基于多种机器学习算法的波士顿房价预测

□田润泽　郑州市郑东新区外国语中学

【摘要】　波士顿房价预测问题在人工智能领域中属于回归问题，回归问题是机器学习中很重要的一个研究方向。解决回归问题，较为常见的算法模型有 Ridge Regression 模型，基于集成学习方法的 Random Forest、AdaBoost 模型，基于深度学习的 DNN 模型等等。在具体问题中，选取的模型不同相应的效果也会不同，本研究根据“波士顿房价预测”这一实际问题，分别采用了不同的机器学习模型进行训练和测试，从多个角度对比了不同算法模型在房价预测这一回归问题上的综合表现，给出分析与总结。对上述不同算法模型的算法原理、实际表现以及特点进行了深入分析，总结了不同算法模型在波士顿房价预测这一实际问题上的表现，对不同模型的优缺点进行横向对比，对效果差异进行了分析与总结。

【关键字】　房价预测　Ridge Regression　Random Forest　AdaBoost　DNN

一、引言

对于回归模型的研究是机器学习算法研究领域的重要部分，除了传统的线性回归与 Ridge Regression 等算法，集成学习算法也得到了广泛应用，比较典型的代表是基于 Bagging 的 Random Forest 算法和基于 Boosting 的 AdaBoost 算法。此外，以神经网络（DNN）为代表的深度学习方法在回归问题的解决中也表现良好。

DNN 是机器学习中一类很重要的学习方法，可以认为是对神经元结构从信息处理角度经过一定的抽象处理之后得出的算法，针对具体问题建立数学模型，依据连接方式以及算法的不同划分成不同的网络结构。DNN 本质上是由大量具有一定运算规则的神经元及其连接关系构建的运算模型。在神经网络中，每个神经元会对输入信号通过激活函数进行非线性变换，也就是激活函数（activation function）。在连接关系中，每两个神经元之间有连接表示对于通过的数据进行加权处理，构建的即权重，权重参数可以看人工神经网络对于数据的记忆。人工神经网络本质上往往是对自然界存在的某种算法 / 函数的拟合，或者逻辑表达。

本研究依据波士顿房价预测问题，采用了上述多种机器学习算法进行建模，对模型进行训练和测试，通过对模型表现以及训练过程的总结，探究不同机器学习算法在这一具体问题中的应用。

1.1 数据集特征

波士顿房价数据集是机器学习算法探究的经典数据集。该数据集共 2920 条房价数据，按照 1：1 的比例划分成训练集和测试集，其中训练集的数据为 1460*80，80 列特征中包含标签列，测试集数据为 1460*79，79 列特征中不包含标签列，并且数值类型的特征占 35 列，类别类型特征占 44 列。

1.2 数据集预处理

1.2.1 缺失值处理

通过对训练数据和测试数据进行整体分析，在整个波士顿房价数据集中，含有缺失值的特征有 19 列。需要在数据预处理阶段对含有缺失值的特征进行处理，根据不同缺失程度，相应的处理方式不同，通常按照以下方式：对特征和标签进行关联度分析，对于在预测过程中提供帮助较小的缺失特征通常会采取直接删除的方式，如果缺失特征的缺失量较高（大于 15%）也会选择删除特征。本数据集中的 Alley 特征未缺失的数据占比只有 6.23%，应该直接删除。对于其他缺失量较小的特征，一般通过缺失值补全的方式将数据补充完整，数据补全方式依据具体的缺失值类型进行选择，常用的有取均值、中位数或取模等方式。此外，如果包含缺失值的特征只在一个或很少几个样本中是缺失的，并且样本数量较大的情况下，只删除这些样本即可。如果样本数较少，并且对样本具有较高精确度要求的情况下，还可以采用建模的方式对缺失值进行预测，即构建模型，通过样本不缺失的特征推断缺失特征的具体内容。此种数据补全方法较为复杂，

启示，分别从地震预防、地震救灾、震后恢复三个方面提出一些建议。日本处在的地震高发带的地震类型以板间地震为主，又因为日本是发达国家，其防震减灾方面的法律体系相对比较健全。而我国是发展中国家，且因其幅员辽阔，地震类型较为复杂、多变，板内地震多发。因此，我国要想建立一套完整的地阵防范应对措施难度更大，多借鉴学习日本这样国家的优秀的体系和成熟的经验，并完善适合我国的经验，再运用在自己身上。最后，面对地震这样的自然灾害，人类还需要再进一步探索其中的奥秘，让人与自然更好的和谐共处。

参 考 文 献

[1] 闫恩辉 , 杨锐 , 张晋辉 , 等 . 日本应急救灾体系 [J]. 国际地震动态 , 2016(1):14–23.
[2] 李玉江 , 吴荣辉 , 卢振恒 . 日本、美国等国家防震减灾法律法规制度比较研究 [J]. 国际地震动态 , 2015(2):15–22.
[3] 庄涛 . 我国地震防灾减灾科普教育的瓶颈及对策分析 [J]. 国际地震动态 , 2013(4):30–34.

因此前两种方法较为常用。

1.2.2 数据变换

进行完缺失值处理之后，将数据集导入，针对数据类型进行相应处理。需要将类别类型的特征转化成数值类型的特征。对类别类型的特征进行一元方差分析，以得到各个类别类型特征对房价方差的影响，从中挑出影响较大的类别特征，依据这些特征的取值将类别用相应的等级数字表示，按照这种方法将类别类型数据用数值型的有序变量表示出来。

1.2.3 数据归一化

数据归一化也是数据预处理阶段常用的数据处理方式，归一化主要是用来调整数据的分布，加快模型收敛速度，简化计算过程。

1.3 数据集划分

1–1460 条数据作为训练集，包含 79 个特征列和 1 个标签列；1461–2920 条数据作为测试集，包含 79 个特征列。

二、构建回归模型

2.1 构建 Ridge Regression 模型

2.1.1 基本原理

Ridge Regression 模型依据最小二乘法，将样本送入模型得到预测值，与相应的真实值进行比对，采用平方损失函数，该平方损失函数对应存在最小值，于是上述回归预测问题就转换成了函数求极值的问题。对于损失函数最小值的求解，采用了梯度下降算法，每次沿着梯度下降最快的方向优化损失函数，使模型参数得到优化 [1]。

2.1.2 Ridge Regression 模型在波士顿房价数据集上的应用

根据房价与房屋数据之间的关系映射到高维空间中即拟合出房价的高维曲线，将预测房价和实际房价采用平方损失函数进行计算，对损失函数通过梯度下降的算法进行优化，修改模型参数，即求得最小损失函数误差所对应的模型参数。Ridge Regression 模型在最小二乘法的基础上中加入一个小扰动 λI，将最小二乘法不能解决的问题解决掉，使不满秩的矩阵变为满秩矩阵，使得问题有稳定解，岭回归就是尽可能的在方差和误差之间找一个都比较小的权衡点。

2.1.3 结果分析

可以很明显看出，选取 10–20 之间的 Alpha 值时，损失可以下降到 0.135，如图 2–1 所示。

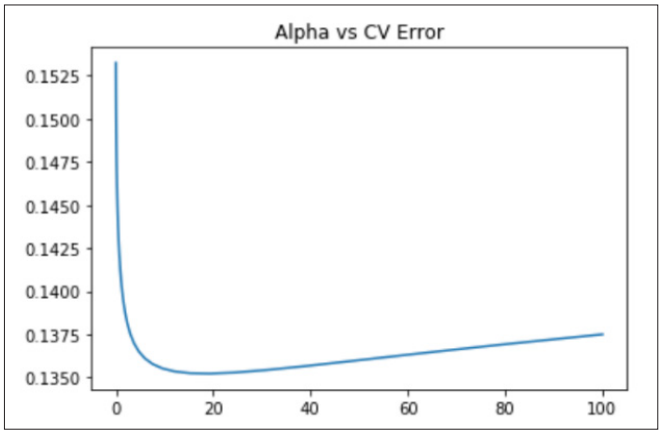

图 2–1　Ridge Regression 模型

2.2 基于 Bagging 的集成学习模型 Random Forest

2.2.1 Random Forest 基本原理

Bagging 通常用来提高学习算法的准确率，采用集成学习的思想，构造一系列预测函数，采用特定的集成方法将这些函数结合起来，对数据集要求有一定程度的“不稳定性”，也就是指函数系列中的每个模型函数所采用的数据集需要有轻微的变动，能够对结果产生较为显著的影响。常用的基学习器有决策树和神经网络等。Bagging 简单来讲基于概率论中有放回抽样的思想，在产生每个模型函数所需要的训练数据集时不是直接将原始数据集送入模型，而是从原始数据集中构建出一个新的数据集，与原始数据集有轻微的不同，构建的过程就是随机抽样的过程。从原始数据集中随机抽取出一个样本放到采样集，随后将抽取到的样本放回原始数据集，这样一来抽过的样本在下次抽样时仍有可能被选中。重复采样操作，通常一个采样集的大小选择和原始数据集大小一致。即原始数据集大小为 m，采样集大小也为 m。重复上述构建采样集的过程，就可以得到多个不同的采样集，分别作为每个基学习器的训练数据集。

经相关研究发现，选择这种有放回采样的方式会使每个采样集中的数据呈现以下规律，原始数据集中大约 63.2% 的数据在采样集中出现大于等于一次，大约 36.8% 的数据在采样集中不出现。原始数据集重复上述采样操作 K 次，得到 K 个采样集，基于每个采样集采用同种机器学习算法去训练基学习器，最后将训练好的基学习器通过一定的结合方法集成起来作为最终的预测模型，集成方法通常为求和求平均、简单投票法、加权投票法等，将构建的全部基学习器结合起来 [2]。

基于 Bagging 思想的 Random Forest 算法的基学习器是决策树，在上述基础的 Bagging 思想上加入了对特征进行随机选择这一操作，在依据原始数据产生采样集的同时，针对每个基学习器所采用的特征也进行了随机抽样，从原始特征集中按照一定比例随机抽取一些特征作为构建模型的特征。

2.2.2 Random Forest 模型在波士顿房价数据集上的应用

从原始数据集中随机抽取出一条房价数据，加入第一个采样集再将样本放回。原始数据集包含 1460 个样本，因此，采样集进行了 1460 次采样，得完整的采样集。重复上述操作 K 次，得到 K 个采样集，这种采样的方法相当于在一定程度上扩充了样本数据集。

针对 K 个采样集分别作为训练数据集构建决策树即基学习器，进行训练，将训练好的模型拿到测试集上进行测试，将 K 个模型的预测结果通过一定的集成方法综合起来得到最终的预测结果。

在上述过程中，会有 63.2% 的数据出现至少一次，37% 的数据一次都没出现，在这种情况下可以将未抽到的数据作为包外估计（类似于交叉验证的作用）在训练过程中检验模型的预测能力。包外估计能起到辅助剪枝的作用，通过包外估计比较是否剪枝的效果好坏，来选择最好的模型以及参数，同时是一种降低过拟合的有效方法。

2.2.3 结果分析

如图 2–2，由结果来看，当 Max Features 取值为 0.5 时，能够得到损失函数最小值，在 0.137–0.138 左右。

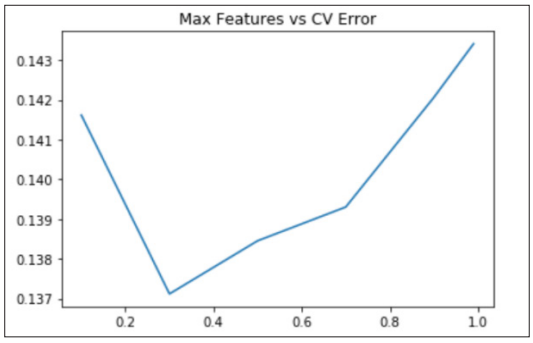

图 2–2　Random Forest 模型

2.3 构建集成学习模型 AdaBoost

2.3.1 AdaBoost 算法的基本原理

Adaboost 算法思想就是基于模型的残差，针对同一批训练数据，后一个模型的构建都是在前一个模型的基础上来完成的。根据前一个模型的预测结果对原始数据集分布进行相应的调整，将改变了的数据集作为下一个模型的输入，以此来得到不同的回归模型。最后根据特定的集成方法将所有的弱回归模型集成起来，得到一个强回归模型，也就是最终的模型。因此，对于 Adaboost 算法而言，并不一定要对样本空间分布进行精确的认识，每一次构建模型都需要在前一个模型的基础上调整样本空间分布，给模型判断正误不同的样本按照一定的规则分配不同的权重，对于正确的样本，降低其在数据集中的权重，对于错误的样本，增加其在数据集中的权重。这样一来，会使新构建的模型更加关注错误的样本，从而更容易对错误样本做出正确回归，而对于已经回归正确的样本则不需要模型过多的关注。Adaboost 是一种自适应的算法，能够自发调整弱学习算法，在经过多轮迭代之后，能够将错误率控制在理想范围内。增加其在最后的集成阶段，Adaboost 采用了加权平均的方式，给准确率较高的弱回归器分配较高的权重，给准确率较低的弱回归器分配较低的权重吧，取加权平均值作为最终结果。

2.3.2 在房价预测数据集上应用 AdaBoost 模型

整个模型构建流程如下：①将原始的房价数据作为输入，基于回归树构建第一个弱回归模型，并进行训练。②根据第一个回归模型的结果，将第一个回归模型拟合偏差小的样本权重降低，将第一个模型拟合偏差大的样本权重增大，改变原始的数据分布。将改变之后的数据作为第二个模型的训练样本输入模型进行训练。③根据第二个模型回归结果继续对训练数据的分布进行调整，在调整后的数据集上构建第三个弱回归模型。重复数据分布调整以及模型构建过程，经过若干次训练后，将所有的弱回归模型按照加权平均的方式集成起来，就得到了错误率较低的强回归模型。即 Adaboost 模型。

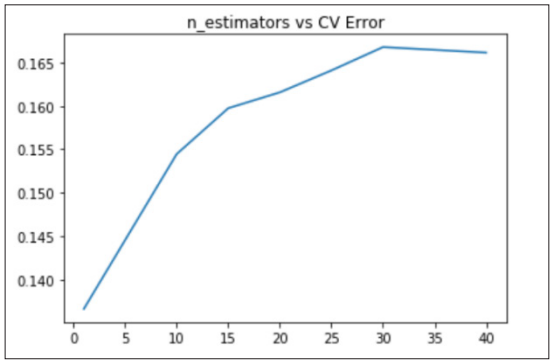

图 2–3　Adaboost 模型

2.3.3 结果分析

如图 2–3，可以看到，n_estimators 取值为 5 附近时，损失函数 score 取到最低，大概为 0.126。

2.4 构建神经网络模型 DNN

2.4.1 基本原理

神经网络是机器学习中一类很重要的学习方法，可以认为是对神经元结构从信息处理角度经过一定的抽象处理之后得出的算法，针对具体问题建立数学模型，依据连接方式以及算法的不同划分成不同的网络结构。DNN 本质上是由大量具有一定运算规则的神经元及其连接关系构建的运算模型。在神经网络中，每个神经元会对输入信号通过激活函数进行非线性变换，也就是激活函数（activation function）。在连接关系中，每两个神经元之间有连接表示对于通过的数据进行加权处理，即权重，权重参数可以看人工神经网络对于数据的记忆。人工神经网络本质上往往是对自然界存在的某种算法 / 函数的拟合，或者逻辑表达。

2.4.2 在房价预测数据集上应用 DNN 模型

构建了一个简单的三层 DNN 网络，每层 10 个神经元，采用 relu 作为激活函数，最后一层输出层设置 1 个神经元。采用了 Adam 优化器，设置了 Dropout 层。在训练过程中 400 轮的时候损失可以降低到 loss:0.12 。

2.4.3 结果分析

训练迭代到 400 轮时，损失函数 score 可以降低到 0.12。

2.5 各模型结果分析

各模型预测结果如下，Ridge Regression 模型在 Alpha=10–20 时，相应的损失函数 score 可以下降到 0.135 左右。Random Forest 模型，当 Max Features 取值为 0.5 时，能够得到损失函数最小值，在 0.137–0.138 左右。Adaboost 模型，n_estimators 取值为 5 附近时，损失函数 score 取到最低，大概为 0.126。DNN 模型，训练迭代到 400 轮时，损失函数 score 可以降低到 0.12。综合分析上述 4 个模型的结果，DNN 模型表现十分突出，增加迭代轮数 loss 还会有所下降。

三、结语

本研究以波士顿房价数据作为基础，在该数据上进行研究，分别采用了 Ridge Regression，Random Forest，Adaboost，DNN。在测试集上分别取得了 0.135，0.137，0.126，0.120 的 loss 取值。验证了选择基于 bagging 思想和基于 boosting 思想的集成学习算法后，相较之前采用的单一学习器，模型的表现在训练集和测试集上准确率都得到了明显提升，所以，集成学习方法比单一模型具有明显的优越性。DNN 算法在回归问题上也具有很明显的优越性。

参 考 文 献

[1]《机器学习》周志华
[2]《统计学习方法》李航