六、问题二的模型建立与求解

6.1 数据预处理

6.1.1 数据标准化优点

对于本题，自变量为母亲的身体指标与心理指标(年龄，婚姻状况，教育程度等)，因变量为婴儿的行为特征，需要对该分类变量进行数据标准化以及独热编码(one-hot)。

在机器学习中，标准化有助于提升模型精度，许多学习算法中的目标函数的基础都是假设所有的特征都是零均值并且具有同一阶数上的方差。如果某个特征的方程比其他特征大几个数量级，那么它就会在学习算法中占据主导地位，从而很难从其他特征中学习。

不仅如此，对数据进行标准化后，以线性模型为例，最优解的寻优过程会变得明显平缓，更容易正确地收敛到最优解。

6.1.2 数据标准化实现

(1) StandardScaler

核心思想为标准化数据通过减去均值后除以标准差：

$$X^* = \frac{X - \mu}{\sigma} \qquad (2-1)$$

其中X为样本数据，μ为样本数据的平均值，σ为样本数据的标准差，X^*是标准化后的样本数据，经过处理后的样本数据符合标准正态分布，即均值为0，标准差为1。

(2) MinMaxScaler

核心思想为对原始数据进行线性变换，使结果映射到$[0,1]$区间

$$X^* = \frac{X_i - \min\limits_{1\le j\le n} X_j}{\max\limits_{1\le j\le n} X_j - \min\limits_{1\le j\le n} X_j} \qquad (2-2)$$

其中X_i为样本数据，$\min X_j$为样本数据的最小值，$\max X_j$为样本数据的最大值，使得最终结果放缩到$[0,1]$区间。

表 2.1 机器学习特征处理方法

类	功能	说明
StandardScaler	无量纲化	标准化，将特征值转换至服从正态分布
MinMaxScaler	无量纲化	区间放缩，将特征值转换至$[0,1]$区间
Normalizer	归一化	将样本向量转换为“单位向量”
OneHotEncoder	哑编码	将定性数据编码为定量数据
Binarizer	二值化	将定量特征按阈值划分
Imputer	缺失值计算	计算缺失值，缺失值可填充为均值等
PolynomialFeatures	多项式数据转换	多项式数据

6.2 KNN 近邻算法

6.2.1 KNN 算法原理

K-最近邻算法是一种非参数、有监督的学习分类器对于新的输入实例，在训练数据集中找到与该实例最邻近的这 K 个实例，这 K 个实例的多数属于某个类，那么就把新加入的实例分类到该类中。

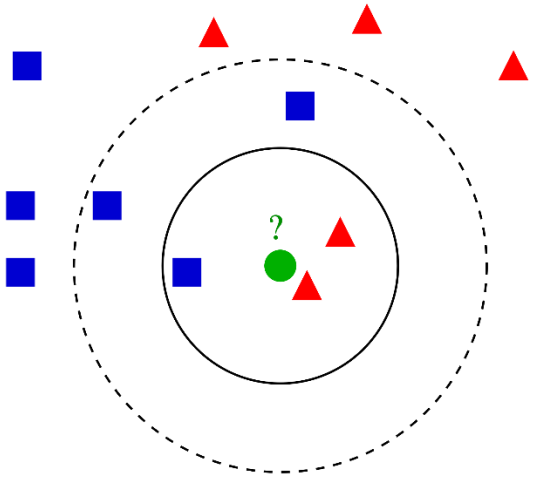

图 2.1 以小球分类解释 KNN

（1） 如果 K=3，绿色圆点的最邻近的 3 个点是 2 个红色小三角形和 1 个蓝色小正方形，基于统计的方法，判定绿色的这个待分类点属于红色的三角形一类。

（2） 如果 K=5，绿色圆点的最邻近的 5 个邻居是 2 个红色三角形和 3 个蓝色的正方形，基于统计的方法，判定绿色的这个待分类点属于蓝色的正方形一类。

6.2.2 KNN 算法流程

（1）对数据进行预处理即标准化归一化等

（2）计算测试样本点到其他每个样本点的距离

（3）对每个距离进行排序，选择出距离测试点数值最小的 K 个点

（4）对 K 个点所属类别比较，将测试样本归入在 K 个点中占比最高的一类

6.2.3 KNN 中 K 值的选取

选择一个最佳的 K 值取决于数据。一般情况下，在分类时，较大的 K 值能够减小噪声的影响，但会使类别之间的界限变得模糊。值得注意的是，一个较好的 K 值能通过各种启发式算法来获取。

6.2.4 KNN 中的距离度量

为了确定哪些数据点最接近给定查询点，需要计算查询点与其他数据点之间的距离。 这些距离度量有助于形成决策边界，而决策边界可将查询点划分为不同的区域。

（1） 欧几里得距离

这是最常用的距离度量，可以测量查询点和被测量的另一个点之间的直线：

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(y_i-x_i)^2} \qquad (2-3)$$

（2） 曼哈顿距离

它用于测量两点之间的绝对值大小，通常用网格可视化，

$$d(x,y)=\sum_{i=1}^{n}|x_i-y_i| \qquad (2-4)$$

6.2.5 KNN 模型求解

本文利用 Python 代码编写 KNN 算法，由于超参数 K 值，训练集的划分难以确定，故通过循环的方式控制超参数并进行枚举，以找到合适的超参数。

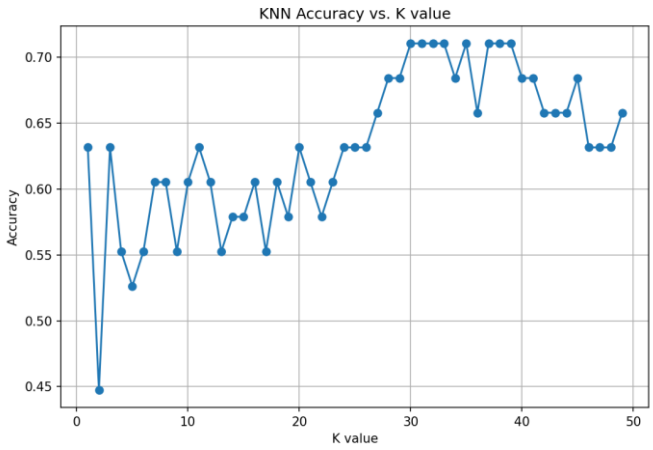

图 2.2 不同 K 值的测试集准确率

固定 K=30 后，以同样的方式寻找超参数测试集的占比

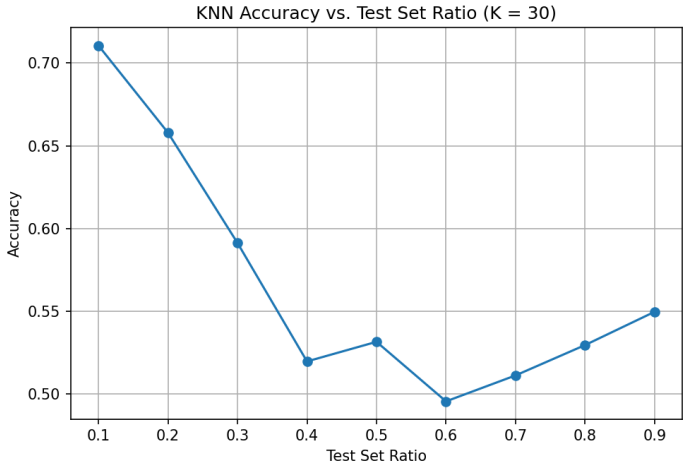

图 2.3 K=30 时不同测试集占比的准确率

此时的输出结果为：

表 2.2 KNN 模型评估结果

	准确率	召回率	精确率	F1
训练集	0.608187135	0.608187135	0.535612318	0.569597267
测试集	0.710526316	0.710526316	0.651251252	0.67959873

上表中展示了交叉验证集、训练集和测试集的预测评价指标，通过量化指标来衡量 K 近邻(KNN)的预测效果。

（1）准确率：预测正确样本占总样本的比例，准确率越大越好。

（2）召回率：实际为正样本的结果中，预测为正样本的比例，召回率越大越好。

（3）精确率：预测出来为正样本的结果中，实际为正样本的比例，精确率越大越好。

（4） F1：精确率和召回率的调和平均，精确率和召回率是互相影响的，虽然两者都高是一种期望的理想情况，然而实际中常常是精确率高、召回率就低，或者召回率低、但精确率高。若需要兼顾两者，那么就可以用 F1 指标。

表 2.3 预测(编号 391-410 号婴儿的行为特征信息

预测结果_Y	预测结果概率_安静型	预测结果概率_中等型	预测结果概率_矛盾型	母亲年龄	婚姻状况	教育程度	妊娠时间（周数）	分娩方式	CBTS	EPDS	HADS
中等型	0.35	0.5	0.15	29	2	4	40	1	7	15	12
矛盾型	0.35	0.25	0.4	29	2	3	42	1	9	14	12
中等型	0.2	0.55	0.25	23	2	2	38.5	1	7	12	7
中等型	0.15	0.75	0.1	27	2	3	36.3	1	8	4	5
中等型	0.3	0.65	0.05	36	2	4	39	1	6	6	8
中等型	0.25	0.6	0.15	30	2	5	41.2	1	5	8	5
中等型	0.3	0.55	0.15	28	2	2	40.6	1	8	11	9
中等型	0.25	0.7	0.05	32	2	5	37	1	3	6	7
矛盾型	0.2	0.3	0.4	28	2	5	38	1	7	11	5
中等型	0.3	0.65	0.05	31	2	4	42	1	4	5	8
中等型	0.25	0.6	0.15	25	2	2	40.5	1	16	22	15
中等型	0.1	0.9	0	27	2	5	40.4	1	4	6	10
中等型	0.2	0.6	0.2	33	2	5	39	1	6	6	4
安静型	0.4	0.4	0.2	25	2	3	39	1	0	4	5
中等型	0.2	0.75	0.05	28	2	2	41	1	9	6	5
安静型	0.4	0.4	0.2	31	2	3	39.5	1	1	4	4
中等型	0.35	0.6	0.05	26	2	2	37	1	4	9	14
安静型	0.6	0.3	0.1	26	2	5	39	1	0	3	3
安静型	0.55	0.4	0.05	27	2	5	41.2	1	0	0	4
中等型	0.25	0.7	0.05	31	2	5	38	1	3	7	7