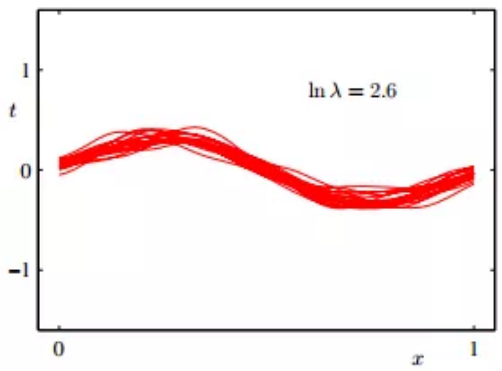
$$\text{variance} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\left\{y^{(l)}(x_n) - \overline{y}(x_n)\right\}^2$$

其中：

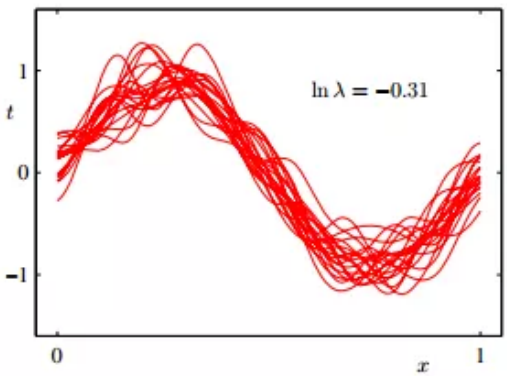
$$\overline{y}(x) = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$$

上式含义：N个测试样本在L个模型的输出方差（请参考方差公式）。

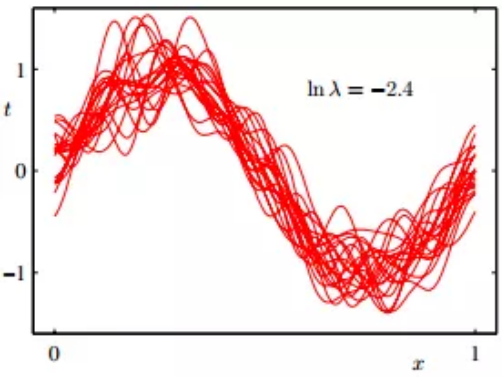
下面三张图表示了复杂度与方差之间的关系：



（1）



（2）



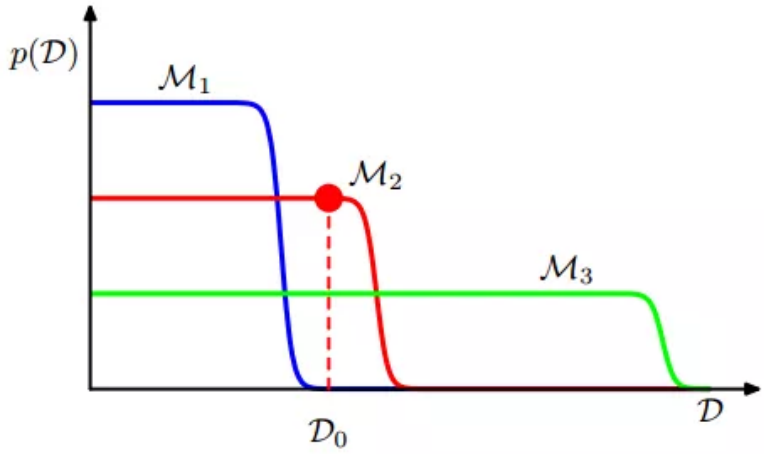
（3）

由上面三张图可知，第三张图的振动最剧烈，即方差最大，根据方差定义来理解复杂度，那么相应的复杂度也越高。

2、数据集分布理解复杂度

若模型越复杂，那么从该模型抽样的数据集变化越大，数据集覆盖的范围也越广。

如下图数据集D在模型M1，模型M2和模型M3的分布情况：



由于数据集D在模型M3分布的范围最广，则模型M3的复杂度越高，M2次之，M1最低。

2、回顾贝叶斯思想

贝叶斯思想是根据当前的观测数据再加上自己的先验知识主观判断事件发生的概率。因此随着观测数据的增加，事件发生的概率会相应的发生改变，同时先验知识是影响主观判断事件发生概率的另一个重要因素。

贝叶斯评估模型参数w分布的公式：

$$P(\mathrm{w} \mid D) = \frac{P(w)P(D \mid w)}{P(D)}$$

$$P(\mathrm{w} \mid D) = \frac{P(D \mid w)P(w)}{\int P(D \mid w)P(w)dw}$$

$\because P(D)$是标准化项

$\therefore P(\mathrm{w} \mid D) \propto P(D \mid w)P(w)$

$P(w)$是参数w的先验分布，$P(D \mid w)$是数据集D的似然函数

由该式可知，后验分布$P(\mathrm{w} \mid D)$由先验分布和似然函数决定

3、贝叶斯角度下的正则项

若模型的复杂度较高，那么通过在损失函数项增加正则项的方式来降低模型的复杂度。
如下图：

$$\widetilde{E}(w)=\frac{1}{2}\sum_{n=1}^{N}(t_n-\vec{w}^{T}\vec{\phi}(x_n))^2+\frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q \qquad (3.1)$$

（1）若q=1时，则正则化项为L1范数，构建的线性回归称LASSO回归。
（2）若q=2时，则正则化项为L2范数，构建的线性回归称Ridge回归。

最小化损失函数$\widetilde{E}(\mathbf{w})$得到的参数w即是模型的最优解。

贝叶斯角度分析损失函数

1、先验分布是高斯分布

由上节可知，贝叶斯估计模型参数w的分布需要知道参数的先验分布和数据集的似然函数，若数据集D已知，参数w的先验分布是均值为0精度为α的高斯分布。

则参数w的后验分布的推导过程如下：

数据集D包含N个高斯分布的样本数据，精度为β
输入数据为$\vec{\mathrm{x}}=(x_1,x_2,\cdots x_N)^T$，输出数据为$\vec{\mathrm{t}}=(\mathrm{t}_1,t_2,\cdots t_N)^T$
假设参数$\vec{\mathrm{w}}$的先验分布是均值为0，精度为α的高斯分布。
后验概率最大时对应的参数是最优模型参数$\vec{w}^{*}$

$$即\vec{w}^{*} = \arg\max_{w} P(w \mid D) = \arg\max_{w} \frac{P(D \mid w) * P(w)}{P(D)}$$

$$= \arg\max_{w} (P(D \mid w) * P(w))$$

取对数

$$\vec{w}^{*} = \arg\max_{w} (\ln(P(D \mid w)) + \ln P(w))$$

$$由 P(D \mid w) = \prod_{n=1}^{N} P(t_n \mid \vec{w}^{T} \overrightarrow{\phi(x_n)}, \beta^{-1}),\quad P(w) = N(w \mid 0, \alpha^{-1})$$

$$得: \vec{w}^{*} = \arg\min_{w} (-\frac{\beta}{2} \sum_{n=1}^{N} (t_n - \vec{w}^{T} \overrightarrow{\phi(x_n)})^2 - \frac{\alpha}{2} \vec{w}^{T} \vec{w} + const)$$

$$即后验概率分布 P(\mathrm{w} \mid D) = -\frac{\beta}{2} \sum_{n=1}^{N} (t_n - \vec{w}^{T} \overrightarrow{\phi(x_n)})^2 - \frac{\alpha}{2} \vec{w}^{T} \vec{w} + const$$

等价于：

$$P(\mathrm{w} \mid D) = -\frac{1}{2} \sum_{n=1}^{N} (t_n - \vec{w}^{T} \overrightarrow{\phi(x_n)})^2 - \frac{\alpha}{2\beta} \vec{w}^{T} \vec{w} + const$$

$$令 \frac{\alpha}{\beta} = \lambda$$

$$P(\mathrm{w} \mid D) = -\frac{1}{2} \sum_{n=1}^{N} (t_n - \vec{w}^{T} \overrightarrow{\phi(x_n)})^2 - \frac{\lambda}{2} \vec{w}^{T} \vec{w} + const \quad (3.2)$$

由3.2可知，**后验概率最大化等于包含L2正则化项损失函数的最小化。**

式（3.1）第一项表示损失函数，第二项表示惩罚函数。

式（3.2）第一项表示数据D的似然函数，第二项表示参数的先验分布。

比较两式可知，参数的先验分布对应于正则化项。当参数的先验分布为高斯分布时，则正则化项为L2范数，构建的回归模型称为Ridge回归。

2、先验分布是拉普拉斯分布

推导过程类似，这里只给出结论部分。

当参数的先验分布为拉普拉斯分布，则正则化项为L1范数，构建的回归模型称为LASSO回归。

小结：贝叶斯定理的后验分布与似然函数和先验分布相关，不考虑先验分布时，则损失函数不包含正则化；考虑先验分布时，则损失函数包含正则化；最大化后验分布等同于最小化正则化的损失函数。

正则项降低模型复杂度的方法

降低模型复杂度的方法主要包括减少模型参数的个数和降低模型参数的值。本节介绍正则项降低模型复杂度的方法。

$$\widetilde{E}(w)=\frac{1}{2}\sum_{n=1}^{N}(t_n-\vec{w}^T\overrightarrow{\phi(x_n)})^2+\frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

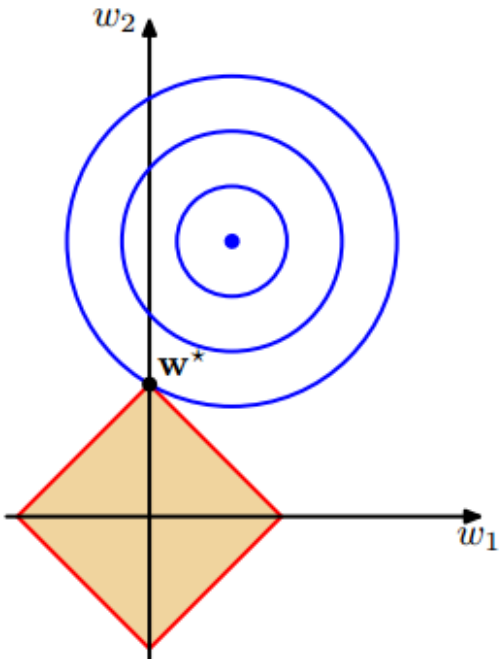
最小化 $\widetilde{E}(w)$ 等价于:

$$(E(w))_{\min}=\left\{\frac{1}{2}\sum_{n=1}^{N}(t_n-\vec{w}^T\overrightarrow{\phi(x_n)})^2\right\}\min \quad (1)$$

$$\sum_{j=1}^{M}|w_j|^q\le\eta \quad (2)$$

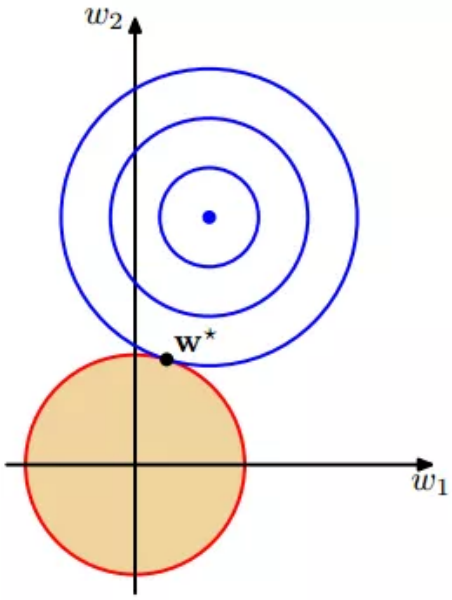
即在式（2）的条件下，求（1）的最小值，L1和L2正则项都是利用这种思想来求最优参数。

1、正则项是L1范数



如上图含L1正则项的损失函数，蓝色线为损失函数，红色线为L1正则项包含的区域。**当处于交点** $\mathbf{w}^*$ **时，含正则项的损失函数最小。由图可知该交点的w2为0，则模型参数个数较少了，相应的模型复杂度降低了。**

2、正则项是L2范数



分析方法与L1类似，该交点所处的坐标为w1较小，即改变了模型参数值的大小，复杂度也相应的降低。

5、总结

本文首先介绍怎么去理解复杂度的概念，然后从贝叶斯角度去分析正则项的含义，即正则项等同于贝叶斯分析的先验分布，最后介绍了正则项降低模型复杂度的两种方法。

参考：

Christopher M.Bishop <<Pattern Reconition and Machine Learning>>

推荐阅读文章

深入理解线性回归算法（一）

线性回归：不能忽视的三个问题

浅谈频率学派和贝叶斯学派

浅谈先验分布和后验分布

模型优化的风向标：偏差与方差

-END-