

欠拟合与过拟合

1. 概念

1.1. 欠拟合：模型过于简单，未能充分捕获数据的特征，表现为训练集上效果不好

1.2. 过拟合：模型过于复杂，过分捕获数据特征，把一些特殊特征作为共性特征处理，表现为训练集效果非常好，测试集效果不好

2. 解决方案

2.1 欠拟合解决方案：

- *使用复杂模型
- *添加新特征（多项式扩展）
- *减少正则化系数

多项式扩展

导入：from sklearn.preprocessing import PolynomialFeatures

- 多项式扩展使用展示

```

1 from sklearn.preprocessing import PolynomialFeatures
2
3 X = np.array([[1, 2], [3, 4]])

```

```

4 # degree: 扩展的阶数。阶数越高，则输出特征越多。
5 # include_bias: 是否包含偏置，默认为True。
6 poly = PolynomialFeatures(2, include_bias=True)
7 # 对输入数据进行转换。
8 # 相当于调用fit之后，再调用transform。
9 # poly.fit(X)
10 # r = poly.transform(X)
11 r = poly.fit_transform(X)
12 print("转换之后的结果：")
13 print(r)
14 print("指数矩阵：")
15 print(poly.powers_)
16 print("输入的特征数量：", poly.n_input_features_)
17 print("输出的特征数量：", poly.n_output_features_)
18
19 for x1, x2 in X:
20     for e1, e2 in poly.powers_:
21         print(x1 ** e1 * x2 ** e2, end="\t")
22     print()

```

相当于先调用fit再调用transform

转换之后的结果：

```
[[ 1.  1.  2.  1.  2.  4.]
 [ 1.  3.  4.  9. 12. 16.]]
```

指数矩阵：

```
[[0 0]
 [1 0]
 [0 1]
 [2 0]
 [1 1]
 [0 2]]
```

输入的特征数量： 2

输出的特征数量： 6

1	1	2	1	2	4
1	3	4	9	12	16

2.2 过拟合解决方案：

*增加数据集

*降低模型复杂度

*增大正则化系数

*使用集成方法

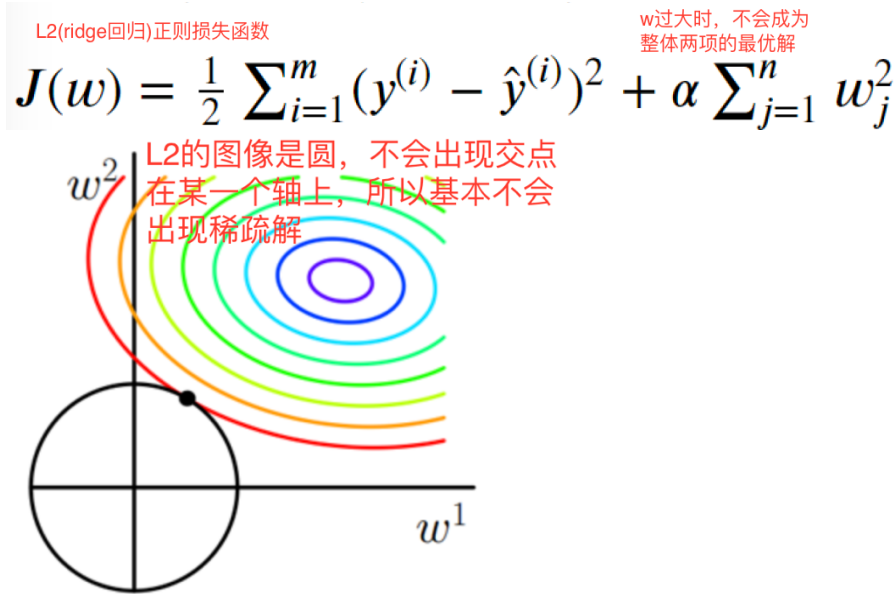
增加正则化系数

L2 正则化

导入：from sklearn.linear_model import Ridge

原理：通过加入系数 w^2 的和，使损失函数变为两项，如果 w 过大，则不会成为整体的最

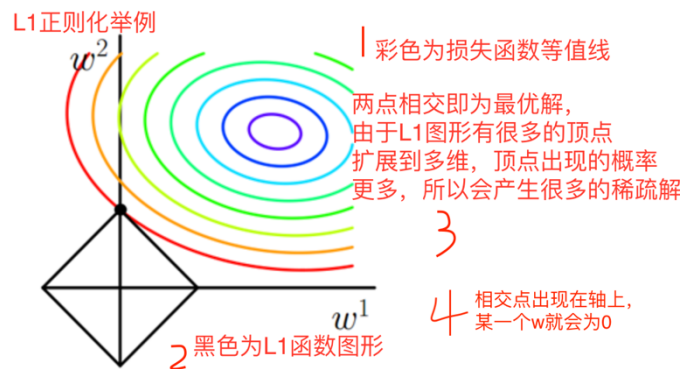
优解，公式如下图



L1 正则化

导入：from sklearn.linear_model import Lasso

$$J(w) = \frac{1}{2} \sum_{i=1}^m (y^{(i)} - \hat{y}^{(i)})^2 + \alpha \sum_{j=1}^n |w_j|$$



elasticNet

导入：from sklearn.linear_model import ElasticNet

$$J(w) = \frac{1}{2} \sum_{i=1}^m (y^{(i)} - \hat{y}^{(i)})^2 + \alpha(p \sum_{j=1}^n |w_j| + (1-p) \sum_{j=1}^n w_j^2)$$

• p: L1正则化的比重 (0 ≤ p ≤ 1)。

L_1 L_2