线性回归——普通最小二乘法 (OLS)

1 线性回归

回归（regression）：描述变量之间在统计学意义上的数量关系。

线性回归（linear regression）：当因变量与自变量之间呈线性相关，则称为线性回归。对于一组数据 $(X_1, y_1), \cdots, (X_j, y_j), \cdots, (X_m, y_m)$，其中 $X = (x_1, \cdots, x_i, \cdots, x_n)$，$x_i$ 代表各特征，我们希望得到一条直线

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

来拟合这组数据，并具有预测的能力。其中，$\hat{y}$ 表示预测值，a_i，$i > 0$ 是代表对应特征影响力或权重的参数，a_0 表示直线的截距。

2 普通最小二乘法

普通最小二乘法（Ordinary Least Squares）通过最小化数据集中因变量与预测值之间的均方误差来进行参数估计。对于单变量回归 (特征 x 只有一个)，均方误差可以表示为均方残差：

$$S(a) = \frac{1}{m} \sum_{j=1}^{m} (y_j - \hat{y})^2$$

我们都在高数里都学过求解多元函数极值，分别令 $\frac{\partial S}{\partial a_1}$ 和 $\frac{\partial S}{\partial a_0}$ 等于 0，易解得：

$$a_1 = \frac{\sum_{j=1}^{m} (x_j - \overline{x})(y_j - \overline{y})}{\sum_{j=1}^{m} (x_j - \overline{x})^2}$$

$$a_0 = \overline{y} - a_1 \overline{x}$$

这里的 x_j 表示第 j 个数据，$\overline{x}$ 和 $\overline{y}$ 表示数据均值。

对于多变量回归，为了方便，我们添加 $x_0 = 1$ 对应 a_0，则回归方程可以表示为

$$\hat{y} = XA^T$$

其中 $A = (a_0, a_1, \cdots, a_n)$

此时，我们用欧氏距离表示误差：

$$S(a) = \frac{1}{m}\sum_{j=1}^{m} ||y_j - X_j A^T||^2$$

与单变量类似，分别令 $\frac{\partial S}{a_i} = 0$，可得正规方程：

$$X^T X A^T = X^T y$$

当 X 列满秩时（绝大多数情况下都满足，一般训练数据集数量 m 会远大于特征数量 n），可解得：

$$A^T = (X^T X)^{-1} X y$$

列满秩条件是为了保证 $X^T X$ 存在逆矩阵，以便直接求解。因此要求特征 x_i 不存在较强的相关性，如果存在某两个或多个特征强相关，要选择性剔除重复的。

对于数据集 X 为稀疏矩阵的情况，sklearn 另外提供了 LSQR 算法，即最小二乘 QR 分解算法。