机器学习项目1

2031536 苏宇

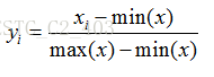
# 时间序列预测

## 数据预处理与特征选择说明

根据题意说明我们需要将段数据整合为一次特征数据，而每一段原始数据的特征维度高达数千维，包含了成百上千个传感器记录的数据，因此我们不能简单地将多段原始数据直接未经处理进行拼接作为一个特征数据，因为这样会导致一个特征数据的维度高达上万维，而我们全部的数据量也不过几千，这样直接训练会导致非常严重的欠拟合。

因此数据预处理的第一个目标就是，为原始数据降维，通过对数据的进一步观察分析可以发现一段数据的N维表示多个商品/用户/传感器在一段时间内销量/用电量/读数的变化，也就是说这些数据记录的是同一种类型的在不同物体中的多次记录，因此我们可以用一个均值来表示N维的一段数据。

但是由于每一个商品/用户/传感器的读数的数据大小差距比较大，所以我们不能直接对同一时间的所有读数进行均值操作，我们可以首先对一个物体所有时间段的读数进行归一化操作进行缩放如下图所示：



缩放之后可以保证同一时间段所有物体的读数在同一数据维度下，接着我们就可以使用算好的均值生成时间序列的特征值和标签。取t时刻对应的归一化后的原始记录数据的均值作为标签。时间段对应的数据均值作为特征值，除此之外还加入了这一时间段内数据均值的最大值和最小值，最终一个特征数据的维度为维，为选定的时间间隔，我们这里令，特征数据的维度远远低于数据量，可以实现很好的拟合效果。

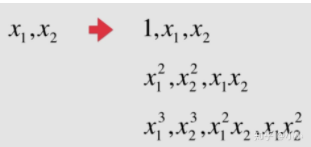
## 用到的模型说明

1. 线性回归

线性回归是很基础的机器学习算法，它属于机器学习中的回归问题，它可以在 2 个或者多个变量之间存在“线性关系”，然后就可以通过历史数据，摸清变量之间的“套路”，建立一个有效的模型，来预测未来的变量结果。线性回归的系数值可以通过朴素的矩阵的逆求解的方式快速算出，因此训练速度非常快。线性回归的系数计算公式是

1. 多项式回归

多项式回归是基于线性回归的变种回归模型，因为线性回归只能拟合多个数据点之间的线性关系，所以可以为数据添加新的特征，而这些新的特征是原有的特征的多项式组合，采用这样的方式就能解决非线性问题，这样的思路跟PCA这种降维思想刚好相反，而多项式回归则是升维，添加了新的特征之后，使得模型可以一定程度上拟合一些非线性或者多次性的情况，相对于线性回归来说可以更好地拟合更广泛的数据分布。如果我们选最大次数项为3，参数量为2，则在多项式回归下生成的新参数量如下图所示。

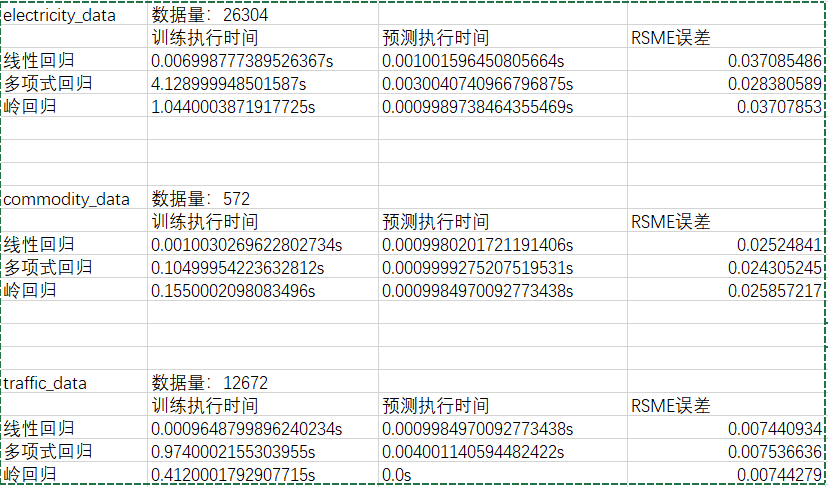


1. 岭回归

岭回归也是一个基于线性回归的变种回归模型，一般的线性回归的回归系数为，这里有一个前提就算要求矩阵X是列满秩的才能计算。即使 列满秩，但是当数据特征中存在共线性，即相关性比较大的时候，会使得标准最小二乘求解不稳定, 的行列式接近零，计算 的时候误差会很大。这个时候我们需要在cost function上添加一个惩罚项，称为L2正则化。此时的回归系数就变为了，其中就是我们选择的岭回归的lambda系数，具体系数值是多少由多次训练计算出最佳的值

## 实验结果与结论

在本实验中，我们针对商品销量，用电量，交通流量分别训练了三种回归模型，并统计了模型训练时的消耗时间，预测的消耗时间，以及最终预测的RSME误差。统计结果如下所示：



从统计结果中可以发现：多项式回归模型相对于岭回归和线性回归模型需要更长的训练时间，预测的速度也略慢于岭回归和线性回归，一方面是因为多项式回归的特征维度是线性回归和岭回归的数倍，另一方面也是因为多项式回归需要多次设置degree计算score，并取一个最优的score对应的degree作为模型的参数值，而线性回归只需要一次简单的矩阵运算就可以得到最优的参数。

# Logistic回归