多项式回归原理

【参考资料】

机器学习笔记八之多项式回归、拟合程度、模型泛化

scikit-learn 官网

多项式回归是普通线性回归的扩展版本，使用多项式的组合来对数据分布进行拟合。

比如对于一维变量，可以使用二次多项式来进行拟合：

$$y = ax^2 + bx + c$$

在scikit-learn中，会使用degree参数来控制多项式的最高维度，用来拟合的多项式是各种特征多项式的组合。

scikit-learn官网的解释如下：

> Generate a new feature matrix consisting of all polynomial combinations of the features with degree less than or equal to the specified degree. For example, if an input sample is two dimensional and of the form [a, b], the degree-2 polynomial features are [1, a, b, a^2, ab, b^2].

以二维变量为例，普通线性回归的做法是

$$\hat{y}(w, x) = w_0 + w_1 x_1 + w_2 x_2$$

而对于degree为2的多项式回归来说，模型如下

$$\hat{y}(w, x) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2$$

相当于是用原始特征构造了新的特征

$$z = \left[x_1, x_2, x_1 x_2, x_1^2, x_2^2\right]$$

需要注意的是，**新构造的特征z需要进行标准化处理**，因为经过多项式操作后不同维度之间的尺度差异会非常大。

所以多项式回归模型可以转化为基本的线性回归的形式

$$\hat{y}(w, x) = w_0 + w_1 z_1 + w_2 z_2 + w_3 z_3 + w_4 z_4 + w_5 z_5$$

新特征中的每一项都是原始特征的多项式组合，同时多项式的最高次幂是degree参数。

原始特征依然是2维，degree为3时，构造的新特征有10个，分别是$1, x_1, x_2, x_1^2, x_2^2, x_1 x_2, x_1^3, x_2^3, x_1^2 x_2$ ，$x_1 x_2^2$。

其他情况以此类推，degree的值越大，用来拟合的多项式次数越高，模型越复杂，所以需要注意过拟合的情况。