 # L1 (lasson regularisation)

什么是正则化：

正则化是用于减少模型过拟合的技术。

过拟合是指模型在给出的训练数据上表现的过好，但是在未见过的新数据上**表现不佳**。

- 过拟合可能是因为模型过度学习了训练数据中的细节和噪音。
- 过拟合的模型往往能获得很高的训练准确性，但是泛化能力差。

泛化（generalization）是指模型对未见过的数据的处理能力。一个具有良好泛化能力的模型在训练集上的表现与在测试集或者实际应用中的表现相似。这意味着该模型能够有效地适应新的情况。

为了训练模型，需要使用一个损失函数计算模型预测值和真实值之间的差异。优化模型就是为了优化损失函数的值。

正则化通过向损失函数中添加一个额外的项来实现，称为正则项。

简单来说，**只要可以减少泛化误差而不是减小训练误差的方法，都可以称为正则化方法**。

L1 lasso正则化

Lasso正则化这个方法向损失函数中添加的正则项**与模型参数的大小有关**。

L1正则化项的值是**模型权重的绝对值之和**。

假设在一个模型中使用的损失函数为MSE，那么带有L1正则化的损失函数通常表示为：

$$
\begin{gathered}
L(w) = MSE + \lambda \sum_{j=1}^{n} |w_j| \\
L(w)\text{是包含}L1\text{正则化的损失函数。} \\
MSE\text{是原始损失函数，代表未加正则项的模型真实损失。} \\
\lambda\text{是正则化的强度参数，决定了正则化项对于总损失的贡献大小。} \\
\sum_{j=1}^{n} |w_j|\text{是模型所有权重的绝对值之和。} \\
w_i\text{表示模型的第}i\text{个权重。}
\end{gathered} \tag{1}
$$

下面让我们具体看看加入L1正则项之后怎么进行优化的，在这里选择用梯度下降进行演示：

$$
\begin{gathered}
L = \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i W)^2 + \lambda \sum_{j=1}^{n} |w_j| \\
W\text{：模型的权重或系数向量，形状通常是}p * 1\text{，其中}p\text{是特征的数量} \\
w_i\text{：权重向量}W\text{中的第}i\text{个元素} \\
X_i\text{：表示第}i\text{行样本数据，形状是}1 * p
\end{gathered} \tag{2}
$$

对加入正则项的损失函数对w_i进行偏导，先对MSE进行：

$$
\begin{gathered}
\frac{\partial MSE}{\partial w_j} = \frac{\partial \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i W)^2}{\partial w_j} = -\frac{1}{n} \sum_{i=1}^{n} X_{ij} (y_i - X_i W) \\
X_{ij}\text{：表示第}i\text{个样本的第}j\text{个特征值}
\end{gathered} \tag{3}
$$

接下来对正则项进行偏导，这里对w_j的导数是一个次梯度，因为如果在绝对值为0处不可导：

$$
\begin{gathered}
\frac{\partial \lambda \sum_{j=1}^{n}|w_j|}{\partial w_j} \\
\Rightarrow \\
\frac{\partial \lambda |w_j|}{\partial w_j} = \begin{cases} -\lambda & \text{if } \beta_j < 0 \\ \lambda & \text{if } \beta_j > 0 \\ [-\lambda, \lambda] & \text{if } \beta_j = 0 \end{cases} \\
\Rightarrow \\
\lambda \, sign(W)
\end{gathered} \tag{4}
$$

$sign$：符号函数

因此，加入了正则项的损失函数的导数实际上是剧方误差部分的导数和正则化部分的**次梯度**的结合。

在实际的梯度下降算法里面，我们通常会先对MSE部分计算并对权重进行最小化更新，然后会使用**软阈值**来处理L1正则化部分。

假设我们现在有一个线性模型，有两个特征，并且提供了两条数据。两个特征分别有两个权重与之对应：

两条数据：

X1	X2	Y
1	2	5
3	4	6

假设初始化权重为：

$$
w_1 = w_2 = 0.5 \tag{5}
$$

学习率为：

$$
learning\_rate = \alpha = 0.1 \tag{6}
$$

正则化参数：

$$
\lambda = 1 \tag{7}
$$

先分别对单个权重计算MSE梯度：

$$
\frac{\partial MSE}{\partial w_1} = \frac{\partial \left(-\frac{1}{2} * (1 * (5 - 1 * 0.5) + 1 * (5 - 2 * 0.5)) + (3 * (6 - 3 * 0.5) + 3 * (4 * 0.5))\right)}{\partial w_1} = -\frac{11}{2} \tag{8}
$$

$$
\frac{\partial MSE}{\partial w_2} = \frac{\partial \left(-\frac{1}{2} * (2 * (5 - 1 * 0.5) + 2 * (5 - 2 * 0.5)) + (4 * (6 - 3 * 0.5) + 4 * (4 * 0.5))\right)}{\partial w_2} = -\frac{17}{2} \tag{9}
$$

得到梯度之后，在加入正则梯度之前，先对权重进行更新：

$$
\begin{aligned}
w_1^{new} &= w_1 - \alpha \frac{\partial MSE}{\partial w_1} = 0.5 - 0.1 * (-\frac{11}{2}) = 1.05 \\
w_2^{new} &= w_2 - \alpha \frac{\partial MSE}{\partial w_2} = 0.5 - 0.1 * (-\frac{17}{2}) = 1.35
\end{aligned} \tag{10}
$$

使用新计算的权重进行正则项的**软阈值**操作：

$$w_j^{new} = sign(w_j^{old}) * max(0, |w_j^{old}| - \alpha * \lambda)$$

$$\begin{gathered} w_j^{old}：是根据原损失函数更新后的权重 \\ sign(w_j^{old})：if\ w_j^{old} > 0\ then\ sign(w_j^{old}) = 1; \\ if\ w_j^{old} < 0\ then\ sign(w_j^{old}) = -1; \\ if\ w_j^{old} = 0\ then\ sign(w_j^{old}) = 0 \end{gathered} \tag{11}$$

我们在上面对加入正则项的损失函数已经计算了求导，和这里的软阈值操作相比，可以发现逻辑并没有发生改变。仍然是对初始权重减去学习率和总损失梯度的乘积。

但是值得注意的是，在这里额外增添了一个ReLU，这是为了实现L1正则化能够进行特征选择的特性。

权重经过梯度更新和阈值减少之后减小至0或以下，则会认为不足以影响模型的决策过程，我们就要从模型中剔除这个特征，所以将其系数（权重）设置为0.

$$w_1^{new} = sign(w_1^{old}) * max(0, |w_1^{old}| - \alpha * \lambda) = sign(1.05) * max(0, |1.05| - 0.1 * 1) = 0.95 \tag{12}$$

$$w_2^{new} = sign(w_2^{old}) * max(0, |w_2^{old}| - \alpha * \lambda) = sign(1.35) * max(0, |1.35| - 0.1 * 1) = 1.25 \tag{13}$$

经过一轮迭代，这就是我们最终获得的更新的权重值。

L1正则化的关键特性是**促使一些权重减少到0**，这样模型就只有剩余的非零权重进行预测。

这使得L1正则化成为一种**特征选择**的方法。能够降低模型的复杂度并且提高模型泛化的能力。

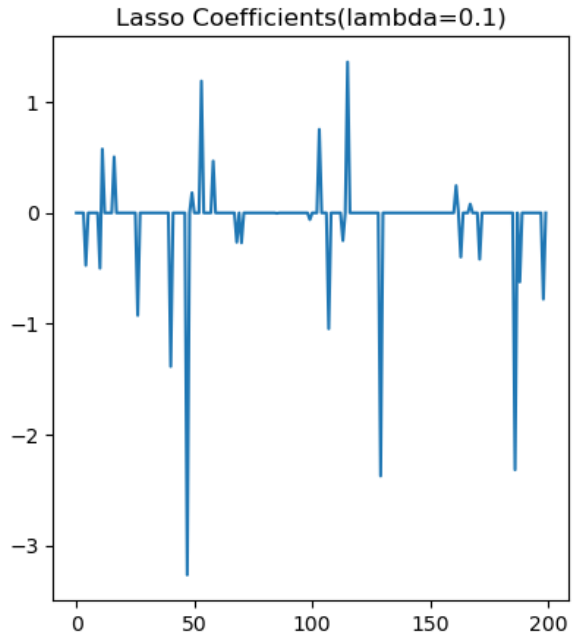
L1正则化能够减少过拟合的原因：

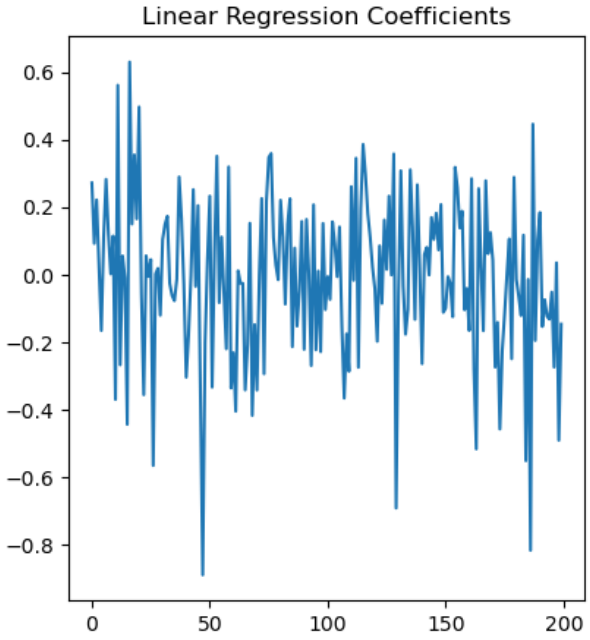
1. 通过对损失函数添加一个正则项，实际上是对模型训练增加了一个约束。使得模型的训练中去缩小权重值。过大的权重值会让模型对某些特征过于敏感导致过拟合。
2. 通过对模型权重的约束，能够让模型提升对数据噪音，细节的容忍性，降低模型的复杂度。
3. L1正则化能够返回稀疏权重。即对模型进行特征选择，迫使模型选择使用那些最具有影响力的特征。

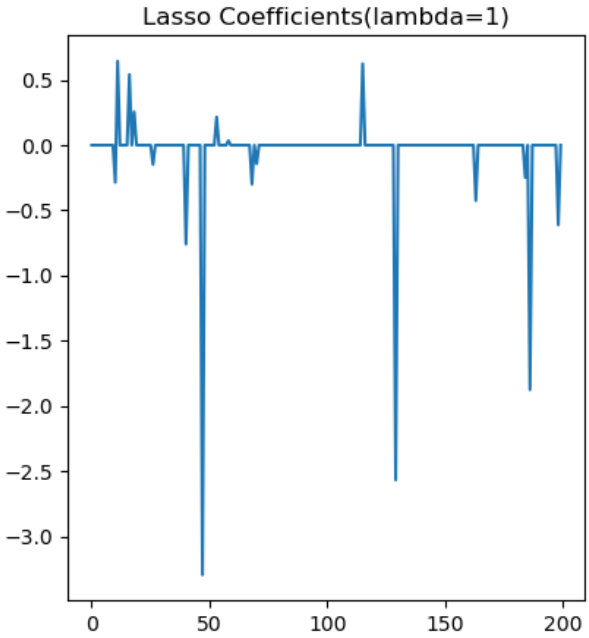
在代码中使用L1正则化，我们需要手动设施正则化参数，通常如果设置越大的正则化参数，会导致模型的权重越稀疏，意味着更多的模型特征的系数会被设置为0.下面提供了几幅plot进行对比（X轴代表了模型的每个系数，Y轴代表了对应系数的数值）。

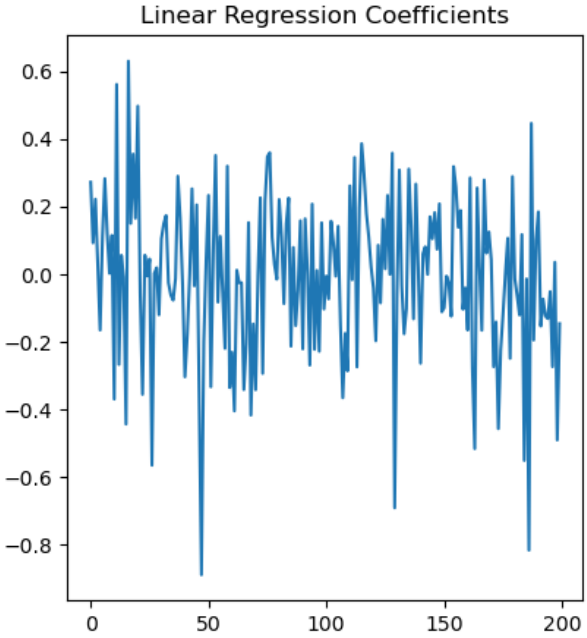
每个plot的左侧subplot是使用了L1正则项，右侧的subplot是简单了线性回归系数。

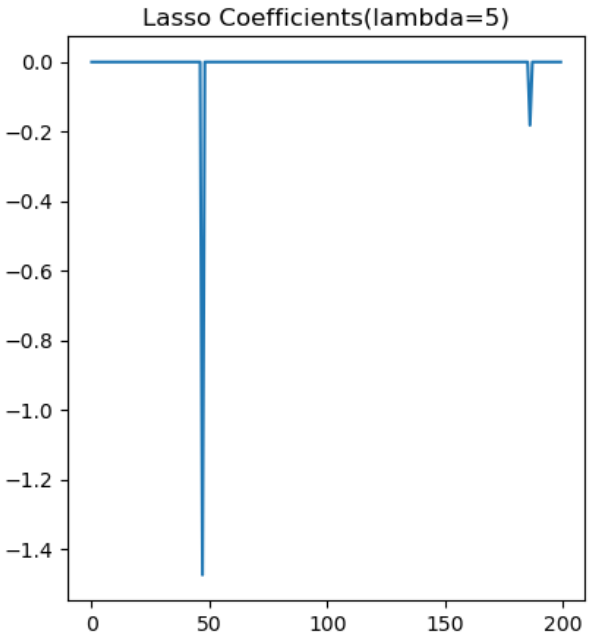
可以看到，随着正则化参数的增加，越来越多的特征被舍弃了。选择合适的正则化参数很重要。

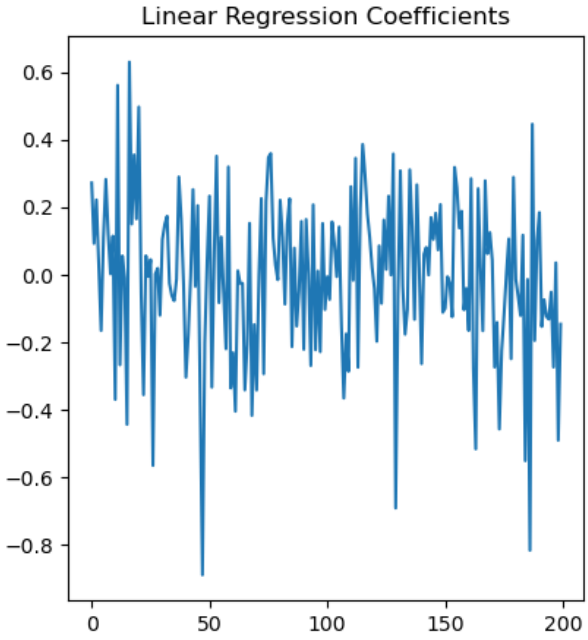

Lasso Coefficients(lambda=0.1)
1
0
−1
−2
−3
0
50
100
150
200


Linear Regression Coefficients
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0
50
100
150
200


Lasso Coefficients(lambda=1)
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
0
50
100
150
200


Linear Regression Coefficients
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
0
50
100
150
200





L1正则化实际上与拉格朗日对偶是相关的，如果有时间，可以尝试从拉格朗日对偶的数学逻辑方面进行解释。