

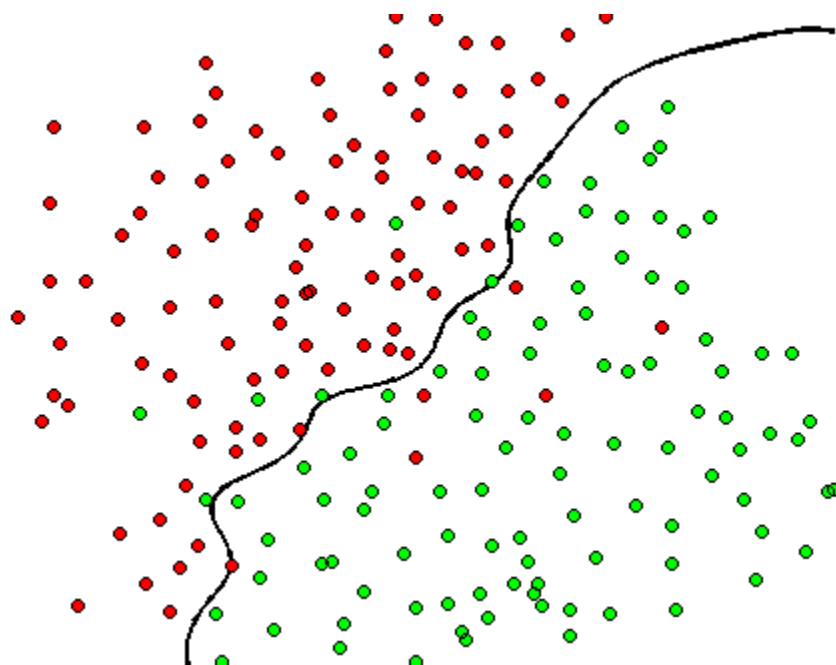
# 法律声明

---

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。



关注 **小象学院**



## 分类模型 (2)

--Robin

# 目录

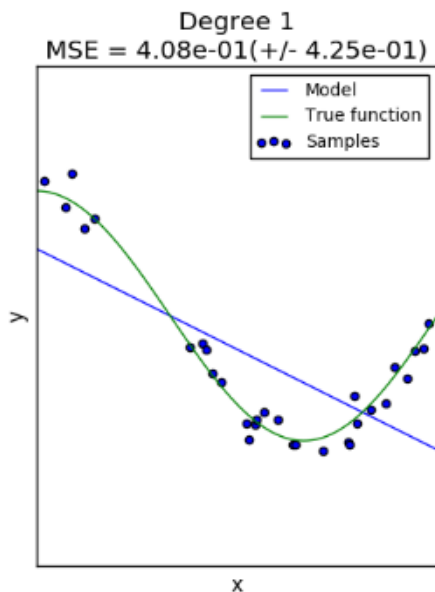
---

- 逻辑回归
- 正则化
- 支持向量机

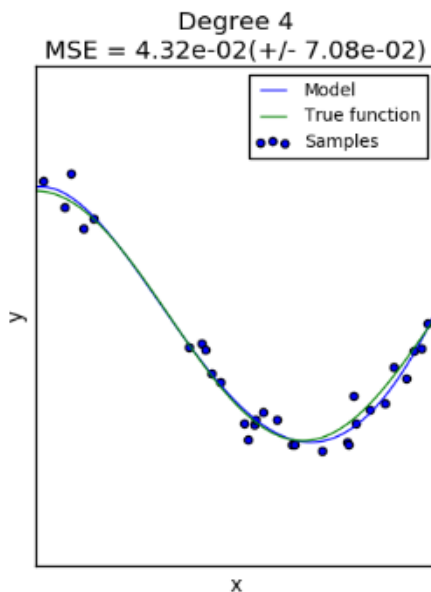
# 正则化

- 过拟合

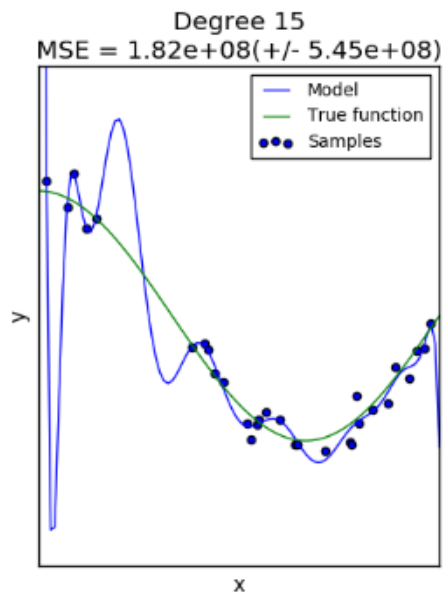
- 是指在调适一个统计模型时，使用**过多**参数。模型对于训练数据拟合**程度过当**，以致太适应训练数据而非一般情况。
- 在训练数据上表现非常好，但是在测试数据或验证数据上表现很差。



欠拟合



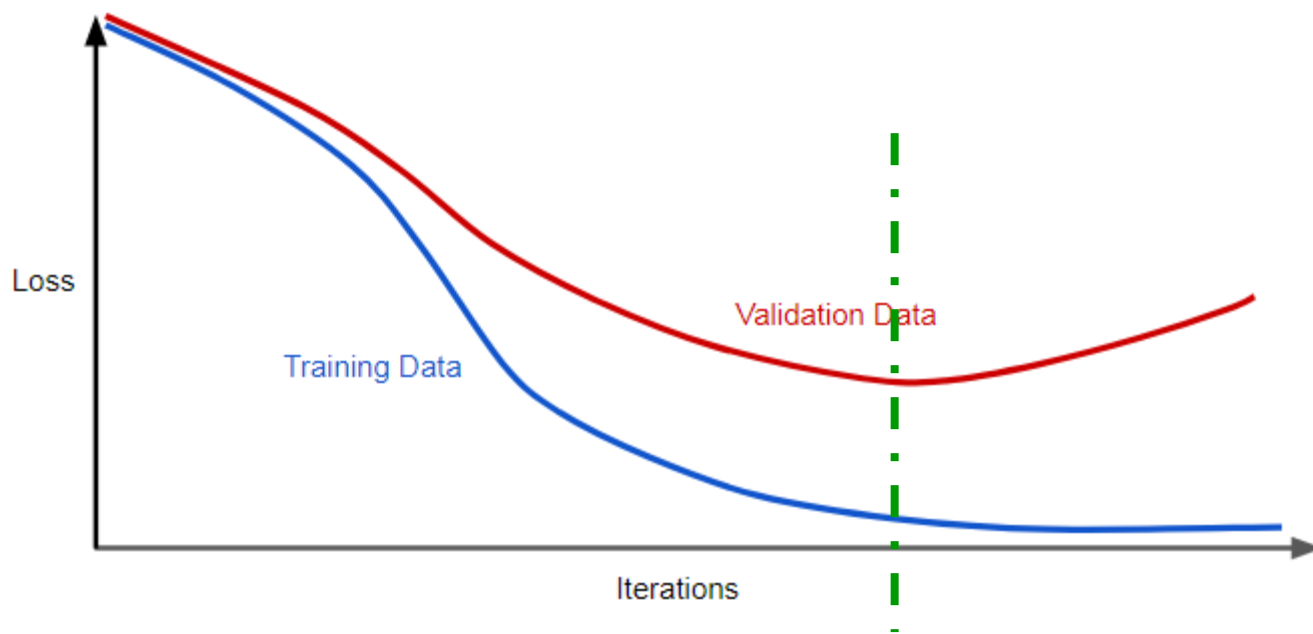
“刚刚好”



过拟合

# 正则化

- 正则化

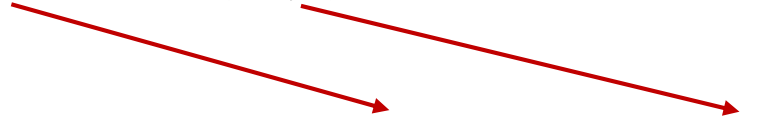


# 正则化

---

- 正则化

- 控制模型复杂度，模型复杂度越高，越容易过拟合
- 平衡损失函数与模型复杂度


$$\text{minimize: } Loss(Data | Model) + complexity(Model)$$

- 衡量模型复杂度

- 模型学习得到的权重越大，模型复杂度越高
- L2 正则化
  - $complexity(model) = \text{sum of the squares of the weights}$
  - 惩罚特别大的权重项

# 正则化

---

- 正则化 A Loss Function with  $L_2$  Regularization

$$L(\mathbf{w}, D) + \lambda || \mathbf{w} ||_2^2$$

Where:

$L$ : Aim for low training error

$\lambda$ : A scalar value that controls how weights are balanced

$\mathbf{w}$ : Balances against complexity

$_2^2$ : The square of the  $L_2$  normalization of  $\mathbf{w}$

- $\lambda$  值越大，正则化越强，表示需要更多关注模型的复杂度，适用于测试集中的样本与训练集中的样本相差比较大时；
- $\lambda$  值越小，正则化越弱，表示需要更多关注损失函数，适用于测试集中的样本与训练集中的样本相差不是很大

# 正则化

---

- 正则化
- 例子：

$$L_2 \text{ regularization term} = \|\mathbf{w}\|_2^2 = w_1^2 + w_2^2 + \dots + w_n^2$$

$$\{w_1 = 0.2, w_2 = 0.5, w_3 = 5, w_4 = 1, w_5 = 0.25, w_6 = 0.75\}$$

$$\begin{aligned} & w_1^2 + w_2^2 + \mathbf{w_3^2} + w_4^2 + w_5^2 + w_6^2 \\ &= 0.2^2 + 0.5^2 + \mathbf{5^2} + 1^2 + 0.25^2 + 0.75^2 \\ &= 0.04 + 0.25 + \mathbf{25} + 1 + 0.0625 + 0.5625 \\ &= 26.915 \end{aligned}$$

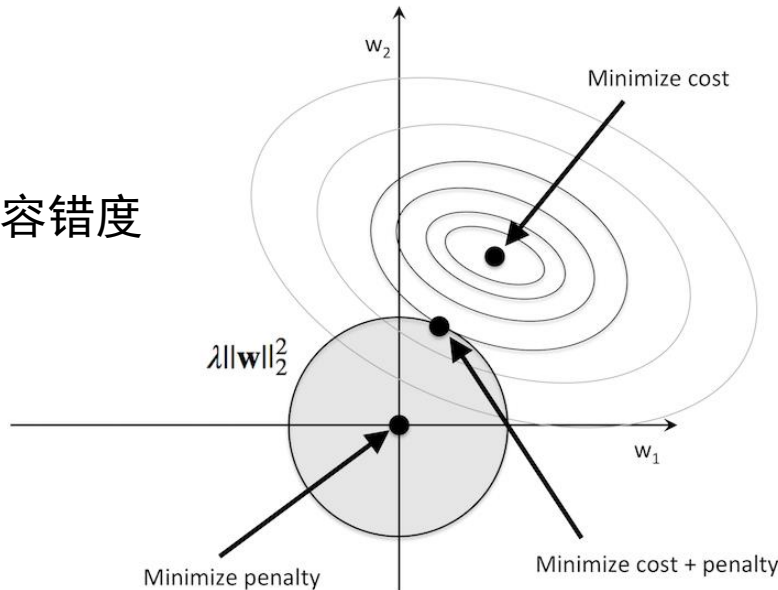


# 正则化

- 正则化

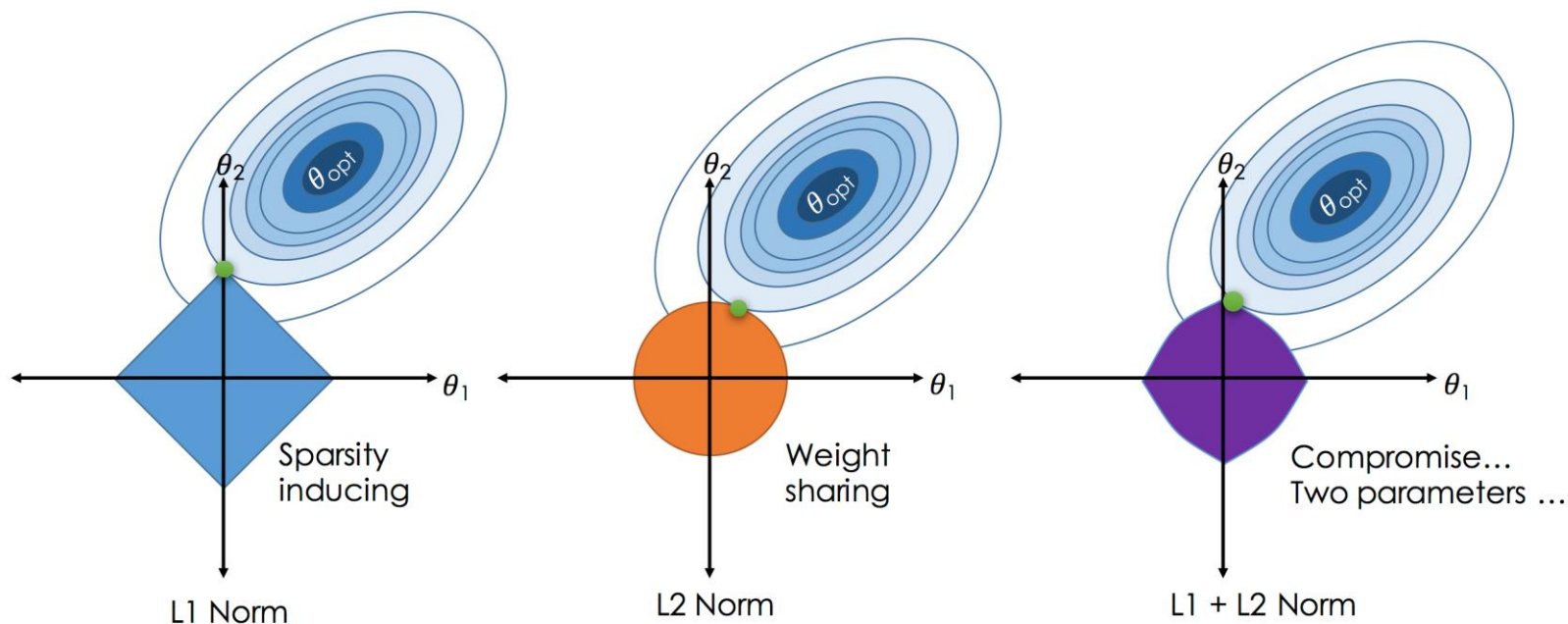
$$L = - \sum_{i=1}^n \log g(y_i z_i) + \frac{\lambda}{2} \sum_{k=1}^l w_k^2 \quad \text{正则项}$$

- 注意：sklearn中，logistic regression的参数C是正则项系数的倒数， $C=1/\lambda$
- 正则项中的C值决定了正则化的强度
- $\lambda$  值越大（C值越小），正则化越强
  - 对于单个样本的错误分类具有较强的容错度
- $\lambda$  值越小（C值越大），正则化越弱
  - 尽可能地去拟合训练样本的数据
  - 对于分类器来说，每个样本都很重要



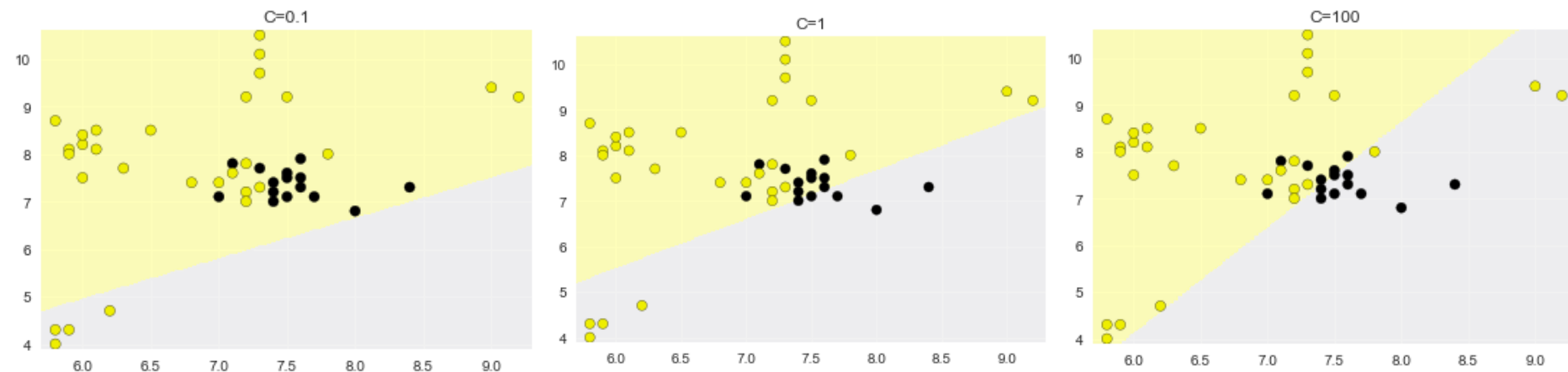
# 正则化

- 正则化



# 正则化

- 正则化



# 联系我们

---

小象学院：互联网新技术在线教育领航者

— 微信公众号：**小象学院**

