MLAPP（Machine Learning A Probabilistic Perspective）网络读书会第二次活动

第 7 章 Linear regression

主讲人 红烧鱼

（新浪微博 @红烧鱼_机器学习）

QQ 群 177217565

读书会微信公众平台请扫描下面的二维码

红烧鱼(403774317) 20:02:08

开始吧

今天讲第 7 章，线性回归

7.1 Introduction

Linear regression is the "work horse" of statistics and (supervised) machine learning. When augmented with kernels or other forms of basis function expansion, it can model also non-linear relationships. And when the Gaussian output is replaced with a Bernoulli or multinoulli distribution, it can be used for classification, as we will see below. So it pays to study this model in detail.

这段话有三层意思

1. 在监督机器学习中的地位；2. 可以建模非线性关系；3. 如何转化为分类问题。

2. 如何建模非线性关系？kernels or other forms of basis function expansion

3. 如何转化为分类问题？gaussian output -> Bernoulli/Multinoulli output

7.2 模型定义

$$p(y|\mathbf{x},\boldsymbol{\theta}) = \mathcal{N}(y|\mathbf{w}^T\mathbf{x},\sigma^2)$$

$$p(y|\mathbf{x},\boldsymbol{\theta}) = \mathcal{N}(y|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}),\sigma^2)$$

下面的公式通过 phi（x）建模非线性关系

一个例子：

$$\phi(x) = [1, x, x^2, \ldots, x^d]$$

7.3 如何求解 - 最大似然估计

$$\hat{\boldsymbol{\theta}} \triangleq \arg\max_{\boldsymbol{\theta}} \log p(\mathcal{D}|\boldsymbol{\theta})$$

D 是 evidence/obeserved data

求解方法一般是求导

展开公式

$$\mathrm{NLL}(\mathbf{w}) = \frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w}) = \frac{1}{2}\mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w} - \mathbf{w}^T(\mathbf{X}^T\mathbf{y})$$

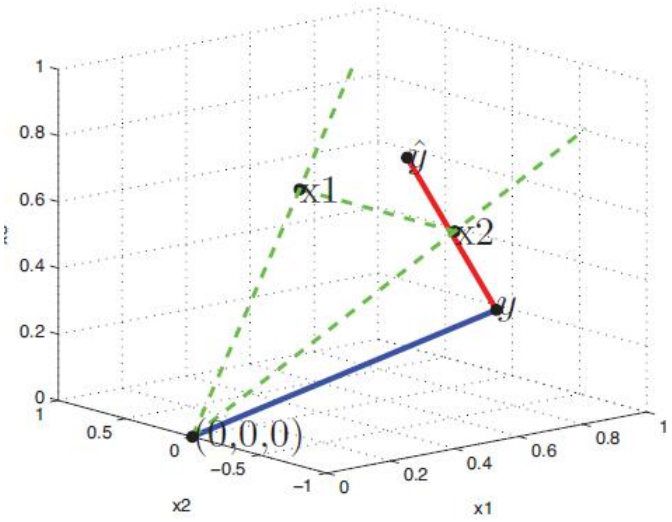
求导

$$\mathbf{g}(\mathbf{w}) = [\mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{y}] = \sum_{i=1}^{N}\mathbf{x}_i(\mathbf{w}^T\mathbf{x}_i - y_i)$$

导数为 0，得

$$\hat{\mathbf{w}}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

7.3.2 物理意义



最小化残差(rss/sse/mse),相当于最小化那条红线

7.3.3 凸性

我们之前列的公式中，函数具有唯一的最小值。

类似的函数叫凸函数。

斯坦福大学有专门一门课程讲凸函数，boyd。

接下来，定义了什么是凸函数。

$$\lambda\theta + (1-\lambda)\theta' \in \mathcal{S}, \quad \forall\, \lambda \in [0,1]$$

公式定义

(a) (b)

几何定义

左边凸，右边非凸。

怎么看出来的呢，结合公式

闭区域内，任选两点，连直线，如果直线整个在闭区域内，那么凸。

这也是公式的几何意义

纽约熊光 MaPhyCSer(939268445) 20:16:27

theta 是多维变量吧？

红烧鱼(403774317) 20:18:05

说是多维变量不太合适

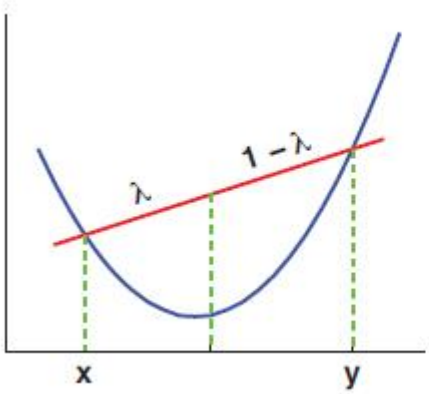
是 set 中的一个

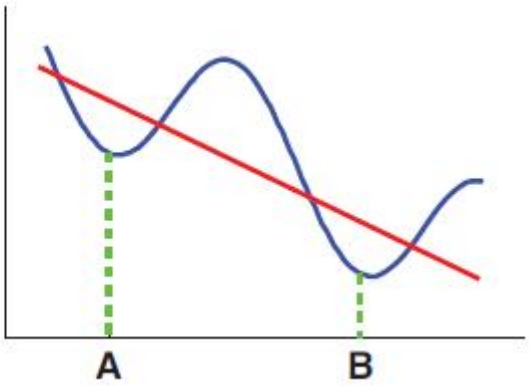
接着给出了凸函数的定义

从凸集->凸函数
定义类似

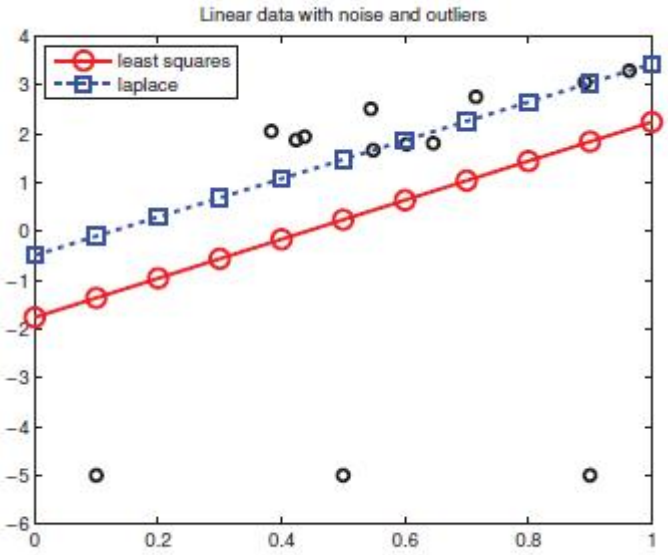
$$f(\lambda\theta + (1-\lambda)\theta') \leq \lambda f(\theta) + (1-\lambda) f(\theta')$$

几何表现就是





注意：不同的书对凹凸性的定义不一样
左边这个图，有些书定义为凹，有些定义为凸。
不必那么较真。
秦淮/sun 人家(76961223) 20:21:08
一般定义左边那个是凸吧
红烧鱼(403774317) 20:21:37
凹凸函数最重要的性质：具有唯一的极值点
gump(915537522) 20:21:43
左边是凹吧
国外的定义好像和国内相反。
红烧鱼(403774317) 20:21:57
所以，不必纠结。不同的书定义不一样。
秦淮/sun 人家(76961223) 20:22:34
说凸是从上往下看，因为有一个上镜图的概念，上镜图是凸的。
红烧鱼(403774317) 20:22:38
这个性质非常重要。
秦淮/sun 人家(76961223) 20:22:41
鱼哥，继续吧
旅程(920995583) 20:22:46
反正有极值点就对了。。凸凹什么的不用太较真
红烧鱼(403774317) 20:24:48
7.4 鲁棒的线性回归
问题：对 outlier 敏感



(a)

怎么解决：

One way to achieve **robustness** to outliers is to replace the Gaussian distribution for the response variable with a distribution that has **heavy tails**. Such a distribution will assign higher

为什么会对 outlier 敏感：

Figure 7.6(a). (The outliers are the points on the bottom of the figure.) This is because squared error penalizes deviations quadratically, so points far from the line have more affect on the fit than points near to the line.

大家还记得刚才的梯度/导数公式吗

里面是二次方

所以越远的影响越大，这也是这句话的意思

接下来是非常重要的岭回归

7.5 ridge regression

针对问题：最大似然估计容易导致 overfitting

亘古不变的话题

一种解决方法：高斯先验的 MAP 估计

为什么说容易 overfit 呢？因为很容易求出来这样的结果

红烧鱼(403774317) 20:30:15

```
6.560, -36.934, -109.255, 543.452, 1022.561, -3046.224, -3768.013,
8524.540, 6607.897, -12640.058, -5530.188, 9479.730, 1774.639, -2821.526
```

.(1208227795) 20:31:59

为什么回归要用 L2norm 来衡量好坏，不采用 L1 或者 L0 呢

红烧鱼(403774317) 20:32:00

显然，12640.058 的系数相比 6.560 系数，对结果的影响要大得多。

不是衡量好坏

l2 解决过拟合，l0 是为了解决稀疏问题，但是 np 难，所以折中 l1

下一章还是下下章会讲 lars,lasso

雪落冬夜(693156402) 20:33:27

这个 ridge 由来是什么啊
.(1208227795) 20:34:32

$$\mathrm{NLL}(\mathbf{w}) \quad = \quad \frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})^T(\mathbf{y}-\mathbf{X}\mathbf{w}) = \frac{1}{2}\mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w} - \mathbf{w}^T(\mathbf{X}^T\mathbf{y})$$

不是 l2 么~~为什么 l2 解决过拟合呢，嘿嘿
红烧鱼(403774317) 20:34:49
好问题
为什么 l2 能解决过拟合呢？
秦淮/sun 人家(76961223) 20:35:14
最早的起源不是为了放过拟合，貌似是为了让 X^T X 可逆
红烧鱼(403774317) 20:35:15
说的不严谨，不是解决
是 amaliorate
是的，是针对 inverse 问题提出
.(1208227795) 20:36:37
还是不明白跟过拟合有什么关系呀~~
红烧鱼(403774317) 20:36:56
因为加了惩罚

$$\hat{\mathbf{w}}_{ridge} = (\lambda \mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

好奇心害死薛定谔的猫(337025583) 20:37:11
岭回归讲完了？
就是加了 L2 ？
红烧鱼(403774317) 20:37:30
这里面加的 lambda
秦淮/sun 人家(76961223) 20:37:41
形象点说，如果 w 很大，那么线容易起伏很大，起伏越大拟合训练集越容易
红烧鱼(403774317) 20:37:59
可以这么说，这个时候应该上 prml 上那张经典的图。。
秦淮/sun 人家(76961223) 20:38:12

DarkScope(530138084) 20:38:17
我们得到的数据是有测量误差的，参数为 x，要优化的为:||Ax-y||，其实是
||(A'+delta)x-y||=||A'x-y+delta*x|| ，x 越小，delta 对模型的影响越小
木交示申(1300510320) 20:38:23
最早的起源是为了把 ill-posed 问题转化为 well-posed 问题
正则项相当于光滑先验 光滑导致连续 就是为了让 w 变化小的时候 y 不会变化很剧烈
红烧鱼(403774317) 20:38:44

黄浩军<littleblack1988@foxmail.com> 20:39:01
在限定条件下 L2 等同于数据加入了高斯噪音 使得模型不太能受到噪音的影响
prml 书中有证明
红烧鱼(403774317) 20:39:26
浩军来了
秦淮/sun 人家(76961223) 20:39:33
不是高斯先验么？高斯噪音由平凡损失导出
雪落冬夜(693156402) 20:39:35
是好像有这个一说，贝叶斯学派的观点就是正则相当于先验的约束
黄浩军<littleblack1988@foxmail.com> 20:39:47
要在限定的条件下
FreeMind(409331172) 20:39:51
对参数加了个高斯先验可以叫加入高斯噪音么？
黄浩军<littleblack1988@foxmail.com> 20:40:05
我说法可能有问题
.(1208227795) 20:40:19
先验是什么意思啊
黄浩军<littleblack1988@foxmail.com> 20:40:41
翻下书先
秦淮/sun 人家(76961223) 20:41:43
这么说吧，开始我们对 w 没有任何知识，那么假定他接近 0，如果我们确定他很接近 0，那么我们应该让他离 0 远的概率小。怎么控制？高斯先验，0 为均值，小方差让他离 0 远的概率小
雪落冬夜(693156402) 20:42:19
对
尘绳聋(523201603) 20:43:09
bayesian 认为 w 是未知的，服从某个概率分布，然后就假定是高斯，然后 MLE，先验对应的就是 L2 了。
尘绳聋(523201603) 20:43:30
l1 对应的是 laplace？
雪落冬夜(693156402) 20:44:14

Gaussian distribution of the form

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\} \tag{1.65}$$

ss(912111026) 20:44:31
“对 w 没有任何知识，那么假定他接近 0。” 这个 0 是随意假设的吗？
雪落冬夜(693156402) 20:44:33

maximum posterior, or simply *MAP*. Taking the negative logarithm of (1.66) and combining with (1.62) and (1.65), we find that the maximum of the posterior is given by the minimum of

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}. \tag{1.67}$$

尘绳聋(523201603) 20:44:45
恩。。MAP
秦淮/sun 人家(76961223) 20:45:24
@ss 这也是我的疑惑。
旅程(920995583) 20:45:25
L2 是二范数的意思？
秦淮/sun 人家(76961223) 20:45:32
难道因为数据都标准化过了么
雪落冬夜(693156402) 20:47:56
估计是因为 w 均值多少，只影响目标函数值的大小
秦淮/sun 人家(76961223) 20:48:40
一般来说，我们做回归的时候，数据和目标值都标准化了，均值都是 0
尘绳聋(523201603) 20:49:12
数据的均值跟 w 的均值怎么联系起来的。。。？
没搞懂
秦淮/sun 人家(76961223) 20:50:18
怎么样让数据的均值 0 乘上 w，然后目标均值也是 0 呢，只能 w 也为 0
尘绳聋(523201603) 20:50:47
我不做标准化也可以 linear regression 的啊，如果你是用 normal equation 的话。标准化只不过是为了
然 gradien descent 更快收敛，提高数值稳定性
尘绳聋(523201603) 20:53:31

$$\hat{\mathbf{w}}_{ridge} = (\lambda \mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

就是这个，这就不用标准化的，均值，方
差的都不用管
TK(384880403) 20:54:03
这不 ridge regression 么。。。
也需要标准化……
秦淮/sun 人家(76961223) 20:54:14
不过我盲目地认为，如果数据不标准化，对回归有影响，假如有一维数据只会从 1000 开始，实际上 1500
应该是 2000 效果的 1/2 而不是 3/4
这些 X 一般都是默认标准化过的，不过不标准化确实可以回归。
TK(384880403) 20:54:59
当然可以回归，效果不好而已。。
秦淮/sun 人家(76961223) 20:55:13
对。
尘绳聋(523201603) 20:55:34
@TK 用 Normal equation 也要？
.(1208227795) 20:55:37
那一半 ridge regression 也要先把数据归一化到 N(0,1)的形式吗

.(1208227795) 20:55:41
一般
尘绳聋(523201603) 20:56:17
我记得直接 normal equation 求解的话，是可以不用的。如果有 gd 就要
TK(384880403) 20:57:46

A justification for choosing this criterion is
by solving the **normal equations**

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}.$$

TK(384880403) 20:57:48
指的是这个么？
尘绳聋(523201603) 20:58:14

$$\hat{\mathbf{w}}_{ridge} = (\lambda \mathbf{I}_D + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

是的，对应 ridge 就是
TK(384880403) 20:58:57
这个虽然说看起来好像求个解就行了，没啥影响，但是 matlab 之类的内置的包很多也用迭代方法求解
尘绳聋(523201603) 20:59:43
哈！ 没想到这个
后面再翻翻资料看看 @红烧鱼 要不先继续吧？
黄浩军<littleblack1988@foxmail.com> 21:00:25
这个问题 可以等下再讨论
TK(384880403) 21:00:32
像 logistic regression + L2 之类的就更不用说了
红烧鱼(403774317) 21:00:35
ok，看大家讨论的学到很多呀
7.5.2 数字计算稳定问题
为了数值稳定性，常常避免求幂。这里为了 fitting 模型，用了一个 trick

$$\tilde{\mathbf{X}} = \begin{pmatrix} \mathbf{X}/\sigma \\ \sqrt{\boldsymbol{\Lambda}} \end{pmatrix}, \quad \tilde{\mathbf{y}} = \begin{pmatrix} \mathbf{y}/\sigma \\ \mathbf{0}_{D\times 1} \end{pmatrix}$$

第一步先扩展原始数据

where $\boldsymbol{\Lambda} = \sqrt{\boldsymbol{\Lambda}}\sqrt{\boldsymbol{\Lambda}}^T$ is a **Cholesky decomposition** of $\boldsymbol{\Lambda}$.

OLS estimates) that is more numerically robust. We assume the prior has the form $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda}^{-1})$, where $\boldsymbol{\Lambda}$ is the precision matrix. In the case of ridge regression, $\boldsymbol{\Lambda} = (1/\tau^2)\mathbf{I}$. To

第二步证明了，求解这个扩充后的数据等同于求解原始数据。
秦淮/sun 人家(76961223) 21:05:36
这个好高端

红烧鱼(403774317) 21:06:38

第三步，对扩充后的数据求解。通过 QR 分解后，原来的求逆问题转化为

$$\hat{\mathbf{w}}_{ridge} = \mathbf{R}^{-1}\mathbf{R}^{-T}\mathbf{R}^{T}\mathbf{Q}^{T}\tilde{\mathbf{y}} = \mathbf{R}^{-1}\mathbf{Q}\tilde{\mathbf{y}}$$

R is easy to invert since it is upper triangular

从而避免了 invert $(\mathbf{\Lambda} + \mathbf{X}^T\mathbf{X})$.

这部分完了

红烧鱼(403774317) 21:08:07

既然能 QR 分解，那么是否能 SVD 分解呢？

7.5.3 和 pca 的关系

对数据进行 svd 分解

Let $\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ be the SVD of $\mathbf{X}$. From Equation 7.44, we have

$$\hat{\mathbf{w}}_{ridge} = \mathbf{V}(\mathbf{S}^2 + \lambda\mathbf{I})^{-1}\mathbf{S}\mathbf{U}^T\mathbf{y}$$

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\hat{\mathbf{w}}_{ridge} = \mathbf{U}\mathbf{S}\mathbf{V}^T\mathbf{V}(\mathbf{S}^2 + \lambda\mathbf{I})^{-1}\mathbf{S}\mathbf{U}^T\mathbf{y} \\ &= \mathbf{U}\tilde{\mathbf{S}}\mathbf{U}^T\mathbf{y} = \sum_{j=1}^{D}\mathbf{u}_j\tilde{S}_{jj}\mathbf{u}_j^T\mathbf{y}\end{aligned}$$

这个是 svd 分解后的表达式

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}}_{ls} = (\mathbf{U}\mathbf{S}\mathbf{V}^T)(\mathbf{V}\mathbf{S}^{-1}\mathbf{U}^T\mathbf{y}) = \mathbf{U}\mathbf{U}^T\mathbf{y} = \sum_{j=1}^{D}\mathbf{u}_j\mathbf{u}_j^T\mathbf{y}$$

这个是 least squares 预测的表达式

$$\tilde{S}_{jj} \triangleq [\mathbf{S}(\mathbf{S}^2 + \lambda I)^{-1}\mathbf{S}]_{jj} = \frac{\sigma_j^2}{\sigma_j^2 + \lambda}$$

其中

这两个比较起来能有什么结论呢？

If σ_j^2 is small compared to λ, then direction $\mathbf{u}_j$ will not have much effect on the prediction.

When $\lambda = 0$, $\mathrm{dof}(\lambda) = D$, and as $\lambda \to \infty$, $\mathrm{dof}(\lambda) \to 0$.

Let us try to understand why this behavior is desirable. In Section 7.6, we show that $\mathrm{cov}[\mathbf{w}|\mathcal{D}] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$, if we use a uniform prior for $\mathbf{w}$. Thus the directions in which we are most uncertain about $\mathbf{w}$ are determined by the eigenvectors of this matrix with the smallest eigenvalues, as shown in Figure 4.1. Furthermore, in Section 12.2.3, we show that the squared singular values σ_j^2 are equal to the eigenvalues of $\mathbf{X}^T\mathbf{X}$. Hence small singular values σ_j correspond to directions with high posterior variance. It is these directions which ridge shrinks the most.

这段大家怎么看

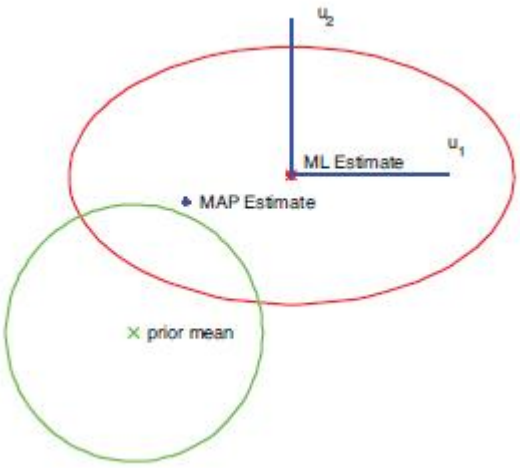
we show that the

squared singular values σ2

j are equal to the eigenvalues of XTX.

smallest eigenvalues, as shown in Figure 4.1. Furthermore, in Section 12.2.3, we show that the squared singular values σ_j^2 are equal to the eigenvalues of $\mathbf{X}^T\mathbf{X}$. Hence small singular values σ_j

squared singular values σ_j^2 are equal to the eigenvalues of $\mathbf{X}^T\mathbf{X}$. Hence small singular values σ_j correspond to directions with high posterior variance. It is these directions which ridge shrinks the most.



岭回归的几何解释

Figure 7.9 Geometry of ridge regression. The likelihood is shown as an ellipse, and the prior is shown as a circle centered on the origin. Based on Figure 3.15 of (Bishop 2006b). Figure generated by `geomRidge`

椭圆的是最大似然估计的结果

绿色的是先验

可以看出，先验平滑了 ml 的估计结果

就像刚才某位网友说的，可以看做是平滑。

接下来，引出了 shrinkage 的概念

This process is illustrated in Figure 7.9. The horizontal w_1 parameter is not-well determined by the data (has high posterior variance), but the vertical w_2 parameter is well-determined. Hence w_2^{map} is close to $\hat{w}_2^{mle}$, but w_1^{map} is shifted strongly towards the prior mean, which is 0. (Compare to Figure 4.14(c), which illustrated sensor fusion with sensors of different reliabilities.) In this way, ill-determined parameters are reduced in size towards 0. This is called **shrinkage**.

谁能帮忙解释一下？

刚才的那个图中

TK(384880403) 21:26:54

$$\frac{\sigma_j^2}{\sigma_j^2+\lambda}$$ 这个是 shrinkage 程度的话，当 sigma_j 小的时候，shrinkage 程度较高，w_j 更趋向于0？

啊，不对……

不太理解它说的 w is well-determined 是什么意思。。

红烧鱼(403774317) 21:31:46

这个概念貌似在其他地方很少见，它这里貌似只是根据(has high posterior variance)来定义

TK(384880403) 21:31:58

我理解的 ridge regression 是这样的。。

红烧鱼(403774317) 21:32:05

估计是因为容易 shift，所以 not well determined

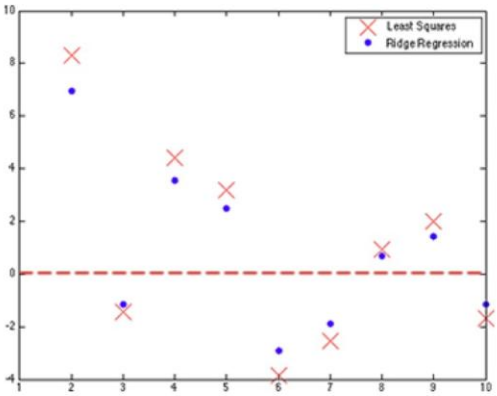
TK(384880403) 21:32:31

Singular Value Decomposition (SVD)

- $N = 100, p = 10$

$$\mathbf{X}\hat{\beta}^{ls} = u_1 \times (u_1^T y) + u_2 \times (u_2^T y) + \cdots + u_p \times (u_p^T y)$$

$$\mathbf{X}\hat{\beta}^{ridge} = u_1 \times (u_1^T y)\frac{d_1^2}{d_1^2+\lambda} + u_2 \times (u_2^T y)\frac{d_2^2}{d_2^2+\lambda} + \cdots + u_p \times (u_p^T y)\frac{d_p^2}{d_p^2+\lambda}$$



Least squares $u_1^T y$

Ridge Reression $u_p^T y \frac{d_p^2}{d_p^2+\lambda}$

$\frac{d_p^2}{d_p^2+\lambda}$ 这个是 shrinkage term。。。

红烧鱼(403774317) 21:34:28

赞上面那张图

旅程(920995583) 21:34:45

这个图从哪截取的？

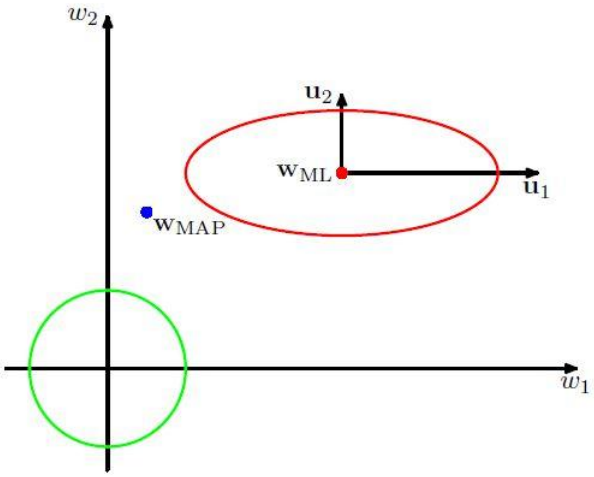
TK(384880403) 21:35:07

自己画的……

黄浩军<littleblack1988@foxmail.com> 21:35:21

这是 prml 上的

Figure 3.15 Contours of the likelihood function (red) and the prior (green) in which the axes in parameter space have been rotated to align with the eigenvectors $\mathbf{u}_i$ of the Hessian. For $\alpha = 0$, the mode of the posterior is given by the maximum likelihood solution $\mathbf{w}_{\mathrm{ML}}$, whereas for nonzero α the mode is at $\mathbf{w}_{\mathrm{MAP}} = \mathbf{m}_N$. In the direction w_1 the eigenvalue λ_1, defined by (3.87), is small compared with α and so the quantity $\lambda_1/(\lambda_1 + \alpha)$ is close to zero, and the corresponding MAP value of w_1 is also close to zero. By contrast, in the direction w_2 the eigenvalue λ_2 is large compared with α and so the quantity $\lambda_2/(\lambda_2 + \alpha)$ is close to unity, and the MAP value of w_2 is close to its maximum likelihood value.



旅程(920995583) 21:35:29

黄浩军<littleblack1988@foxmail.com> 21:35:38

可以参照理解

红烧鱼(403774317) 21:36:39

牛，这两张图在一起看就能解释了

接下来一段提到主成分回归

There is a related, but different, technique called **principal components regression**. The idea is this: first use PCA to reduce the dimensionality to K dimensions, and then use these low dimensional features as input to regression. However, this technique does not work as well as ridge in terms of predictive accuracy (Hastie et al. 2001, p70). The reason is that in PC regression, only the first K (derived) dimensions are retained, and the remaining $D - K$ dimensions are entirely ignored. By contrast, ridge regression uses a "soft" weighting of all the dimensions.

貌似没什么意思

先用 pca 降维，然后用 low dimensional features 当做 regression 的输入

TK(384880403) 21:38:19

嗯，曾经看起来觉得 PCR 这名字很高大上，结果一看就是 regression on PC……

红烧鱼(403774317) 21:39:20

为什么用 low dimensional features 呢？因为 pca 降维后，剩下一堆维度基本上没用了。。

By contrast, ridge regression uses a "soft" weighting of all the dimensions.

只有 low dimensinal 对应的比较大的特征值对应的 feature 才比较有建模能力

没感觉出来这章讲什么意思。。

还以为和 pca 等价什么的。。

7.5.4regularization

中文貌似翻译成正则化

Regularization is the most common way to avoid overfitting.

l2,l1,l0 都属于 regulatization 的范畴

这里提出了 avoid overfitting 的另外一种有效的方法

to use lots of data。。。

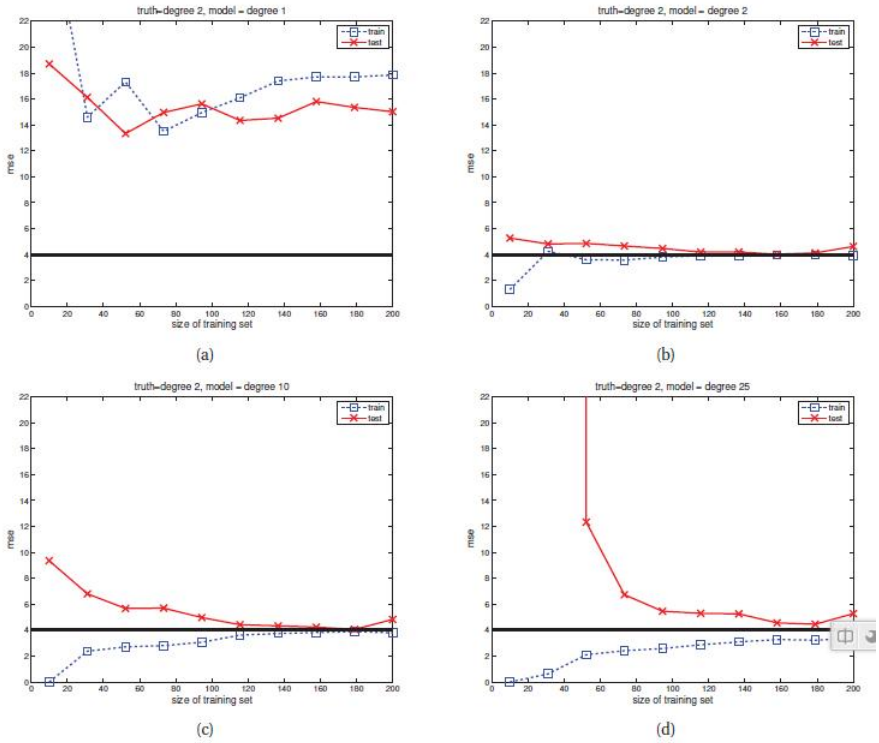
This is illustrated in Figure 7.10, where we plot the mean squared error incurred on the test set achieved by polynomial regression models of different degrees vs N (a plot of error vs training set size is known as a **learning curve**). The level of the plateau for the test error consists of two terms: an irreducible component that all models incur, due to the intrinsic variability of the generating process (this is called the **noise floor**); and a component that depends on the discrepancy between the generating process (the "truth") and the model: this is called **structural error.**

第二段提出了 structural error 的概念

结构错误包涵两部分：

an irreducible component that all models incur, due to the intrinsic variability of the generating process (this is called the noise floor); and a component that depends on the discrepancy between the generating process (the "truth") and the model:

然后给图阐述 structural error



这里面的 degree 1,2,25 都是 polynomial

We see that the structural error for models M2 and M25 is zero, since both are able to capture the true generating process. However, the structural error for M1 is substantial, which is evident from the fact that the plateau occurs high above the noise floor.

In domains with lots of data, simple methods can work surprisingly well (Halevy et al. 2009).

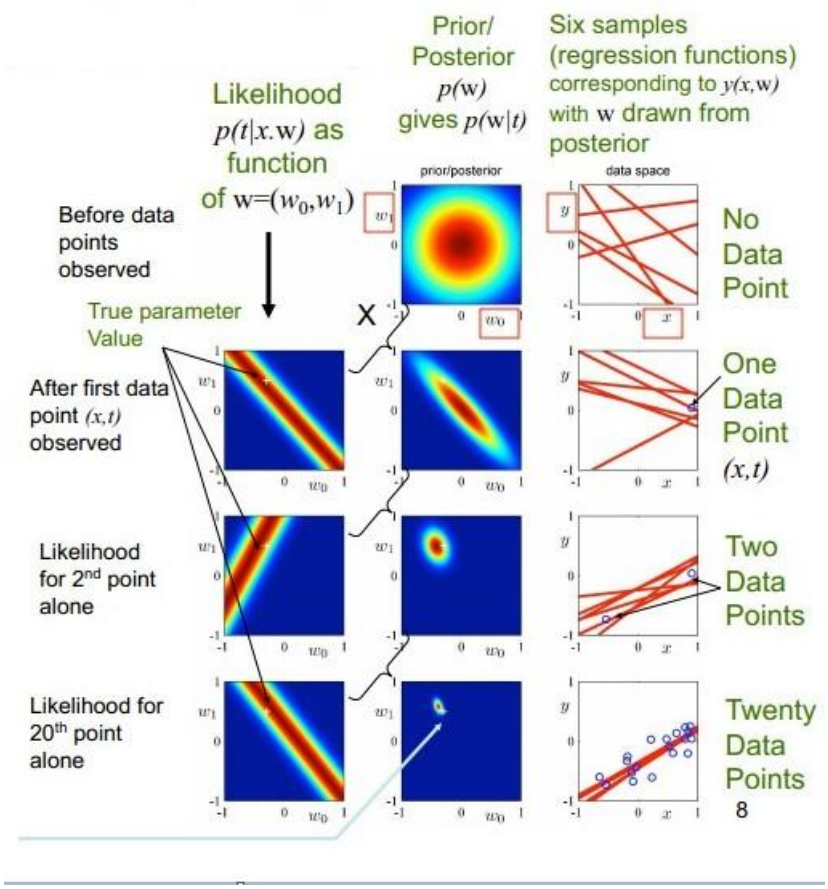
这句话前几年特别火

特别是深度学习国内火起来之前。。。

听工业界几个大佬都说过类似的话

深度学习火起来之后，没人说了。。。

刚才那张图的结论是：Note that for small training set sizes, the test
error of the degree 25 polynomial is higher than that of the degree 2 polynomial, due to overfitting, but
this difference vanishes once we have enough data. Note also that the degree 1 polynomial is too simple
and has high test error even given large amounts of training data.
小数据集上，复杂的模型容易比简单的模型过拟合。
但是当数据量足够时会 varnish
当模型过于简单时，即便给大量数据，正确率仍很高。
7.6 bayesian linear regression
Although ridge regression is a useful way to compute a point estimate, sometimes we want to compute the full posterior over w and sigma^2
TK(384880403) 22:01:56
不太了解，但是看这个意思是
因为，实际上 L1 就是给 w 一个 laplacian 先验，L2 就是给 w 一个 gaussian 先验，估计就是来了数据之后，把先验分布更新一下，得到一个后验分布吧？
红烧鱼(403774317) 22:19:55
先把这部分里面的经典图放上来



这部分下次再给大家讲吧
TK(384880403) 22:26:21

红烧鱼(403774317) 22:26:24

辛苦大家乐

Phinx(411584794) 22:26:36

下一步(289154544) 22:26:44

网络上的尼采(813394698) 22:26:45

感谢红烧鱼给大家带来的精彩讲课 

TK(384880403) 22:26:48

撒花，鼓掌

好奇心害死薛定谔的猫(337025583) 22:26:55

辛苦 

红烧鱼(403774317) 22:27:04

不敢不敢 才疏学浅，希望没有误导大家。

黄浩军<littleblack1988@foxmail.com> 22:27:15