Bias & Variance 正则化

学生创新中心：肖雄子彦

目录 / CONTENTS

01

误差来源
Bias and Variance

学习目标:

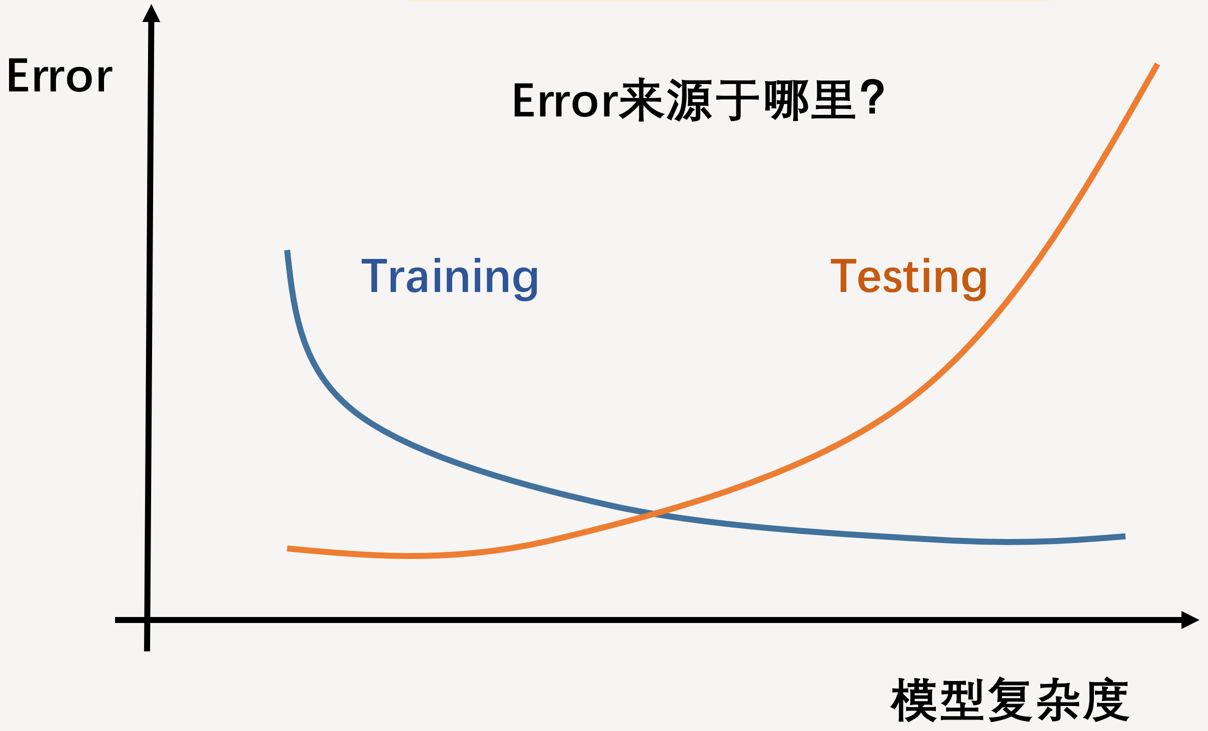
- **理解**Bias and Variance概念
- **能够判断、分析**不同网络模型的误差来源
- **熟悉**应对欠拟合、过拟合的基本改良方法
- **掌握**基础的L1与L2正则化

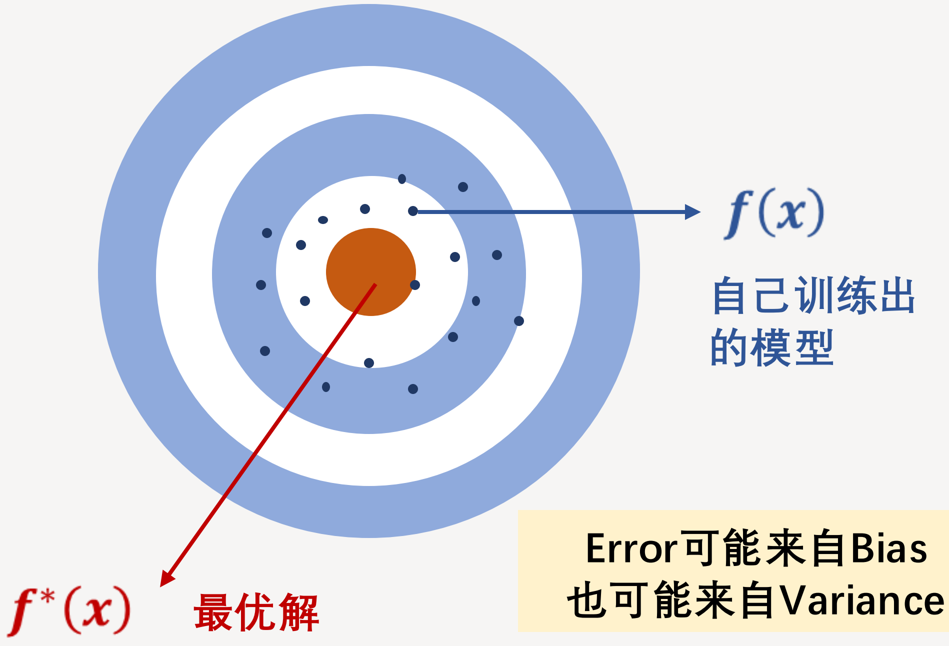
Bias and Variance

我们都知道，越复杂的模型不一定在Training data和 Testing data上都表现的好

判断来源

Bias 和 Variance → 挑选合适的方法进行改良





Error可能来自Bias
也可能来自Variance

Bias and Variance

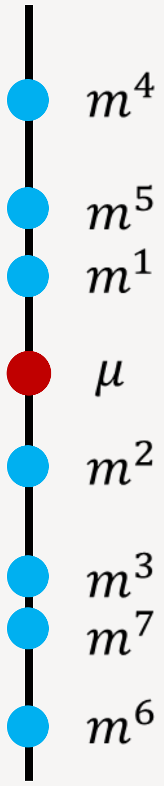
假设有一组变量 X

- 假定他们的平均值Mean是 $\boldsymbol{\mu}$
- 方差Variacne是 $\boldsymbol{\sigma}^2$

给定一组样本 $N \in \{x^1, x^2, x^3 \ldots x^n\}$

$$m = \frac{1}{N}\sum_{n} x^n$$

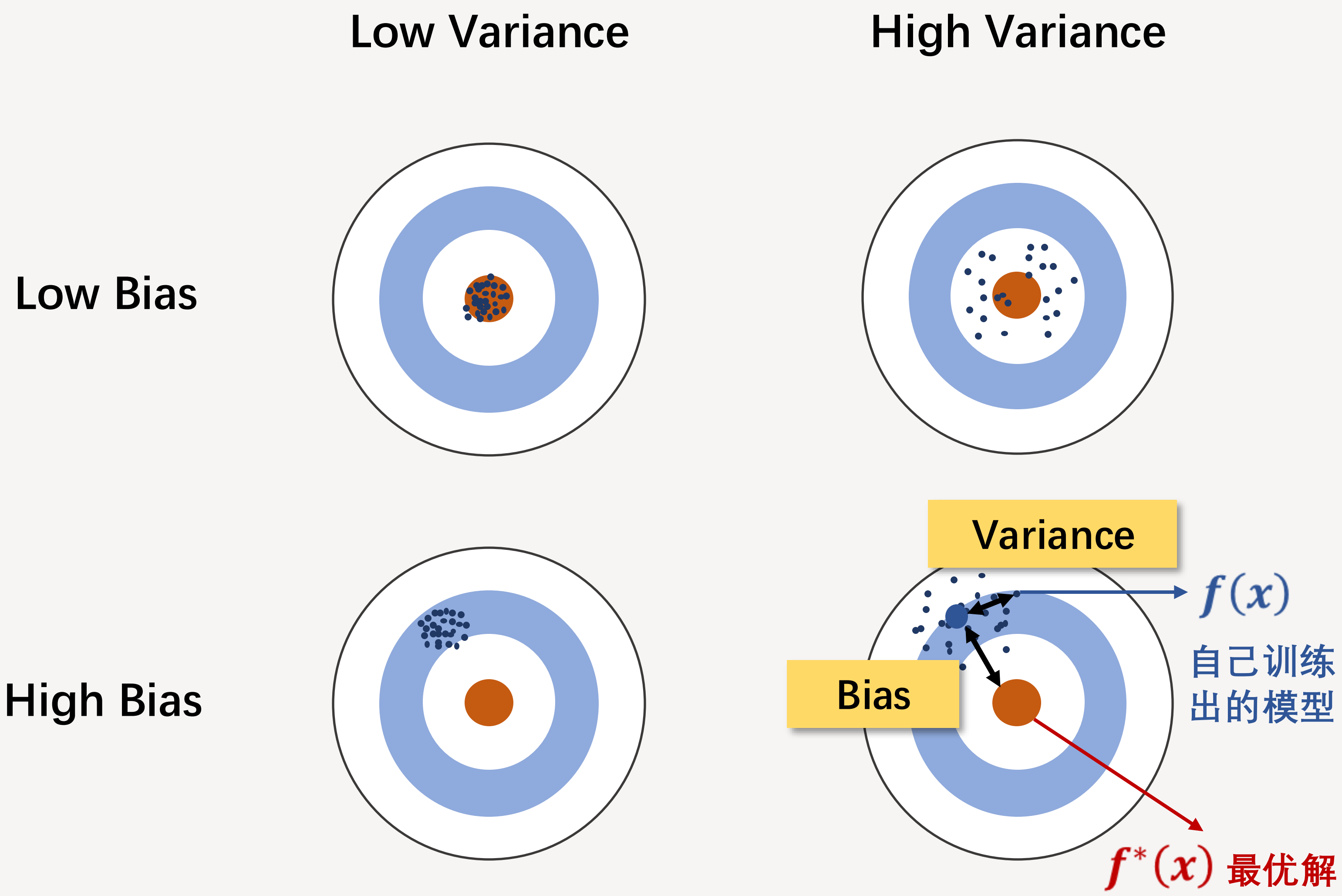
$$E(m) = E\left(\frac{1}{N}\sum_{n} x^n\right) = \frac{1}{N}E(x^n) = \mu$$



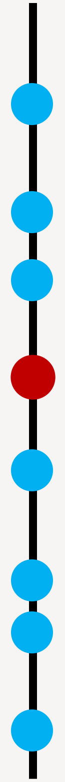
这些蓝色点，就好比我们根据不同的样本训练出来的不同的model。

每一组数据都是以某种分布取样得到的，因此会训练出来以$\boldsymbol{\mu}$为中心的各种f(x)，如果没有以靶心为中心就产生了Bias；分散的有多开由Variance决定。

Bias and Variance



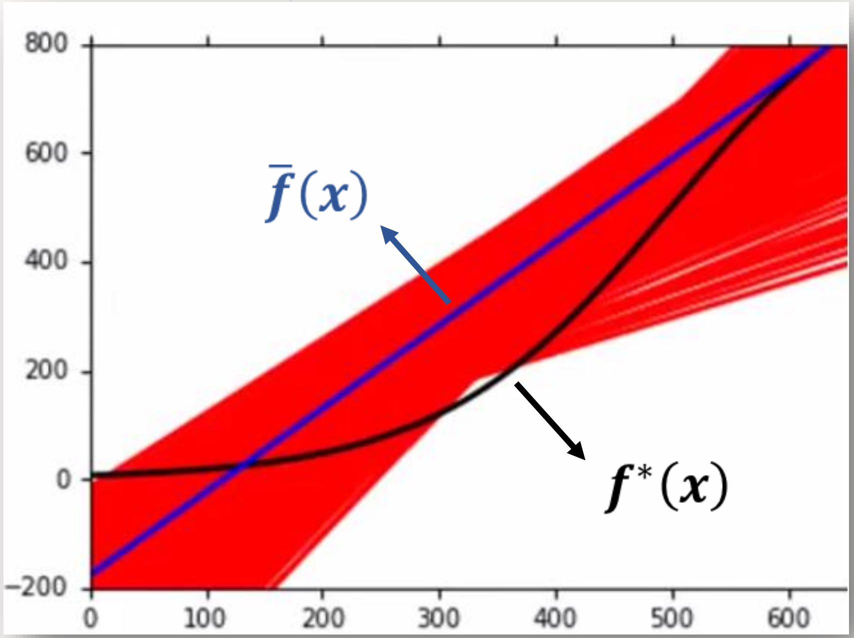
如果没有以靶心为中心就产生了Bias；分散的有多开由Variance决定。



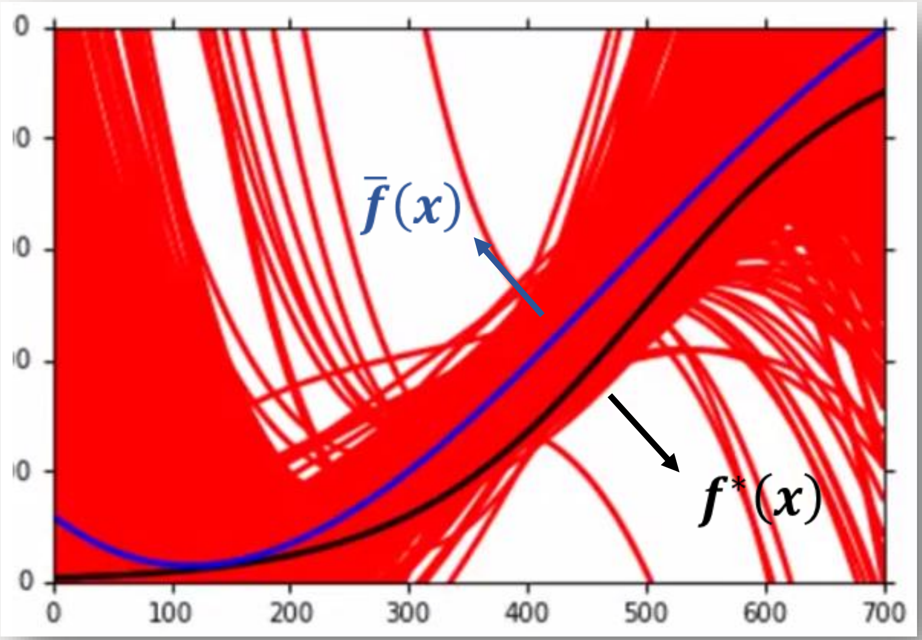
Bias and Variance

Bias 的情况

$$y = wx + b$$



$$y = w_1x_1 + w_2x_1^2 + w_3x_1^3 + w_4x_1^4 + b$$



图来源
台大李宏毅教授

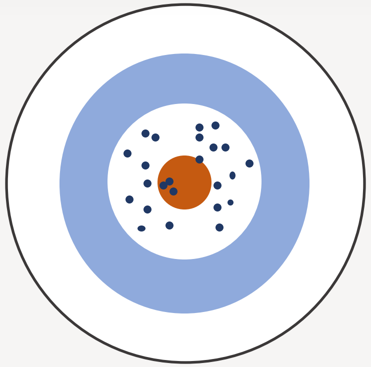
Simple model

拟合的不太好

Bias很大

Complex model

在每一笔训练集上
各自拟合较好

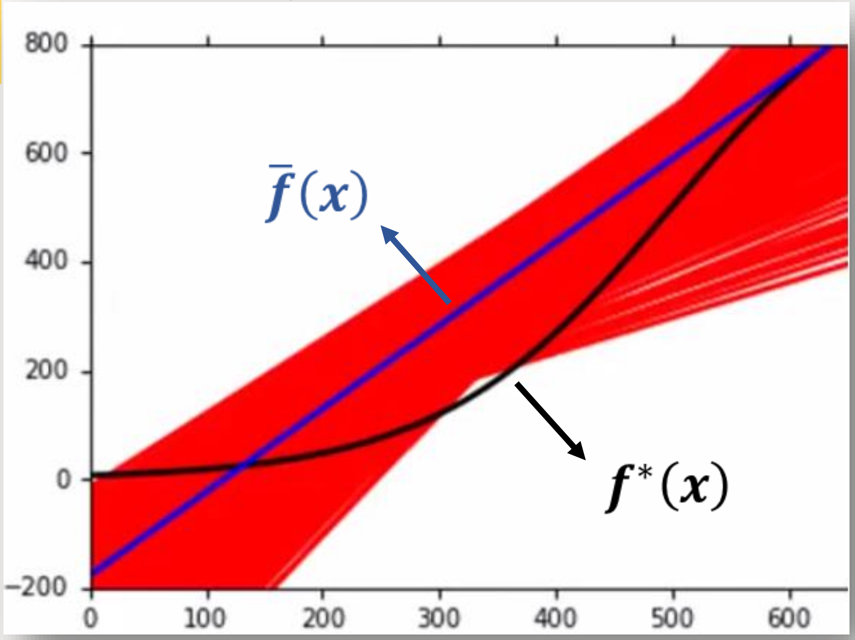


Bias比较小

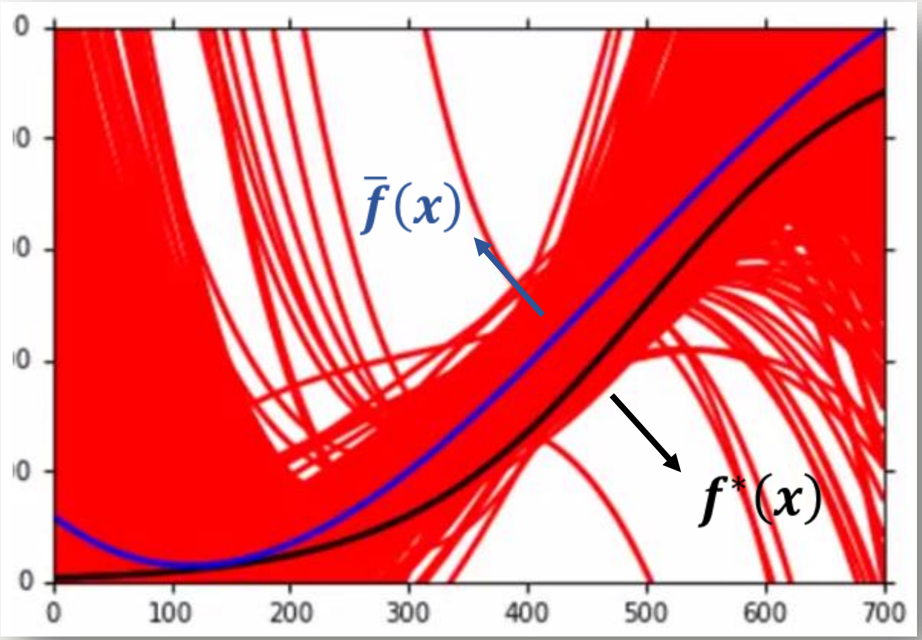
Bias and Variance

Variance
的情况

$y = wx + b$



$y = w_1x_1 + w_2x_1^2 + w_3x_1^3 + w_4x_1^4 + b$



图来源
台大李宏毅教授

Simple model

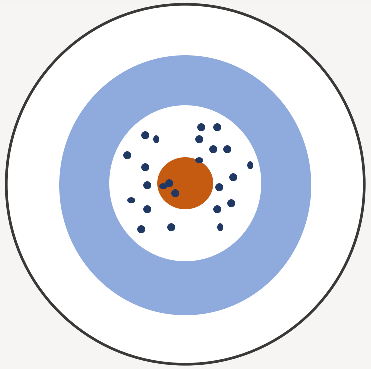
参数较少
泛化能力差不多
易欠拟合

Variance很小

Complex model

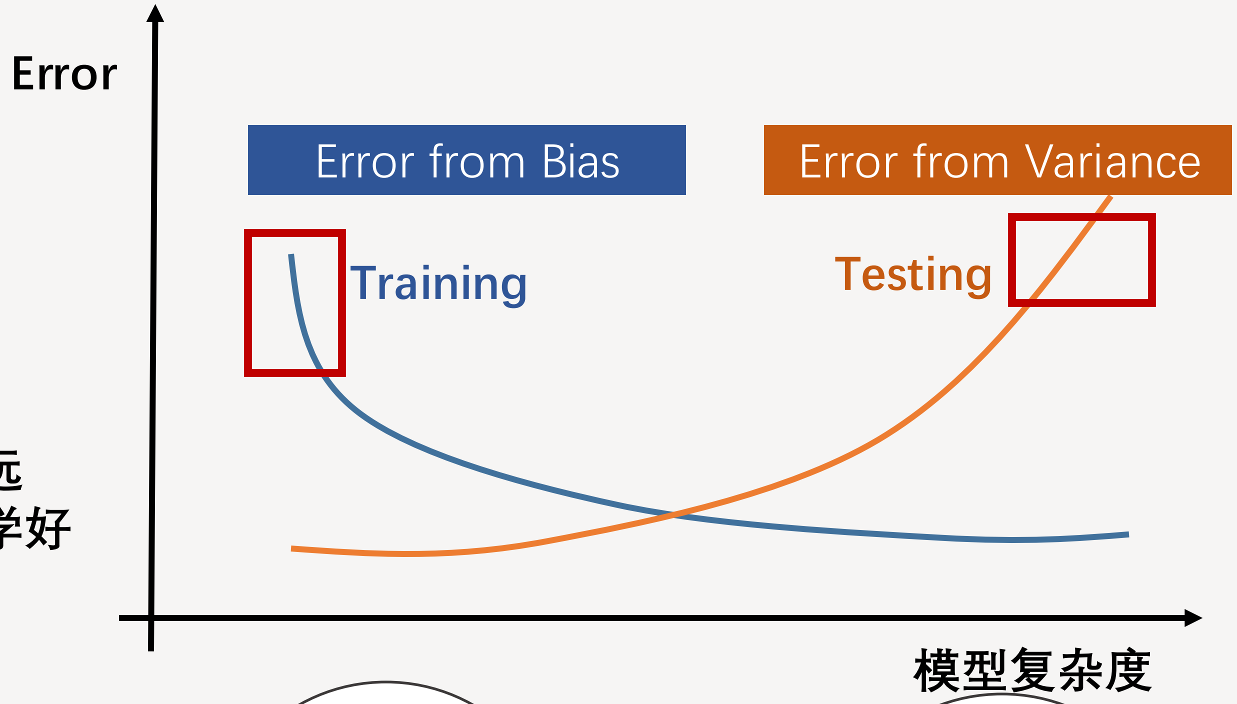
各函数拟合各自数据
易过拟合
所以泛化能力差

Variance较大



Bias and Variance

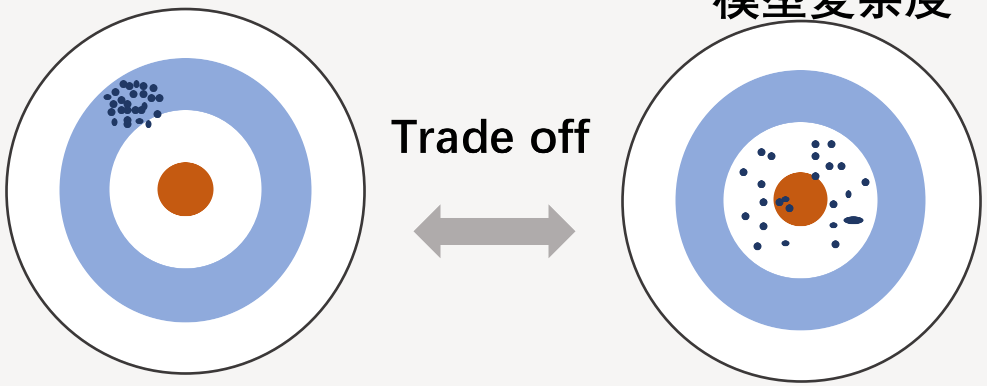
Bias大
离靶心远
模型没学好



Variance大
过度拟合当前样本
模型间泛化能力差

Bias很大
Variance较小

underfitting



Bias很小
Variance较大

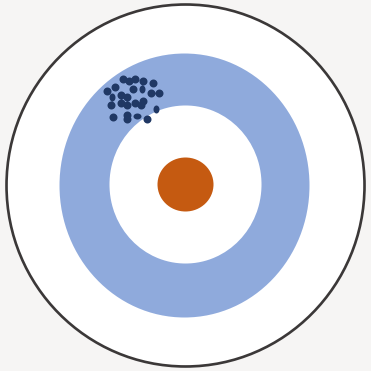
overfitting

Bias and Variance

Underfitting

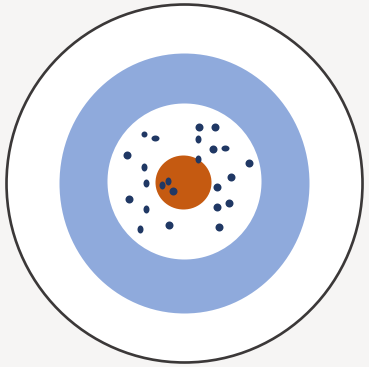
- 增加特性feature
- 选择更复杂的模型
- 优化学习率

Bias很大
Variance较小



Trade off

过程中
相互伤害



Overfitting

- 增加数据量
- Dropout
- 正则化

Bias较小
Variance很大

Boosting：
AdaBoost、XGBoost

Ensemble learning
集成学习

Bagging

Regularization

防止过拟合—— Regularization 正则化

- 重新设计修改 Loss function

原来

$$L(\theta)=\frac{1}{N}\sum_{i=1}^{N}\left(y^i-f(x^i)\right)^2$$

添加正则化项后

$$L(\theta)=\frac{1}{N}\sum_{i=1}^{N}\left(y^i-f(x^i)\right)^2+\boxed{\lambda\|w\|_p}$$

惩罚项：根据p的取值常用的有l_1-norm和l_2-norm

$\lambda\|w\|_p$越小越好，部分weight也尽可能小
因此，让$f(x^i)$整体更加平滑

$$y=w_i\cdot x_i+b$$

$$y \xrightarrow{\Delta w_i x_i} y^{new} \qquad x_i \xrightarrow{\Delta x_i} x^{new}$$

w^i 越趋近于0，$f(w,b)$ 对 x^i 的变化越不敏感，那么函数越平滑

Regularization

范数

范数是衡量某个向量空间（或矩阵）中的每个向量长度或大小。
对实数p≥1， 范数定义如下：

$$\|x\|_p = \left(\sum_{i=1}^{n}|x_i|^p\right)^{\frac{1}{p}}$$

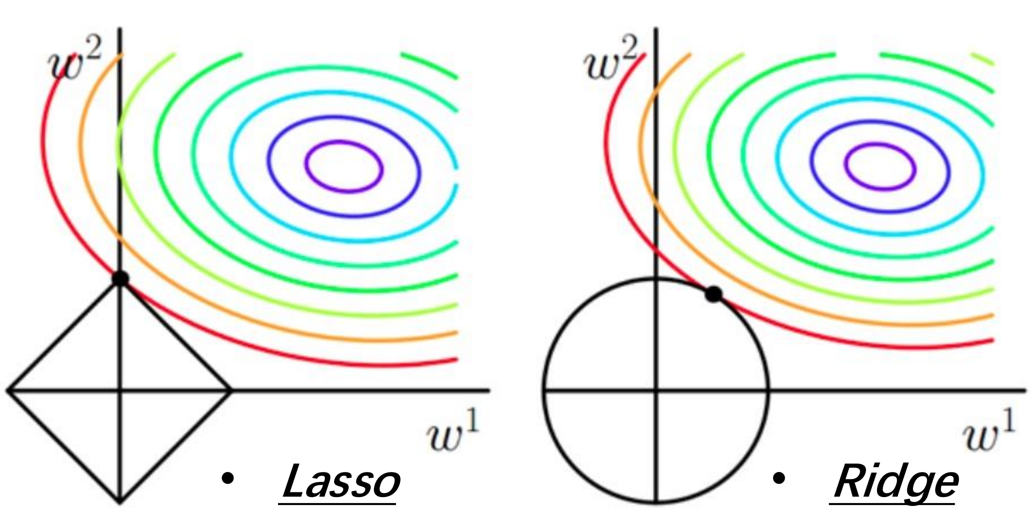
- L1范数：某个向量中所有元素绝对值的和
- L2范数：某个向量中所有元素平方和再开根（欧几里得距离）

- *Lasso回归*

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y^i - f(x^i)\right)^2 + \lambda\|w\|_1$$

- *Ridge回归（岭回归）*

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y^i - f(x^i)\right)^2 + \lambda\|w\|_2{}^2$$



- *Lasso*
- *Ridge*

Regularization

- L1正则（Lasso）

产生稀疏矩阵，用于特征选择

$$L(\theta)=\frac{1}{N}\sum_{i=1}^{N}\left(y^i-f(x^i)\right)^2+\boxed{\lambda\|w\|_1}$$

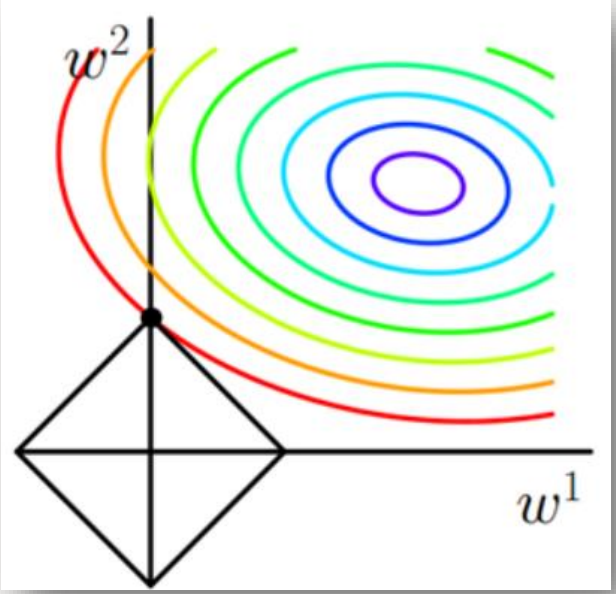
L1正则化是权值的绝对值之和

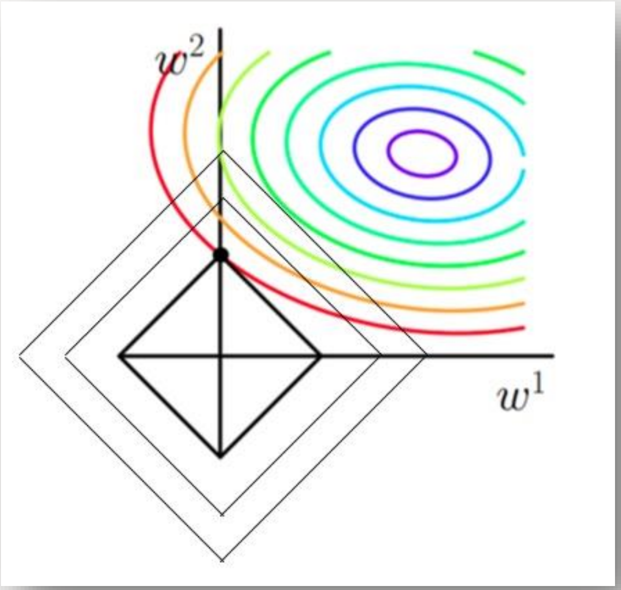
在预测/分类时，特征太多会难以选择。
如果只有部分特征对这个模型有贡献，其他特征w是0或者很小的时候，我们就可以只关注w是非零值的那些特征。

我们的任务：

在L1正则约束下求出$L(\theta)$取最小值的解

- 相交的地方是最优解
- 顶点处，某些权值为0
- λ系数越小，L1正则的方框越大





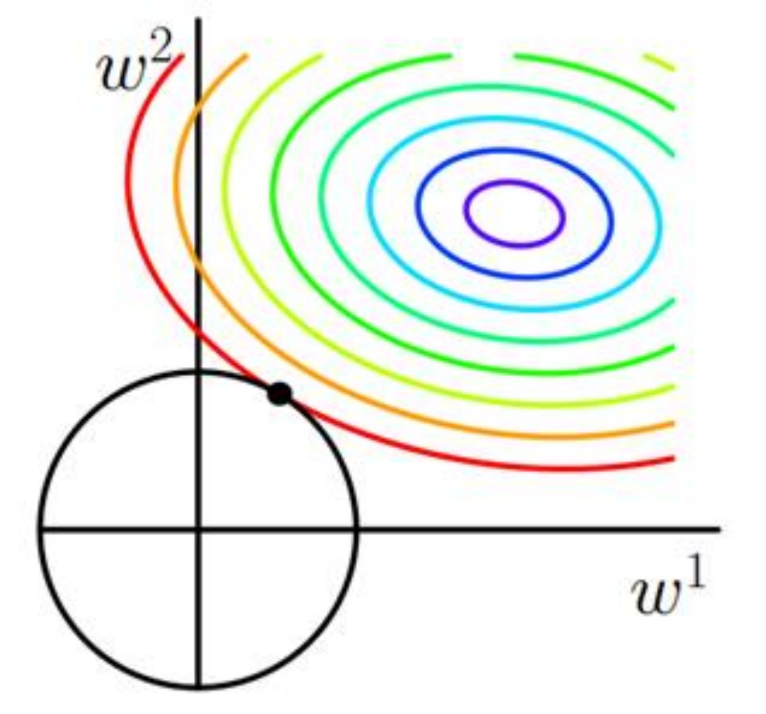
Regularization

- L2正则（Ridge）

不具有稀疏性，防止过拟合overfitting

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y^i - f\left(x^i\right)\right)^2 + \boxed{\lambda\|w\|_2^2}$$

→ L2正则化是各个feature权值的平方和（未开根号）



- 与方形相比，被磨去了棱角，两线相交处feature为0的可能性较小

加入L2正则之后，做梯度下降更新：

Learning rate

$$w := w - \alpha\frac{\partial L_2}{\partial w} = w - \boxed{\alpha}\left(\frac{\partial L}{\partial w} + 2\lambda w\right)$$

$$w := (1 - 2\alpha\lambda)w - \alpha\frac{\partial L}{\partial w}$$

- 每次都不断减小，抗扰动能力强，防止overfitting。

L1 与 L2 对比

- L1正则（Lasso）

$|w_1| + |w_2| + \cdots$

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y^i - f(x^i)\right)^2 + \lambda\boxed{\|w\|_1}$$

$$w := w - \alpha\frac{\partial L_1}{\partial w} = w - \alpha\left(\frac{\partial L}{\partial w} + \lambda\,\mathrm{sgn}(w)\right)$$

$$w := w - \alpha\frac{\partial L_1}{\partial w} = w - \alpha\frac{\partial L}{\partial w} - \alpha\lambda\,\mathrm{sgn}(w)$$

- W为正，减一个固定值，变小
- W为负，加一个固定值，变大

结果：参数有大有小，有稀疏性

- L2正则（Ridge）

$w_1^2 + w_2^2 + \cdots$

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y^i - f(x^i)\right)^2 + \lambda\boxed{\|w\|_2{}^2}$$

$$w := w - \alpha\frac{\partial L_2}{\partial w} = w - \alpha\left(\frac{\partial L}{\partial w} + 2\lambda w\right)$$

$$w := (1 - 2\alpha\lambda)w - \alpha\frac{\partial L}{\partial w}$$

- W无论正负都会按比例减小
- 哪怕很小也会有数值，不容易为0

结果：会保留接近0的值，平均都比较小

Regularization

正则化与Bias和Variance之间的关系？

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(y^i - f\left(x^i\right)\right)^2 + \boxed{\lambda}\|w\|_p$$

- 如果太大：函数会过于平滑，拟合效果比较差，所以Bias会比较大。

- 如果太小：对损失函数没有惩罚，对于过拟合的情况没有改善，因此Variance会比较大。

多选题 5分

设置

关于过拟合的描述正确的是?

A 训练集上loss较低，但测试集上loss较高

B 模型泛化能力较好

C bias很小，但variance很大

D 改善方法有：扩充数据集、正则化、增加模型复杂度等

提交

· Thanks ·

学生创新中心：肖雄子彦