太原理工大学
TAIYUAN UNIVERSITY OF TECHNOLOGY

太原理工大学
大数据学院
COLLEGE OF DATA SCIENCE
TAIYUAN UNIVERSITY OF TECHNOLOGY

深度学习中的正则化

Regularization for Deep Learning

本章内容概况

主要内容

01 测试误差来源分析

02 权重衰减：L1和L2正则化

PART

测试误差来源分析

ONE

Step 1: Model

w and b are parameters
(can be any value)

$y = b + w \cdot x_{cp}$

A set of function

Model

$f_1, f_2 \cdots$

f_1: $y = 10.0 + 9.0 \cdot x_{cp}$

f_2: $y = 9.8 + 9.2 \cdot x_{cp}$

f_3: $y = -0.8 - 1.2 \cdot x_{cp}$

...... infinite

$$f(x) = y$$

x_i: x_{cp}, x_{hp}, x_w, x_h ...

Linear model: $y = b + \sum w_i x_i$

feature

w_i: weight, b: bias

Step 2: Goodness of Function

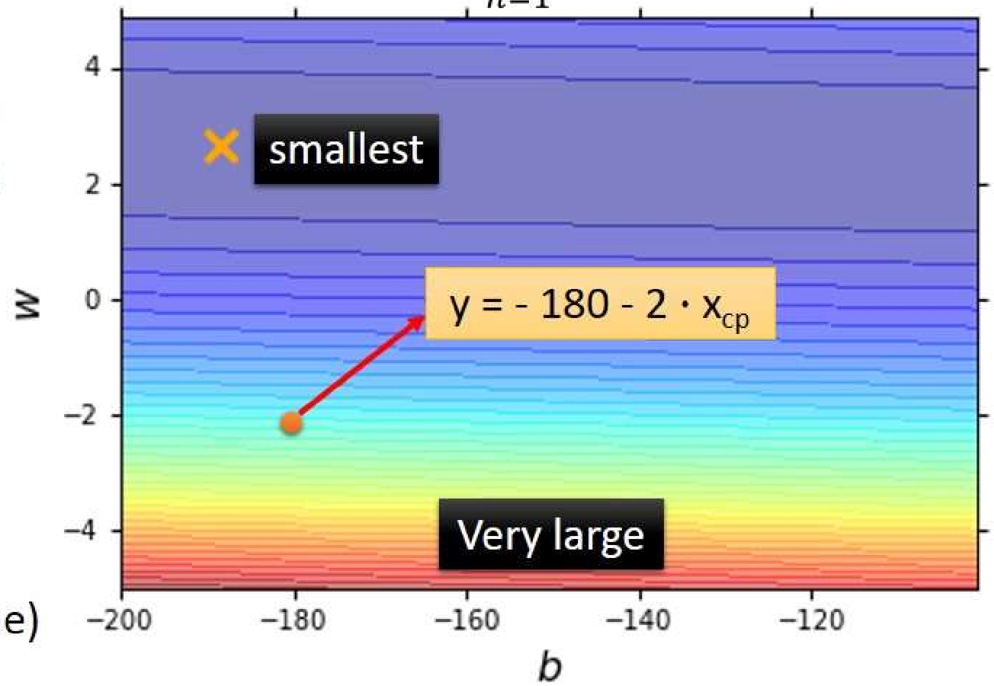
- Loss Function

$$L(w,b)=\sum_{n=1}^{10}\left(\hat{y}^n-\left(b+w\cdot x_{cp}^n\right)\right)^2$$

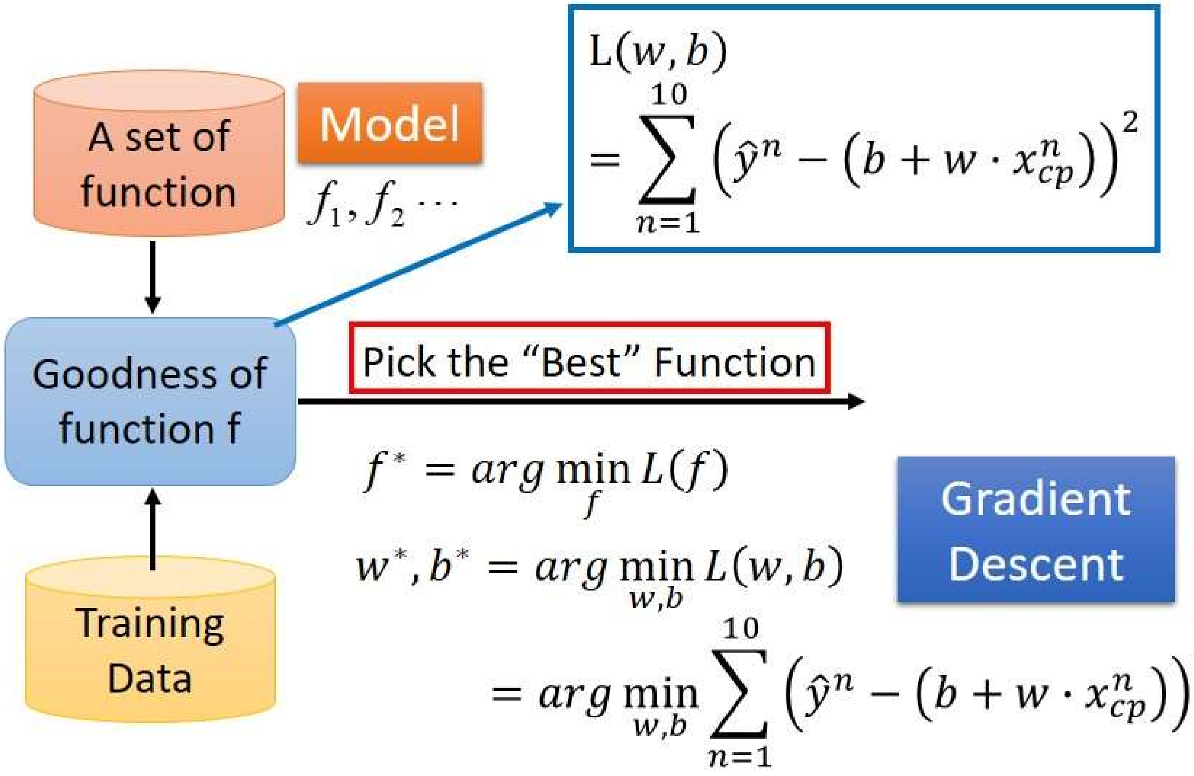
Each point in the figure is a function

The color represents $L(w,b)$.

(true example)



Step 3: Best Function



$f_1, f_2 \cdots$

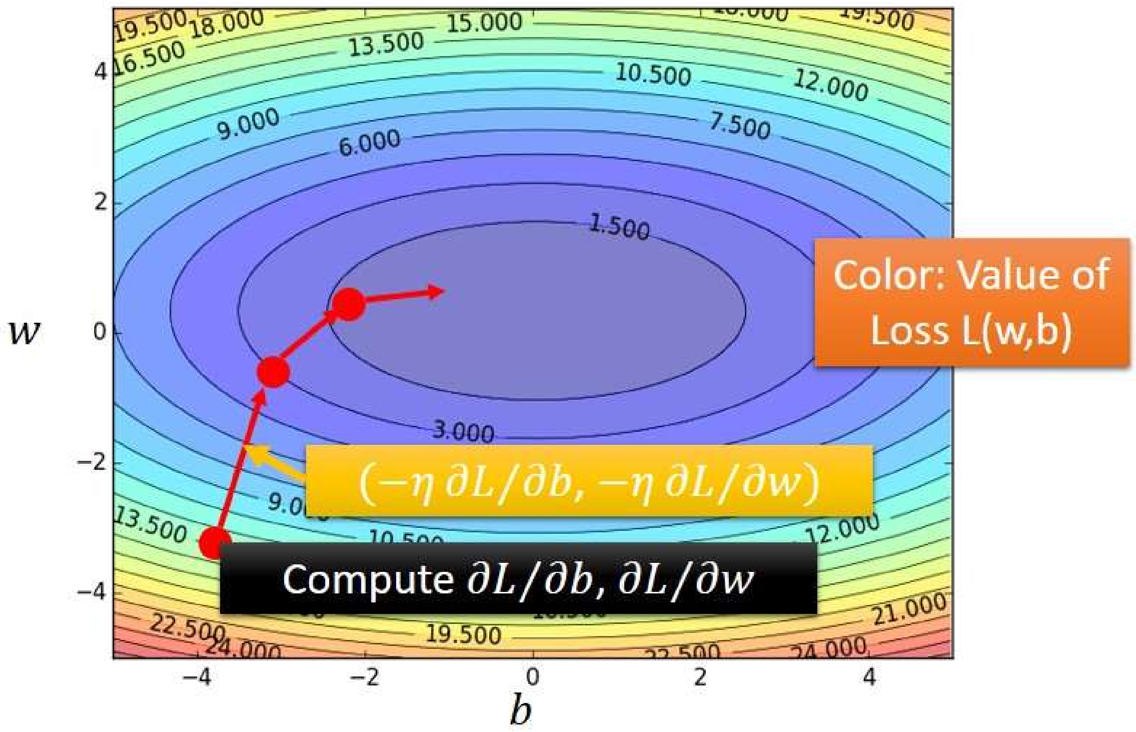
$$L(w,b) = \sum_{n=1}^{10} \left(\hat{y}^n - (b + w \cdot x_{cp}^n)\right)^2$$

$$f^* = arg \min_f L(f)$$

$$w^*, b^* = arg \min_{w,b} L(w,b)$$

$$= arg \min_{w,b} \sum_{n=1}^{10} \left(\hat{y}^n - (b + w \cdot x_{cp}^n)\right)^2$$

Step 3: Gradient Descent

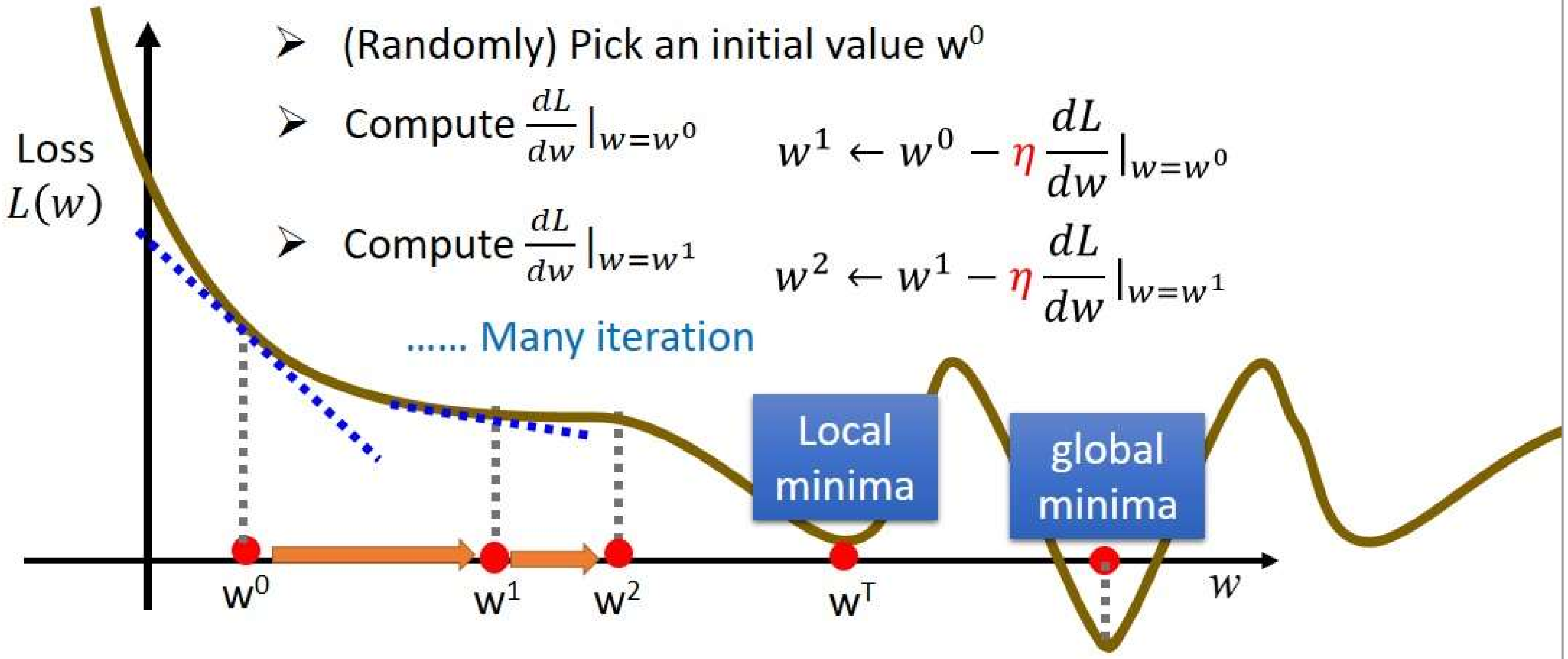


Step 3: Gradient Descent

$$w^* = arg \min_w L(w)$$

- Consider loss function $L(w)$ with one parameter w:
  - ➢ (Randomly) Pick an initial value w^0
  - ➢ Compute $\frac{dL}{dw}|_{w=w^0}$ $\quad w^1 \leftarrow w^0 - \eta \frac{dL}{dw}|_{w=w^0}$
  - ➢ Compute $\frac{dL}{dw}|_{w=w^1}$ $\quad w^2 \leftarrow w^1 - \eta \frac{dL}{dw}|_{w=w^1}$

...... Many iteration

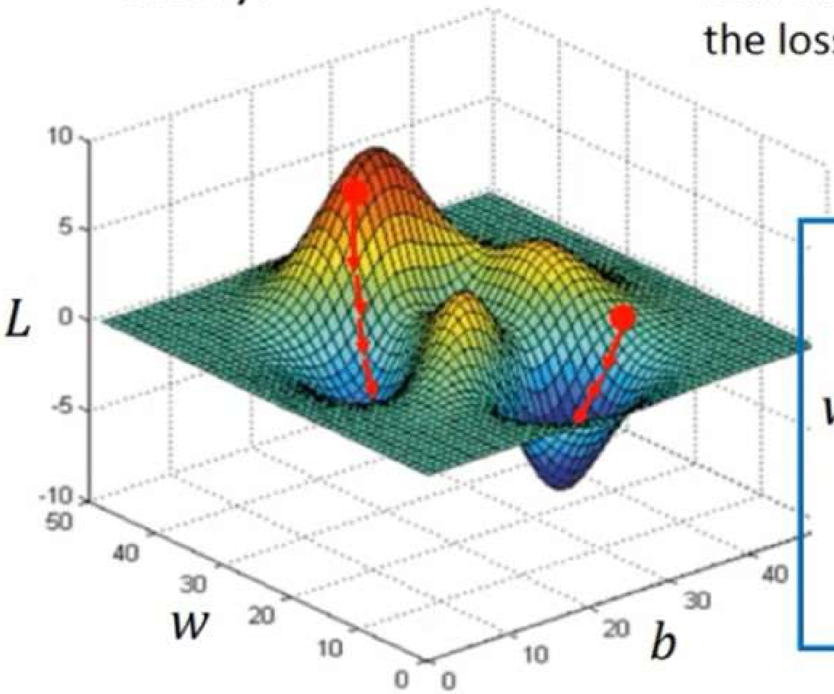


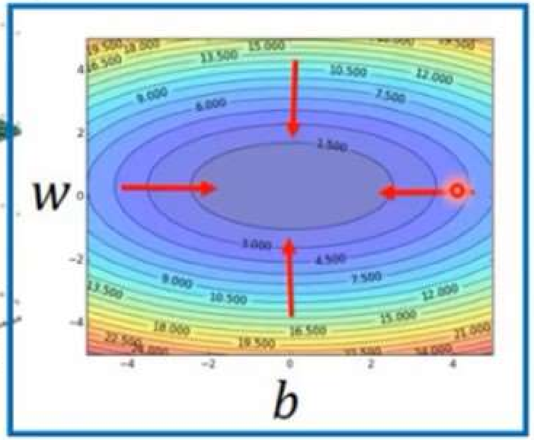
Step 3: Gradient Descent

- Worry?

Don't worry. In linear regression, the loss function L is convex.

No local optimal





引言

线性回归（额外补充）

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

$$min_\beta ||\mathbf{y} - \mathbf{x}\boldsymbol{\beta}||_2^2$$

where $\mathbf{y} \in R^n$, $\mathbf{x} \in R^{n\times p}$ and $\boldsymbol{\beta} \in R^p$.

Pros and cons

- Closed-form solution $\boldsymbol{\beta} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}$ when $n > p$;
- Easy to overfit;
- The solution is not well-defined when $n < p$.

10

引言

线性回归（额外补充）

令

$$X = \begin{bmatrix} — (x^{(1)})^T — \\ — (x^{(2)})^T — \\ \vdots \\ — (x^{(m)})^T — \end{bmatrix} \quad \vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

则

$$X\theta - \vec{y} = \begin{bmatrix} (x^{(1)})^T\theta \\ \vdots \\ (x^{(m)})^T\theta \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} = \begin{bmatrix} h_\theta(x^{(1)}) - y^{(1)} \\ \vdots \\ h_\theta(x^{(m)}) - y^{(m)} \end{bmatrix}$$

所以

$$\begin{aligned} \frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y}) &= \frac{1}{2}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 \\ &= J(\theta) \end{aligned}$$

注：（矩阵推导）

$$\mathrm{tr}ABC = \mathrm{tr}CAB = \mathrm{tr}BCA,$$

$$\mathrm{tr}ABCD = \mathrm{tr}DABC = \mathrm{tr}CDAB = \mathrm{tr}BCDA.$$

$$\mathrm{tr}A = \mathrm{tr}A^T$$

$$\mathrm{tr}(A+B) = \mathrm{tr}A + \mathrm{tr}B$$

$$\mathrm{tr}\,aA = a\mathrm{tr}A$$

$$\begin{aligned} \nabla_A \mathrm{tr}AB &= B^T & (1) \\ \nabla_{A^T} f(A) &= (\nabla_A f(A))^T & (2) \\ \nabla_A \mathrm{tr}ABA^TC &= CAB + C^TAB^T & (3) \\ \nabla_A |A| &= |A|(A^{-1})^T. & (4) \end{aligned}$$

11

引言

线性回归（额外补充）

由（2）（3）知：

$$\nabla_{A^T}\operatorname{tr}ABA^TC = B^TA^TC^T + BA^TC$$

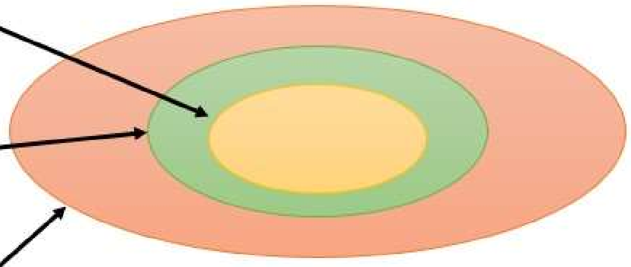
则

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \frac{1}{2}(X\theta-\vec{y})^T(X\theta-\vec{y}) \\
&= \frac{1}{2}\nabla_\theta \left(\theta^TX^TX\theta - \theta^TX^T\vec{y} - \vec{y}^TX\theta + \vec{y}^T\vec{y}\right) \\
&= \frac{1}{2}\nabla_\theta \operatorname{tr}\left(\theta^TX^TX\theta - \theta^TX^T\vec{y} - \vec{y}^TX\theta + \vec{y}^T\vec{y}\right) \\
&= \frac{1}{2}\nabla_\theta \left(\operatorname{tr}\theta^TX^TX\theta - 2\operatorname{tr}\vec{y}^TX\theta\right) \\
&= \frac{1}{2}\left(X^TX\theta + X^TX\theta - 2X^T\vec{y}\right) \\
&= X^TX\theta - X^T\vec{y}
\end{aligned}$$

$$\theta = (X^TX)^{-1}X^T\vec{y}.$$

Model Selection

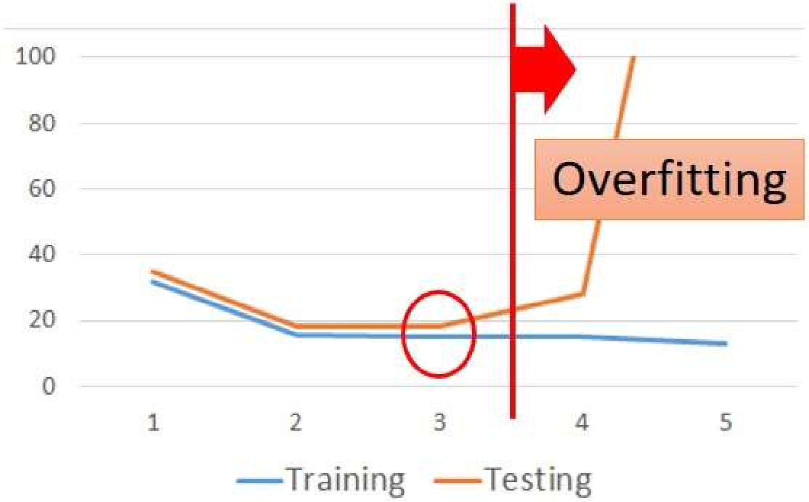
1. $y = b + w \cdot x_{cp}$
2. $y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2$
3. $y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3$
4. $y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4$
5. $y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$



A more complex model yields lower error on training data.

If we can truly find the best function

Model Selection



	Training	Testing
1	31.9	35.0
2	15.4	18.4
3	15.3	18.1
4	14.9	28.2
5	12.8	232.1

A more complex model does not always lead to better performance on ***testing data***.

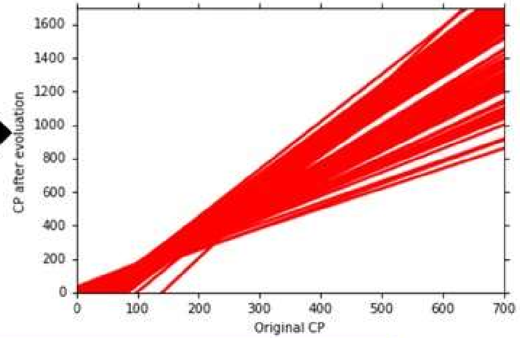
This is ***Overfitting***. → Select suitable model

引言

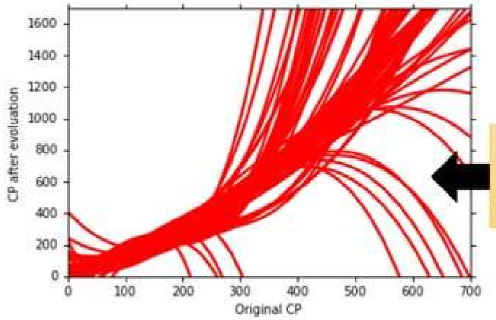
where err: 偏差与方差

f^* in 100 Universes

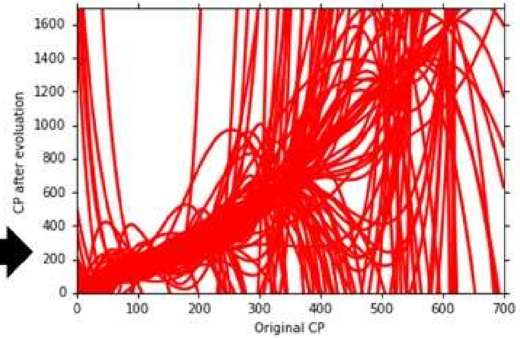
$y = b + w \cdot x_{cp}$



$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3$



$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$



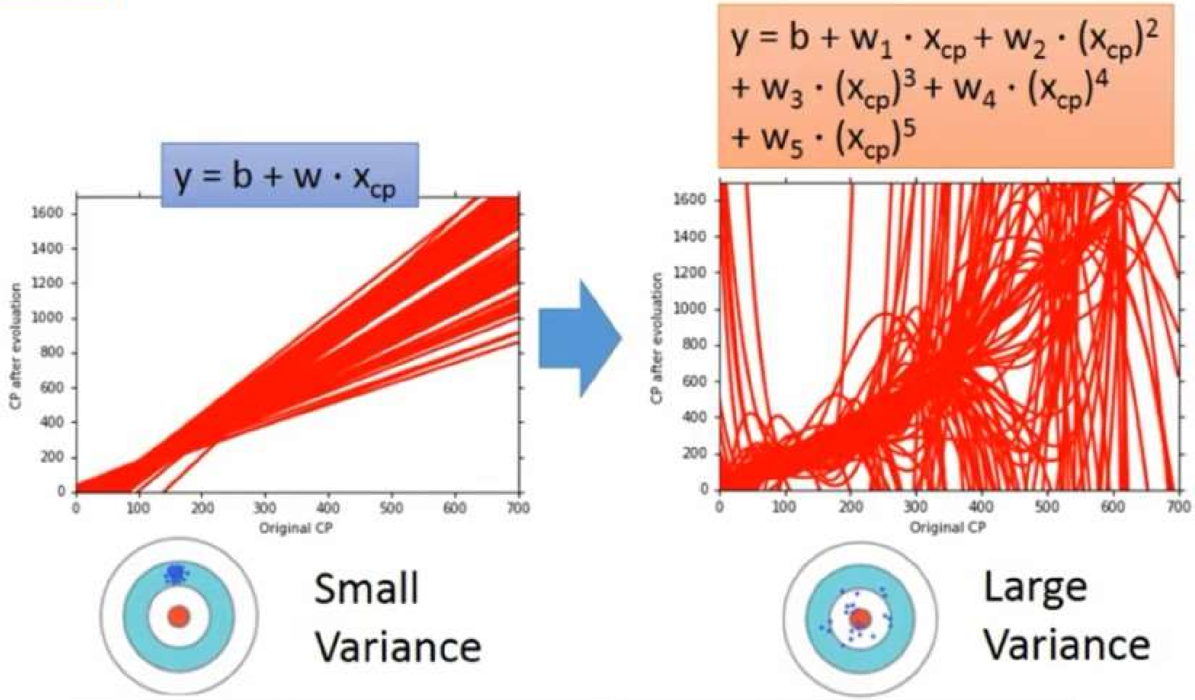
Slide credit: Hung-yi Lee

15

引言

where err: 偏差与方差

Variance



Simpler model is less influenced by the sampled data

Slide credit: Hung-yi Lee

16

引言

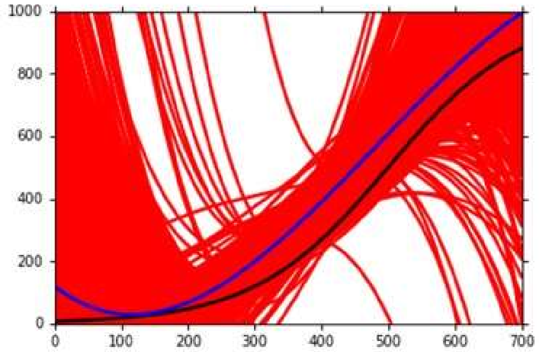
where err: 偏差与方差

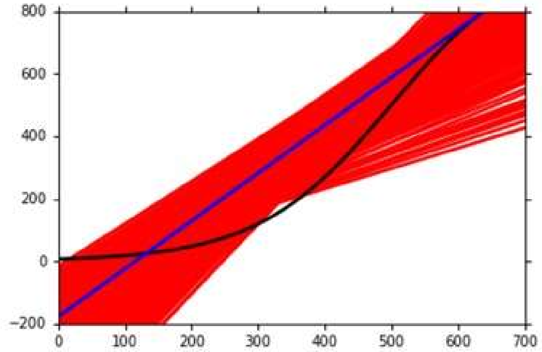
Bias

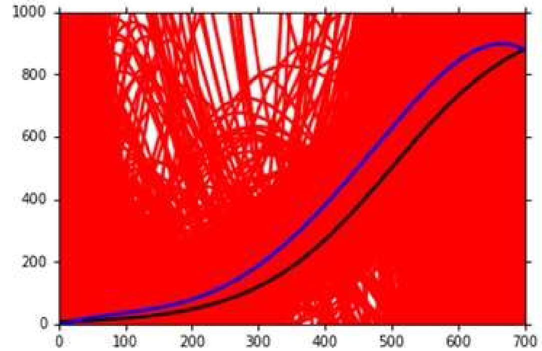
Black curve: the true function $\hat{f}$

Red curves: 5000 f^*

Blue curve: the average of 5000 f^* $= \bar{f}$







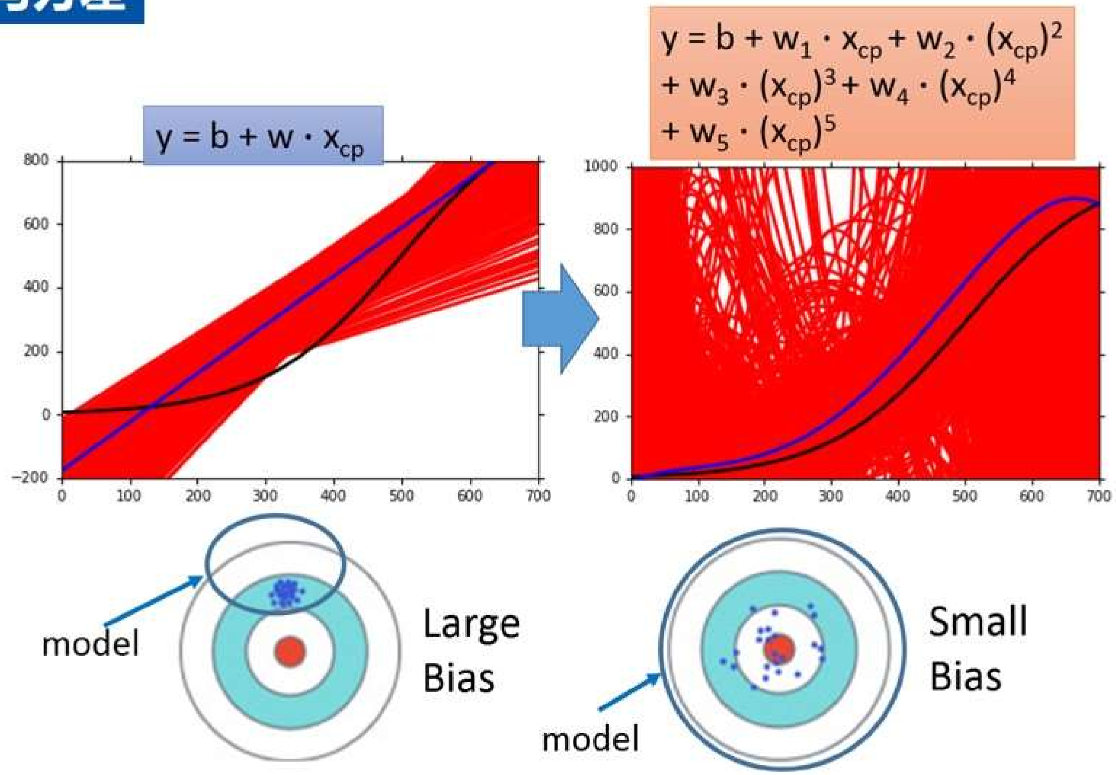
Slide credit: Hung-yi Lee

17

引言

where err: 偏差与方差

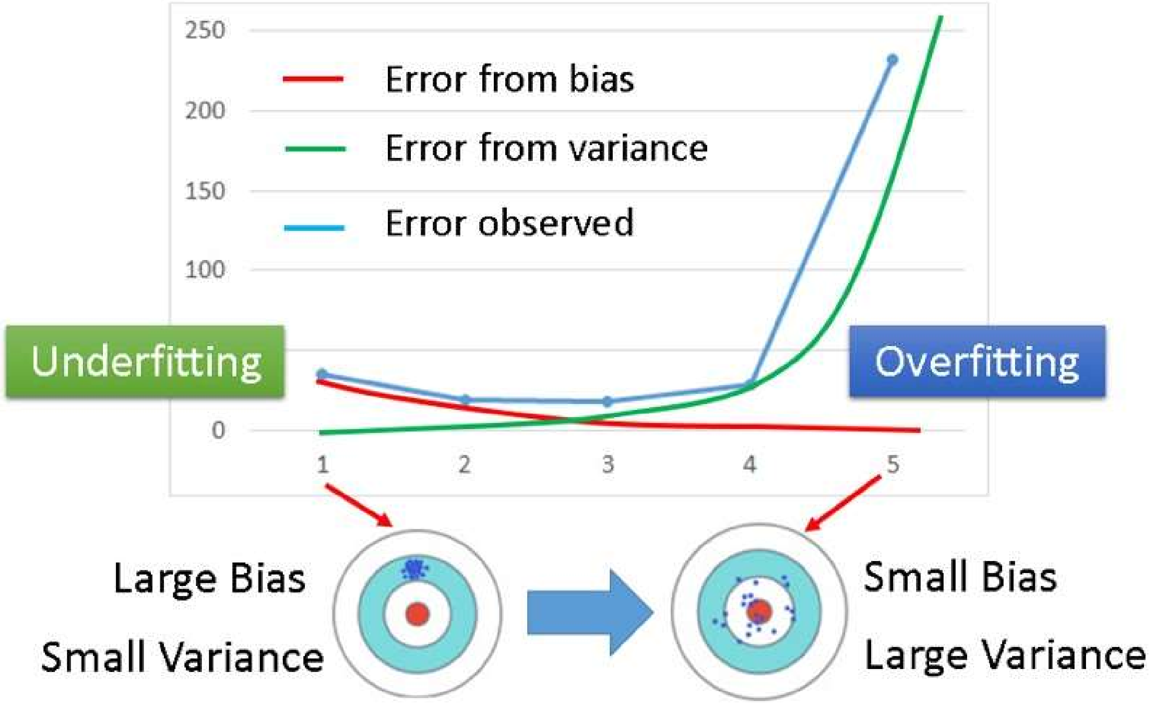
Bias



Slide credit: Hung-yi Lee

引言

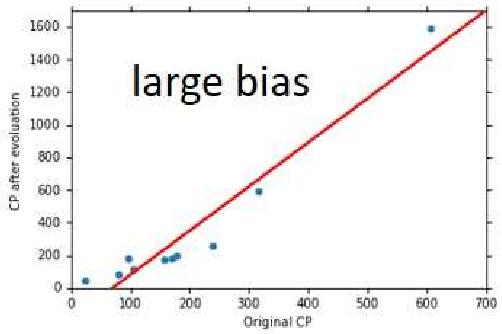
where err: 偏差与方差



Slide credit: Hung-yi Lee

What to do with large bias?

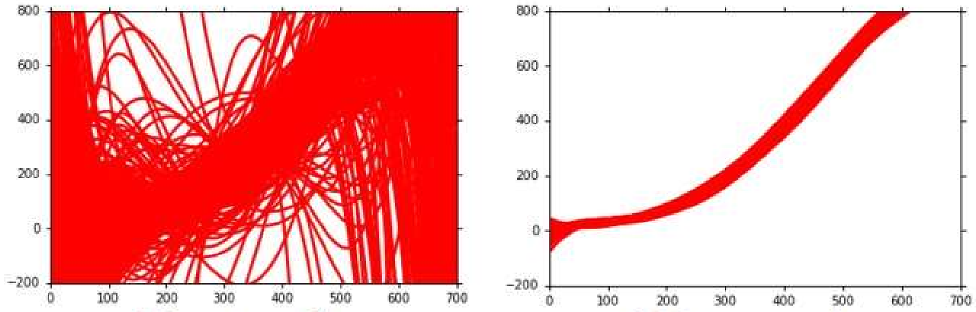
- Diagnosis:
  - If your model cannot even fit the training examples, then you have large bias **Underfitting**
  - If you can fit the training data, but large error on testing data, then you probably have large variance **Overfitting**
- For bias, redesign your model:
  - Add more features as input
  - A more complex model



What to do with large variance?

- More data

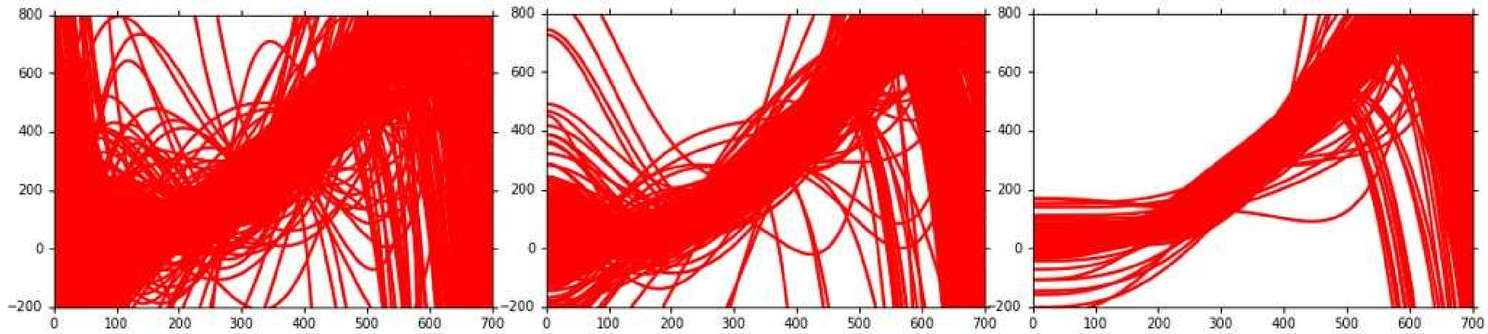
  Very effective,
  but not always
  practical



10 examples 100 examples

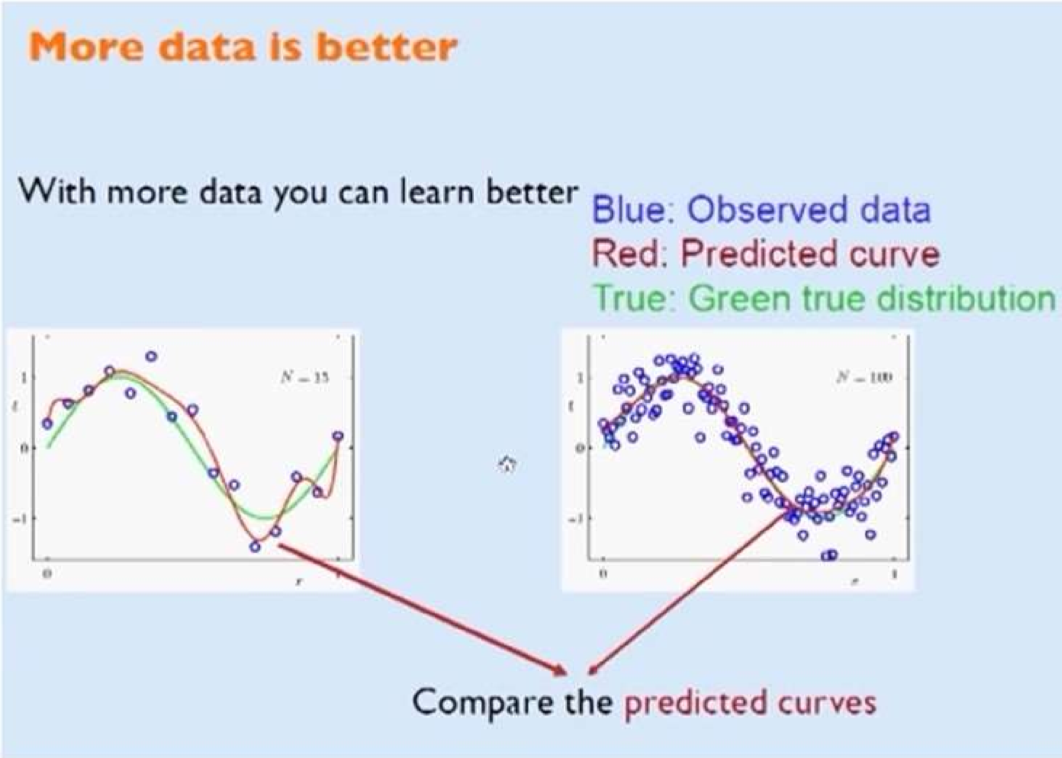
- Regularization

May increase bias



权重衰减

增加数据



More data is better
With more data you can learn better
Blue: Observed data
Red: Predicted curve
True: Green true distribution
N = 15
N = 100
Compare the predicted curves

引言

权重衰减

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

The functions with smaller w_i are better

$$L = \sum_n \left(\hat{y}^n - \left(b + \sum w_i x_i \right) \right)^2 \boxed{+\lambda \sum (w_i)^2}$$

- Smaller w_i means … smoother

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

$$y + \sum w_i \Delta x_i = b + \sum w_i (x_i + \Delta x_i)$$

- We believe smoother function is more likely to be correct

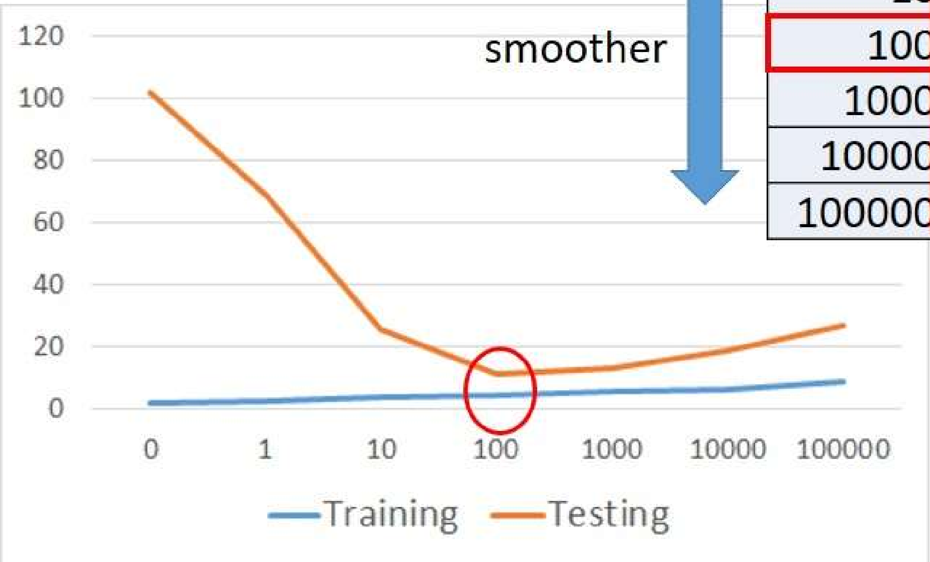
Do you have to apply regularization on bias?

23

引言
权重衰减

$$L = \sum_n \left(\hat{y}^n - \left(b + \sum w_i x_i\right)\right)^2 \boxed{+\lambda \sum (w_i)^2}$$

λ	Training	Testing
0	1.9	102.3
1	2.3	68.7
10	3.5	25.7
100	4.1	11.1
1000	5.6	12.8
10000	6.3	18.7
100000	8.5	26.8



How smooth?

Select λ obtaining the best model

- Training error: largerλ, considering the training error less
- We prefer smooth function, but don't be too smooth.

PART TWO

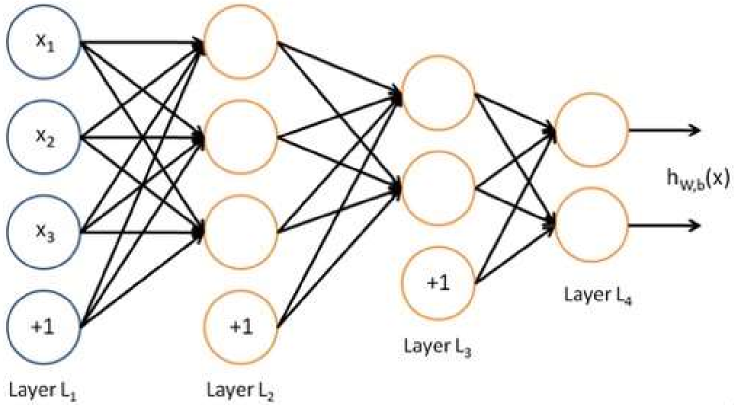
权重衰减
L1和L2正则化

概念

Definition

- 神经网络
  - 过度参数化
  - 拟合能力强

~~泛化性差~~



Zhang C, Bengio S, Hardt M, et al.

Understanding deep learning requires rethinking generalization.

ICLR 2017

Regularization is any modification we make to a learning algorithm that is intended to reduce its generalization error but not its training error

正则化（Regularization）是一类通过限制模型复杂度，从而避免过拟合，提高泛化能力的方法，包括引入一些约束规则，增加先验、提前停止等。

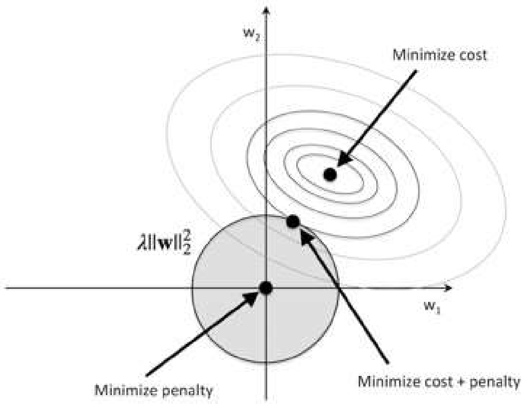
26

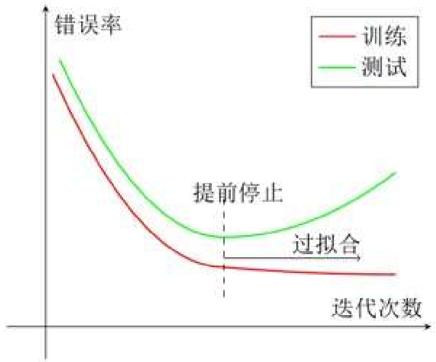
概念

Definition

所有损害优化的方法都是正则化。

增加优化约束	干扰优化过程
L1 / L2约束、数据增强	权重衰减、随机梯度下降、提前停止





深层神经网络的优化和正则化是即对立又统一的关系。

一方面我们希望优化算法能找到一个全局最优解（或较好的局部最优解），

另一方面我们又不希望模型优化到最优解，这可能陷入过拟合。

优化和正则化的统一目标是期望风险最小化。

27

概念

Definition

- 如何提高神经网络的泛化能力
  - L1和L2正则化
  - Early Stop
  - 权重衰减
  - SGD
  - Dropout
  - 数据增强

Slide credit: Xipeng Qiu

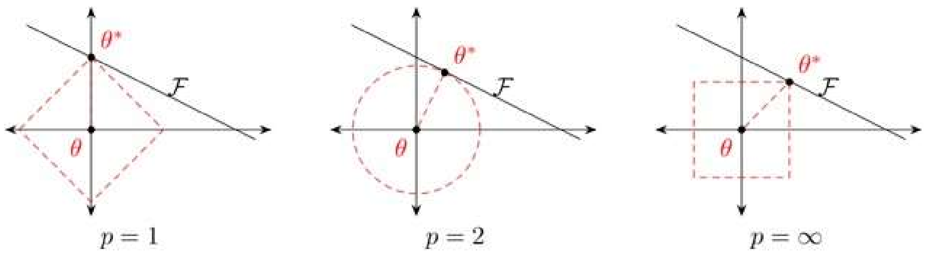
权重衰减

L1和L2正则化

- 优化问题可以写为

$$\theta^* = \arg\min_\theta \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}\big(y^{(n)}, f(\mathbf{x}^{(n)}, \theta)\big) + \lambda \ell_p(\theta)$$

- ℓ_p 为范数函数，p的取值通常为{1,2}代表ℓ_1和ℓ_2范数，λ为正则化系数。



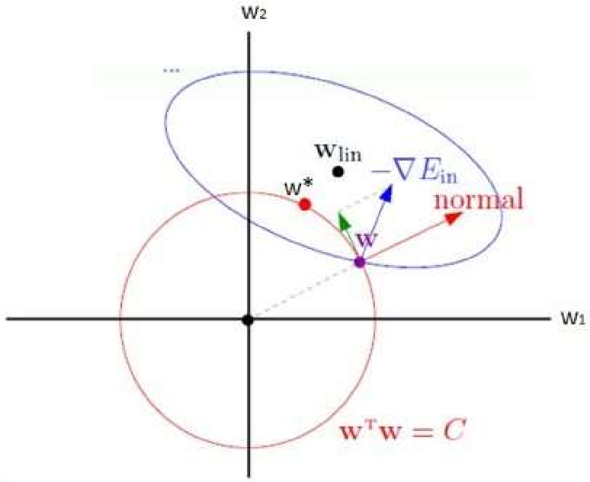
Slide credit: Xipeng Qiu

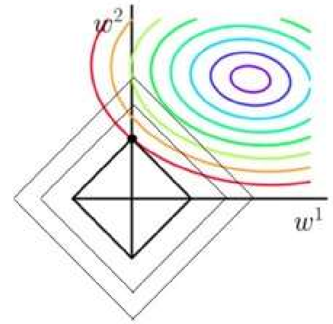
29

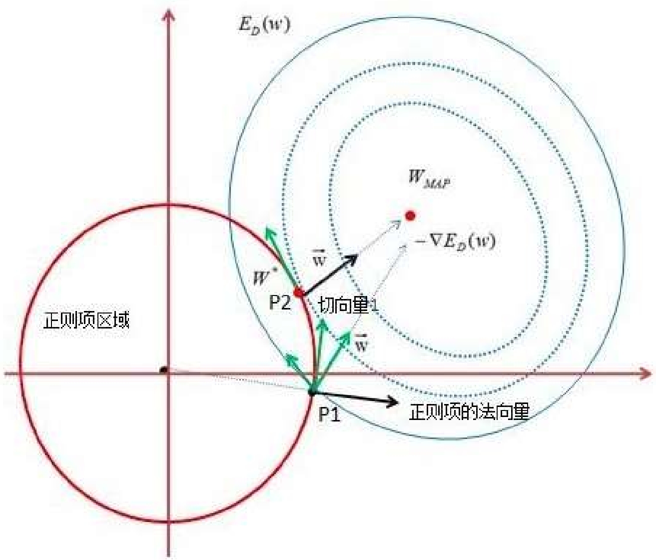
优化角度

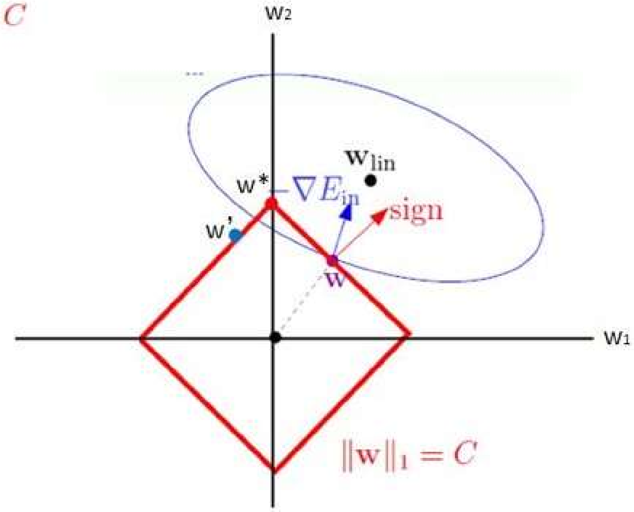
权重衰减

为什么是在相切点







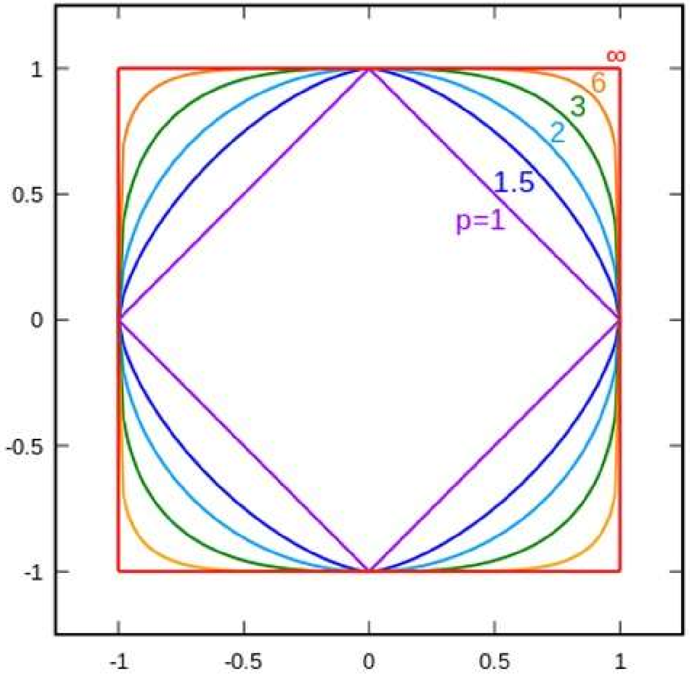


31

雨课堂 Rain Classroom

权重衰减

不同的范数



all p-norms penalize larger weights

p < 2 tends to create sparse (i.e. lots of 0 weights)

p > 2 tends to like similar weights

32

梯度角度

权重衰减

L2正则化

$$\tilde{J}(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y})=\frac{\alpha}{2}\boldsymbol{w}^{\top}\boldsymbol{w}+J(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y})$$

- 在每次参数更新时，引入一个衰减系数w。

$$\boldsymbol{w}\leftarrow\boldsymbol{w}-\epsilon(\alpha\boldsymbol{w}+\nabla_{\boldsymbol{w}}J(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y}))$$

$$\boldsymbol{w}\leftarrow(1-\epsilon\alpha)\boldsymbol{w}-\epsilon\nabla_{\boldsymbol{w}}J(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y})$$

- 在标准的随机梯度下降中，权重衰减正则化和L2正则化的效果相同。
- 在较为复杂的优化方法（比如Adam）中，权重衰减和L2正则化并不等价。

34

权重衰减

L2正则化

$$\tilde{J}(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y})=\frac{\alpha}{2}\boldsymbol{w}^{\top}\boldsymbol{w}+J(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y})$$

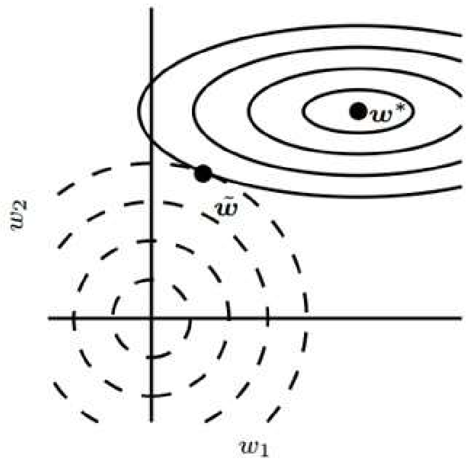


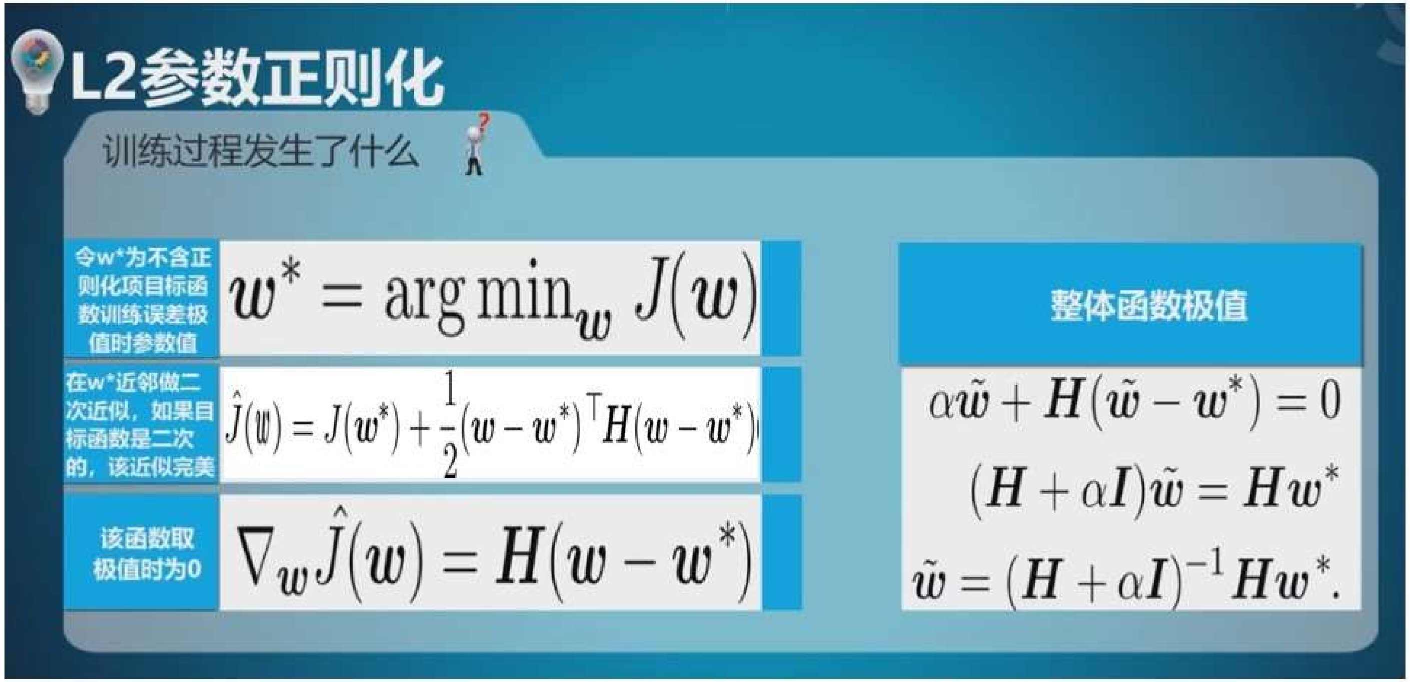
图 7.1: L^2（或权重衰减）正则化对最佳 $\boldsymbol{w}$ 值的影响。实线椭圆表示没有正则化目标的等值线。虚线圆圈表示 L^2 正则化项的等值线。在 $\tilde{\boldsymbol{w}}$ 点，这两个竞争目标达到平衡。目标函数 J 的 Hessian 的第一维特征值很小。当从 $\boldsymbol{w}^*$ 水平移动时，目标函数不会增加得太多。因为目标函数对这个方向没有强烈的偏好，所以正则化项对该轴具有强烈的影响。正则化项将 w_1 拉向零。而目标函数对沿着第二维远离 $\boldsymbol{w}^*$ 的移动非常敏感。对应的特征值较大，表示高曲率。因此，权重衰减对 w_2 的位置影响相对较小。

35

权重衰减

L2正则化

$$\tilde{J}(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y})=\frac{\alpha}{2}\boldsymbol{w}^{\top}\boldsymbol{w}+J(\boldsymbol{w};\boldsymbol{X},\boldsymbol{y})$$



Slide credit: Chun Yuan

纸质版花书(教材)上内容 P_{144}
电子版花书(教材)上内容 P_{200}-P_{201}

36

权重衰减

L2正则化

L2参数正则化

训练过程发生了什么？

$$
\begin{aligned}
\tilde{w} &= (Q\Lambda Q^\top + \alpha I)^{-1} Q\Lambda Q^\top w^* \\
&= \left[Q(\Lambda + \alpha I)Q^\top\right]^{-1} Q\Lambda Q^\top w^* \\
&= Q(\Lambda + \alpha I)^{-1} \Lambda Q^\top w^*.
\end{aligned}
$$

W* 对应第i个特征值的部分，尺度变化：

$$\tilde{w} = \frac{\lambda_i}{\lambda_i + \alpha} w^*$$

Slide credit: Chun Yuan

37

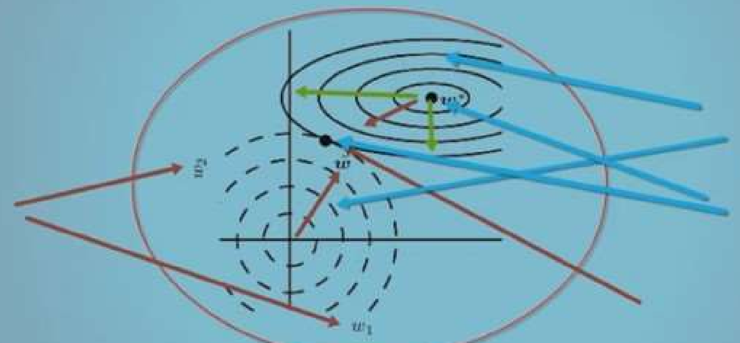
权重衰减

L2正则化

$$\tilde{w} = \frac{\lambda_i}{\lambda_i + \alpha} w^*$$

L2参数正则化

训练时发生了什么？



$\lambda_i \gg \alpha$ 沿着H 特征值较大的方向，正则化的影响较小。

$\lambda_i \ll \alpha$ 沿着H 特征值较小的方向，正则化的影响较大。

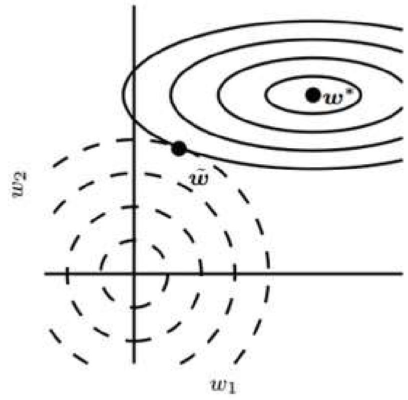
Slide credit: Chun Yuan

38

权重衰减

L2正则化

$$\tilde{w} = \frac{\lambda_i}{\lambda_i + \alpha} w^*$$



L2参数正则化

结论：

L2参数正则化主要针对损失函数特征向量不重要的方向：

对应Hessian矩阵较小的特征值，
改变参数不会显著增加梯度，
不重要方向对应的分量会在训练过程中因正则而衰减；

Slide credit: Chun Yuan

39

权重衰减

L1正则化

L2最常用，但是有时也用L1

$$\Omega(\boldsymbol{\theta}) = \|\boldsymbol{w}\|_1 = \sum_i |w_i|$$

和L2有什么区别呢？采用同样分析法

$$\tilde{J}(\boldsymbol{w}; \boldsymbol{X}, \boldsymbol{y}) = \alpha \|\boldsymbol{w}\|_1 + J(\boldsymbol{w}; \boldsymbol{X}, \boldsymbol{y})$$

$$\nabla_{\boldsymbol{w}} \tilde{J}(\boldsymbol{w}; \boldsymbol{X}, \boldsymbol{y}) = \alpha \operatorname{sign}(\boldsymbol{w}) + \nabla_{\boldsymbol{w}} J(\boldsymbol{w}; \boldsymbol{X}, \boldsymbol{y})$$

正则化对梯度的影响不再是线性地缩放每个Wi；
添加了一项sign(wi)同号的函数；
使用这种形式的梯度后，不一定能得到J(x;y;w)二次近似的直接算术解。

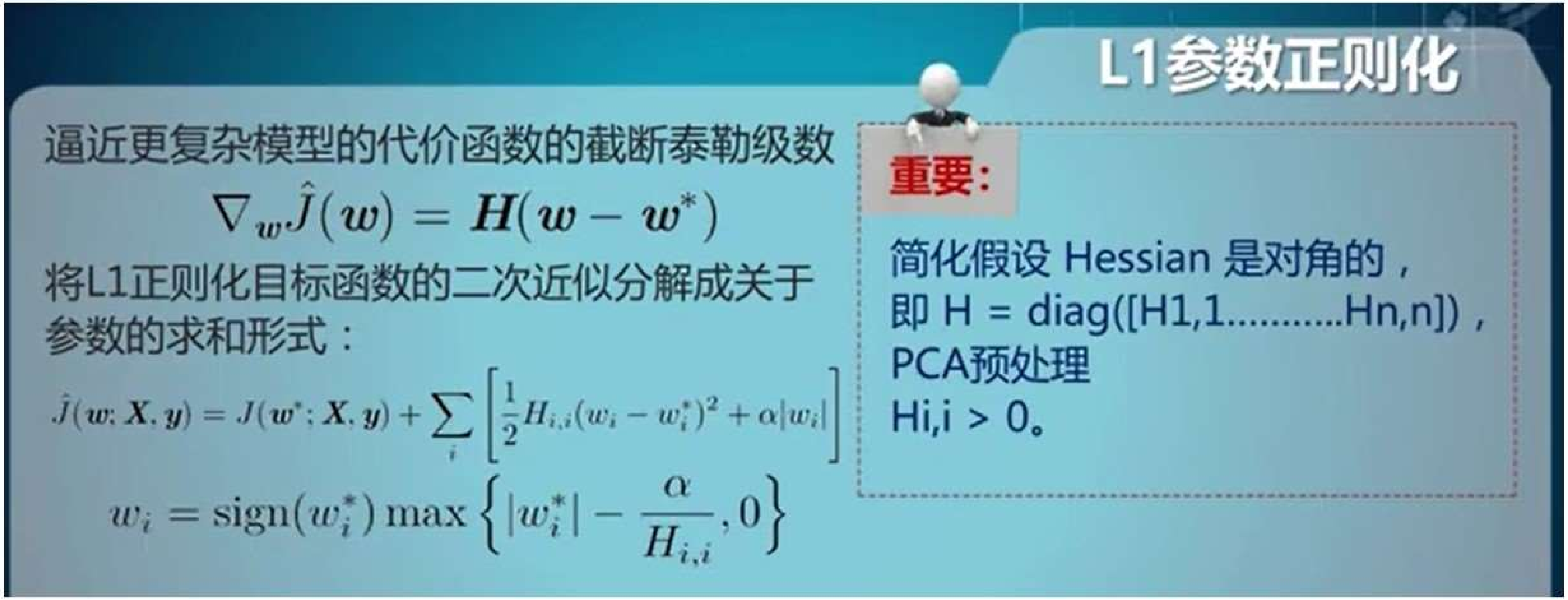
怎么解决?

Slide credit: Chun Yuan

40

权重衰减

L1正则化



Slide credit: Chun Yuan

权重衰减

L1正则化

根据公式： $w_i = \text{sign}(w_i^*) \max\left\{|w_i^*| - \frac{\alpha}{H_{i,i}}, 0\right\}$

分析wi*的情况

$w_i^* > 0$

- $w_i^* \leq \frac{\alpha}{H_{i,i}}$ 贡献小，L1正则化将wi推向0。
- $w_i^* > \frac{\alpha}{H_{i,i}}$ 贡献大，L1正则化将wi移动 $\frac{\alpha}{H_{i,i}}$ 的距离。

$w_i^* < 0$ L1 惩罚项使 wi 更接近 0 (增加 $\frac{\alpha}{H_{i,i}}$) 或者为0。

Slide credit: Chun Yuan

42

概率角度

权重衰减

L1L2

Norm Penalties

- L1: Encourages sparsity, equivalent to MAP Bayesian estimation with Laplace prior
- Squared L2: Encourages small weights, equivalent to MAP Bayesian estimation with Gaussian prior

44

权重衰减

L1L2

Ridge Regression

- Adds an L2 regularizer to Linear Regression

$$J_{\mathrm{RR}}(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) + \boxed{\lambda ||\boldsymbol{\theta}||_2^2}$$

$$= \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{\theta}^T\mathbf{x}^{(i)} - y^{(i)})^2 + \boxed{\lambda\sum_{k=1}^{K}\theta_k^2}$$

prefers parameters close to zero

- Bayesian interpretation: MAP estimation with a **Gaussian prior** on the parameters

$$\boldsymbol{\theta}^{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(y^{(i)}|\mathbf{x}^{(i)}) + \log p(\boldsymbol{\theta})$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ J_{\mathrm{RR}}(\boldsymbol{\theta})$$

where $p(\boldsymbol{\theta}) \sim \mathcal{N}(0, \frac{1}{\lambda})$

60

45

权重衰减

L1L2

LASSO

- Adds an **L1 regularizer** to Linear Regression

$$J_{\text{LASSO}}(\boldsymbol{\theta}) = J(\boldsymbol{\theta}) + \boxed{\lambda||\boldsymbol{\theta}||_1}$$
$$= \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{\theta}^T\mathbf{x}^{(i)} - y^{(i)})^2 + \boxed{\lambda\sum_{k=1}^{K}|\theta_k|}$$

yields sparse parameters (exact zeros)

- Bayesian interpretation: MAP estimation with a **Laplace prior** on the parameters

$$\boldsymbol{\theta}^{MAP} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(y^{(i)}|\mathbf{x}^{(i)}) + \log p(\boldsymbol{\theta})$$
$$= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, J_{\text{LASSO}}(\boldsymbol{\theta})$$

where $p(\boldsymbol{\theta}) \sim \text{Laplace}(0, f(\lambda))$

61

46

权重衰减

L1L2

从贝叶斯先验概率看正则化

假设输入空间是 $X \in \mathbb{R}^n$ 输出空间是 Y ，不妨假设含有 m 个样本数据 $(x^{(1)}, y^{(1)})$、$(x^{(2)}, y^{(2)})$、$\cdots$、$(x^{(m)}, y^{(m)})$ ，其中 $x^{(i)} \in X$、$y^{(i)} \in Y$ 。

贝叶斯学派认为参数 θ 也是服从某种概率分布的，即先给定 θ 的先验分布为 $p(\theta)$ ，然后根据贝叶斯定理 $P(\theta|(X,Y)) = \frac{P((Y,X);\theta) \times P(\theta)}{P(X,Y)} \sim P(Y|X;\theta) \times P(\theta)$ （这里的 $Y|X$ 仅仅是一种记号，代表给定的 X 对应相关的 Y ），因此通过极大似然估计可求参数 θ 。

$$\arg\max_{\theta} \ L(\theta) = \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta)p(\theta)$$

等价于求解对数化极大似然函数

$$\begin{aligned}
\arg\max_{\theta} \ l(\theta) &= \log L(\theta) \\
&= \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta) + \sum_{i=1}^{m} \log p(\theta)
\end{aligned}$$

$$\begin{aligned}
\Leftrightarrow \arg\min_{\theta} \ -l(\theta) &= -\log L(\theta) \\
&= -\sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta) - \sum_{i=1}^{m} \log p(\theta) \\
&= f(\theta) - \sum_{i=1}^{m} \log p(\theta)
\end{aligned}$$

47

权重衰减

L1L2

- **$L1$ 正则化的概率解释**

  假设 θ 服从的先验分布为均值为 0 参数为 λ 的拉普拉斯分布，即 $\theta \sim La(0,\lambda)$ 其中，$p(\theta) = \frac{1}{2\lambda}e^{-\frac{|\theta|}{\lambda}}$ 。因此，上述优化函数可转换为：

$$
\begin{aligned}
&\arg\min_{\theta}\ f(\theta) - \sum_{i=1}^{m}\log p(\theta) \\
&= f(\theta) - \sum_{i=1}^{m}\log\frac{1}{2\lambda}e^{-\frac{|\theta_i|}{\lambda}} \\
&= f(\theta) - \sum_{i=1}^{m}\log\frac{1}{2\lambda} + \frac{1}{\lambda}\sum_{i=1}^{m}|\theta_i| \\
&\Leftrightarrow \arg\min_{\theta}\ f(\theta) + \lambda\|\theta\|_1
\end{aligned}
$$

从上面的数学推导可以看出， $L1$ 正则化可以看成是：通过假设权重参数 θ 的先验分布为拉普拉斯分布，由最大后验概率估计导出。

48

权重衰减

L1L2

- **$L2$ 正则化的概率解释**

  假设 θ 服从的先验分布为均值为 0 方差为 σ^2 的正态分布，即 $\theta \sim \mathcal{N}(0, \sigma^2)$ 其中，$p(\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\theta^2}{2\sigma^2}}$ 。因此，上述优化函数可转换为：

$$
\begin{aligned}
&\arg\min_{\theta} \ f(\theta) - \sum_{i=1}^{m} \log p(\theta) \\
&= f(\theta) - \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\theta_i^2}{2\sigma^2}} \\
&= f(\theta) - \sum_{i=1}^{m} \log \frac{1}{\sqrt{2\pi}\sigma} + \frac{1}{2\sigma^2} \sum_{i=1}^{m} \theta_i^2 \\
&\Leftrightarrow \arg\min_{\theta} f(\theta) + \lambda \|\theta\|_2^2
\end{aligned}
$$

从上面的数学推导可以看出，$L2$ 正则化可以看成是：通过假设权重参数 θ 的先验分布为正态分布，由最大后验概率估计导出。

49

直观展示

权重衰减

Ridge Regression

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}\{t_n - y(x_n, \mathbf{w})\}^2$$

Regularized Regression
(L2-Regularization or Ridge Regularization)

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=0}^{N-1}(t_n - y(x_n, \mathbf{w}))^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

$$\nabla_{\mathbf{w}}(E(\mathbf{w})) = 0$$

$$\nabla_{\mathbf{w}}\left(\frac{1}{2}\sum_{i=0}^{N-1}(y(x_i, \mathbf{w}) - t_i)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2\right) = 0$$

$$\nabla_{\mathbf{w}}\left(\frac{1}{2}\|\mathbf{t} - \mathbf{X}\mathbf{w}\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2\right) = 0$$

$$\nabla_{\mathbf{w}}\left(\frac{1}{2}(\mathbf{t} - \mathbf{X}\mathbf{w})^T(\mathbf{t} - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}\right) = 0$$

$$\nabla_{\mathbf{w}}\left(\frac{1}{2}(\mathbf{t} - \mathbf{X}\mathbf{w})^T(\mathbf{t} - \mathbf{X}\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}\right) = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \nabla_{\mathbf{w}}\left(\frac{\lambda}{2}\mathbf{w}^T\mathbf{w}\right) = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{w} = 0$$

$$-\mathbf{X}^T\mathbf{t} + \mathbf{X}^T\mathbf{X}\mathbf{w} + \lambda\mathbf{I}\mathbf{w} = 0$$

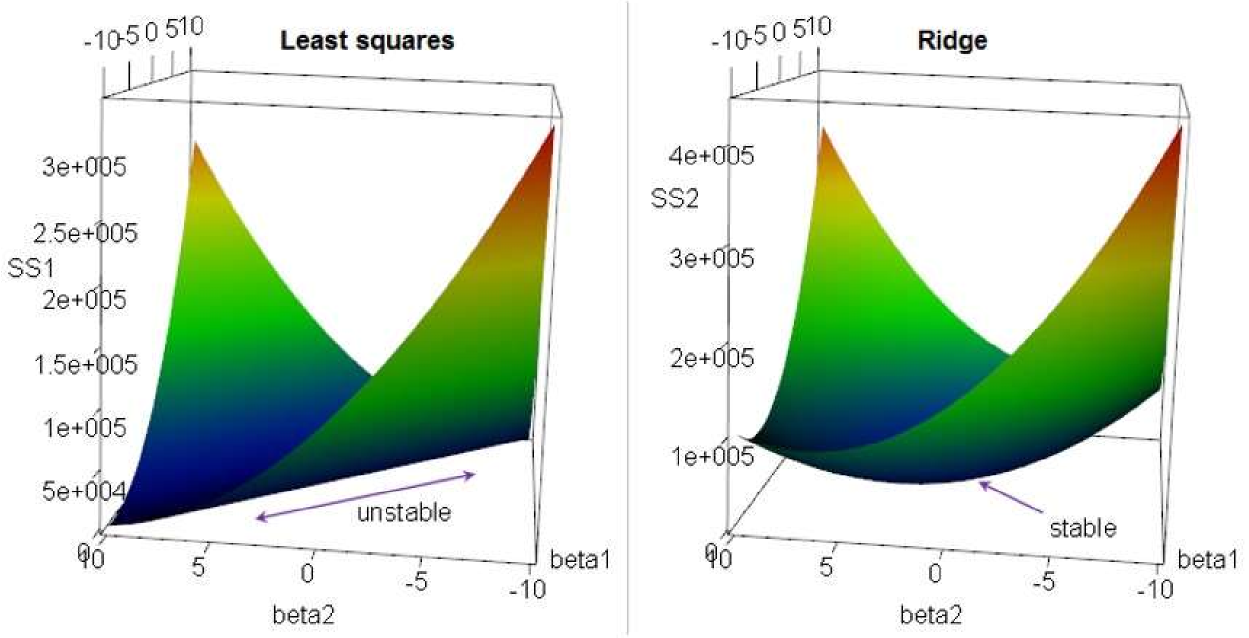
$$-\mathbf{X}^T\mathbf{t} + (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = 0$$

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\mathbf{w} = \mathbf{X}^T\mathbf{t}$$

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{t}$$

权重衰减

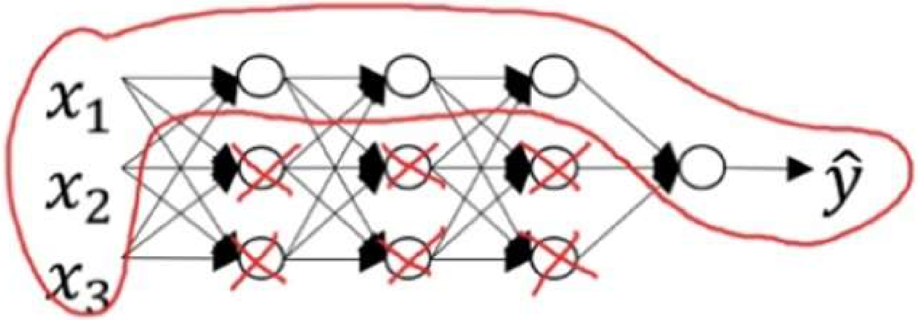
L2

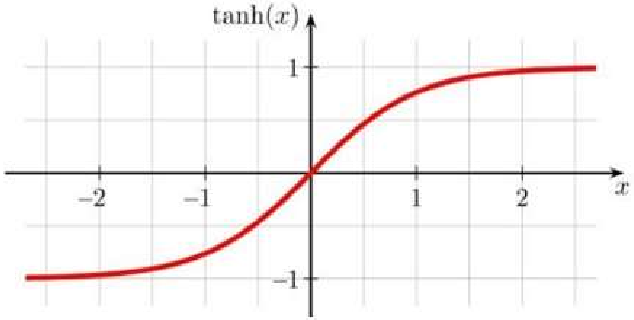


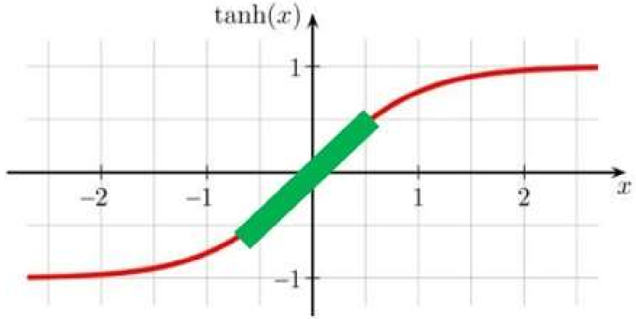
Least squares
SS1
3e+005
2.5e+005
2e+005
1.5e+005
1e+005
5e+004
unstable
beta1
beta2
Ridge
SS2
4e+005
3e+005
2e+005
1e+005
stable
beta1
beta2

权重衰减

神经网络示例







WX+b

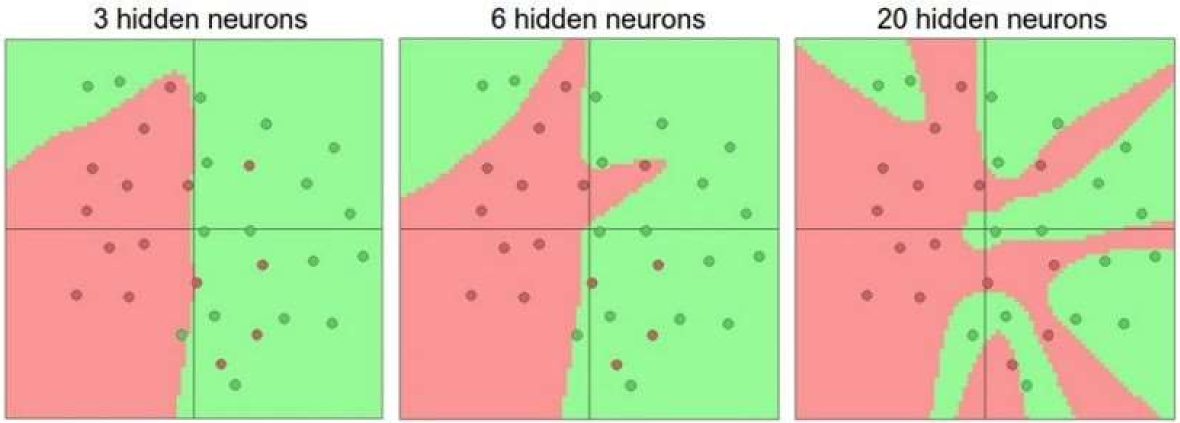
当正则化使W很小时，对于激励函数而言，它在零点附近相当于线性函数，
因此减小了模型的复杂度

53

权重衰减

神经网络示例

- 隐藏层的不同神经元个数

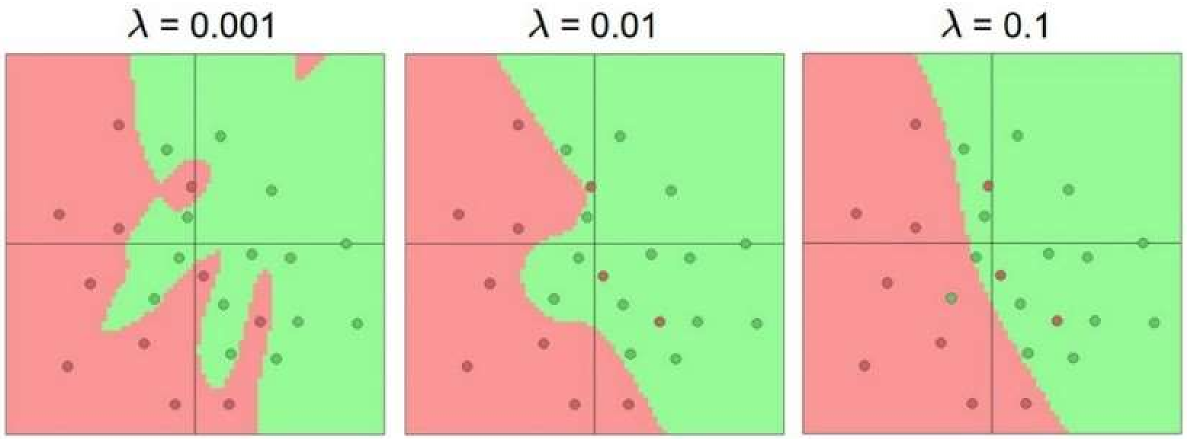


54

权重衰减

神经网络示例

- 不同的正则化系数



THANK YOU
Q&A