摘要

本文分为两大部分。在第一部分，建立了线性回归和多项式回归模型对波士顿房价数据集进行了拟合，使用了最小二乘法进行求解，通过MSE以及R^2进行性能评估，并对结果进行了对比可视化展示。在第二部分，通过PCA对手写数据集进行了降维处理，从64维降低到了3维并进行了可视化展示，此时保留的信息只有40%。进一步，为了保留99%的有用信息，我们使用PCA降低到了41维，从而完成了目标。

我们选取了波士顿房价数据集中25%的数据作为样本，RM列（每栋住宅的房间数）作为自变量，房价作为因变量。分别使用线性回归模型、二次多项式、三次多项式、五次多项式以及十次多项式来对其进行拟合。

模型拟合的公式如下所示。

本节分别采用了MSE以及dd对不同拟合模型进行评估，结果如下。

可以看出随着多项式的最高次数依次递增，MSE逐渐下降，同时包含原数据的信息量越来越多。由于使用的波士顿数据集影响房价的因素有13种之多，因此仅仅使用一列数据作为自变量，拟合的效果不是很好。

本节使用了手写数据集，每幅图片都是8\*8的大小，对应的数据维度为64维。为了能够可视化展现，我们将其降低到了3维。

分析PCA后的三维数据，只包含了原始数据约40%的信息，表明PCA降维过程损失了过多的信息。因此，为了保留99%的有用信息，令PCA自动去选取保留哪些维度的信息，最终发现在保留41维的信息时，可以保留较多有用信息，并且在一定程度上降低了数据量的维度，有利于加速训练过程和防止过拟合现象的出现。