电子科技大学信息与软件工程学院

实 验 报 告

学　　号 201809162008

姓　　名 邓萌达

（实验）课程名称 人工智能

理论教师 周帆

实验教师 周帆

电子科技大学教务处制表

电 子 科 技 大 学

实　验　报　告

学生姓名：邓萌达　　**学号：**2018091620008　**指导教师：**周帆

实验地点：宿舍　　**实验时间：**2021-1-16

一、实验名称：回归任务—地震等级预测

二、实验学时：4

三、实验目的：

- 掌握各种回归的使用场景和使用方法
- 掌握各种回归的理论知识

四、实验原理：

- 线性回归
    - 线性回归是回归问题中的一种，线性回归假设目标值与特征之间线性相关，即满足一个多元一次方程。通过构建损失函数，来求解损失函数最小时的参数 w 和 b。通长我们可以表达成如下公式：

$$\hat{y} = wx + b$$

- 决策树回归
    - 决策树是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。在机器学习中，决策树是一个预测模型，他代表的是对象属性与对象值之间的一种映射关系。
    - 决策树是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一种类别。
- 神经网络算法

- 网络模型包括其输入输出模型、作用函数模型、误差计算模型和自学习模型
- 输入层：输入神经元定义数据挖掘模型所有的输入属性值以及概率。一个感知器可以接收多个输入（x_1,?_2 ……?_?），每个输入上有一个权值?_?，此外还有一个偏置项 b，就是上图中的?_0。
- 隐含层：隐藏神经元接受来自输入神经元的输入，并向输出神经元提供输出。隐藏层是向各种输入概率分配权重的位置。
- 输出层：输出神经元代表数据挖掘模型的可预测属性值。
- 激活函数：所谓激活函数（Activation Function），就是在人工神经网络的神经元上运行的函数，负责将神经元的输入映射到输出端，如 sigmod 函数、tanh 函数等。

- 支持向量机回归
  - 支持向量分类的方法能被推广到解决回归问题，称为支持向量回归。由支持向量分类产生的模型仅依赖训练数据的子集，因为创建模型的代价函数并不考虑超过边界的训练点。类似地，由支持向量回归产生的模型仅依赖训练数据的子集，因为创建模型的代价函数忽略任何接近模型预测的训练数据。
- 回归模型评价标准
  - MSE：该统计参数是预测数据和原始数据对应点误差的平方和的均值，公式如下：

$$MSE = \frac{SSE}{n} = \frac{1}{n}\sum i = 1^m wi(yi - \hat{yi})^2$$

  - RMSE：该统计参数，也叫回归系统的拟合标准差，是 MSE 的平方根，计算公式为：

$$RMSE = \sqrt{MSE} = \sqrt{\frac{SSE}{n}} = \sqrt{\frac{1}{n}\sum i = 1^m wi(yi - \hat{yi})^2}$$

五、实验内容：

- 回归任务—地震等级预测
- 使用均方误差（MSE）和均方根误差（RMSE）指标进行评估

六、实验器材（设备、元器件）：

- MacBook Pro（macOS 16.0）
- PyCharm（2020.03）
- Python 3.7
- Anaconda3

七、实验步骤：

- 加载数据集 quake.dat
- 划分数据集 train&test
- 使用不同的算法进行回归
    - 线性回归
    - 支持向量机回归
    - 决策树回归
    - 神经网络多层感知器回归
- 对不同算法得到的不同结果进行画图
- 使用均方误差（MSE）和均方根误差（RMSE）指标进行评估
- 分析各个算法的优点和缺点，分析得到不同的评估结果的原因

八、实验结果与分析（含重要数据结果分析或核心代码流程分析）

- 加载数据集

```
def load_data():
    data_set = pd.read_table('quake.dat', header=None, sep=',', comment='@')
    x_0_set = data_set[0].values
    x_1_set = data_set[1].values
    x_2_set = data_set[2].values
    y_set = data_set[3].values
    return x_0_set, x_1_set, x_2_set, y_set
```

- 划分数据集

```
def use_three_features(x_0_set, x_1_set, x_2_set, y_set):
    x_set = list(zip(x_0_set, x_1_set, x_2_set))
    x_set = np.array(x_set)
    return train_test_split(x_set, y_set, train_size=2000)
```

- 线性回归

```
def linear_regression(x_train, x_test, y_train, y_test):
    line = linear_model.LinearRegression()
    line.fit(x_train, y_train)
    y_predict = line.predict(x_test)
    draw_compare_line(y_test, y_predict)
    print(mse_metrics(y_test, y_predict))
    print(r_mse_metrics(y_test, y_predict))
```

- 支持向量机回归

```
def svr_regression(x_train, x_test, y_train, y_test):
    clf = SVR(kernel='rbf')
    clf.fit(x_train, y_train)
    y_predict = clf.predict(x_test)
    draw_compare_line(y_test, y_predict)
    print("mse:{}".format(mse_metrics(y_test, y_predict)))
    print("rmse:{}".format(+ r_mse_metrics(y_test, y_predict)))
```

- 决策树回归

```
def decision_tree(x_train, x_test, y_train, y_test):
    clf = tree.DecisionTreeRegressor()
    clf.fit(x_train, y_train)
    y_predict = clf.predict(x_test)
    draw_compare_line(y_test, y_predict)
    print("mse:{}".format(mse_metrics(y_test, y_predict)))
    print("rmse:{}".format(+ r_mse_metrics(y_test, y_predict)))
```

- 神经网络多层感知器回归

```
def neural_network_regression(x_train, x_test, y_train, y_test):
    mpl = MLPRegressor()
    mpl.fit(x_train, y_train)
    y_predict = mpl.predict(x_test)
    draw_compare_line(y_test, y_predict)
    print(mse_metrics(y_test, y_predict))
    print(r_mse_metrics(y_test, y_predict))
```

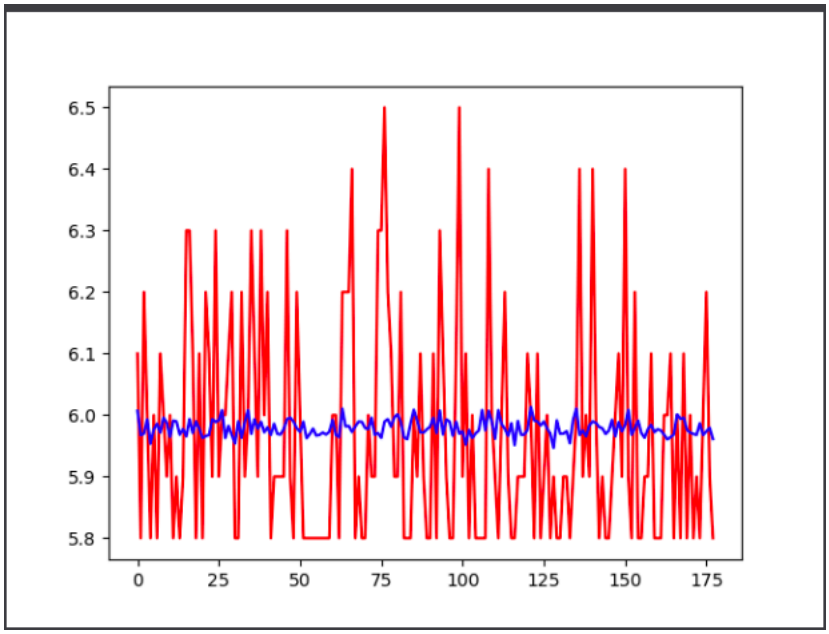
- 画出 predict&true 对比图

```
def draw_compare_line(y_test, y_predict):
    plt.plot(y_test, color='red')
    plt.plot(y_predict, color='blue')
    plt.show()
```

- 使用均方误差和均方根误差进行评估

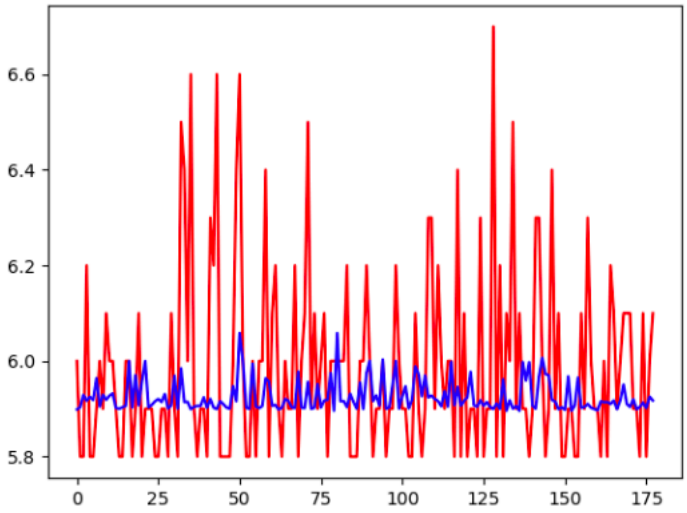
```
def mse_metrics(y_test, y_predict):
    return metrics.mean_squared_error(y_test, y_predict)


def r_mse_metrics(y_test, y_predict):
    return np.sqrt(metrics.mean_squared_error(y_test, y_predict))
```

- 结果：
    - 线性回归对比图以及评估



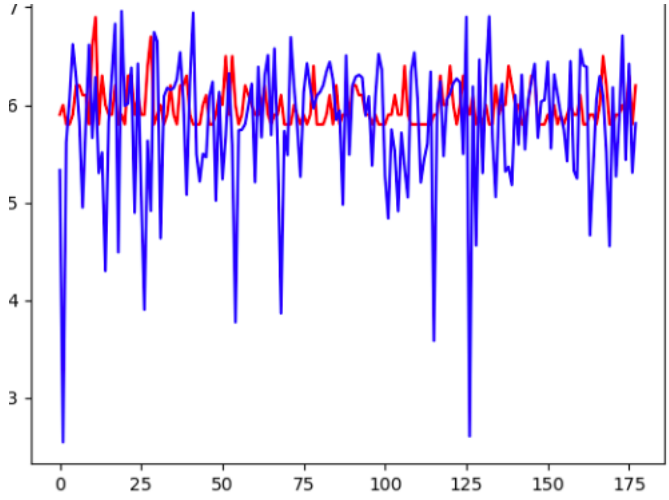

```
/usr/local/anaconda3/envs/ai-project-1/bin/python /Users/kelo/kelo/python/ai-project-1/main.py
mse:0.030655644606524184
rmse:0.1750875341265739

Process finished with exit code 0
```

- 支持向量机回归对比图以及评估



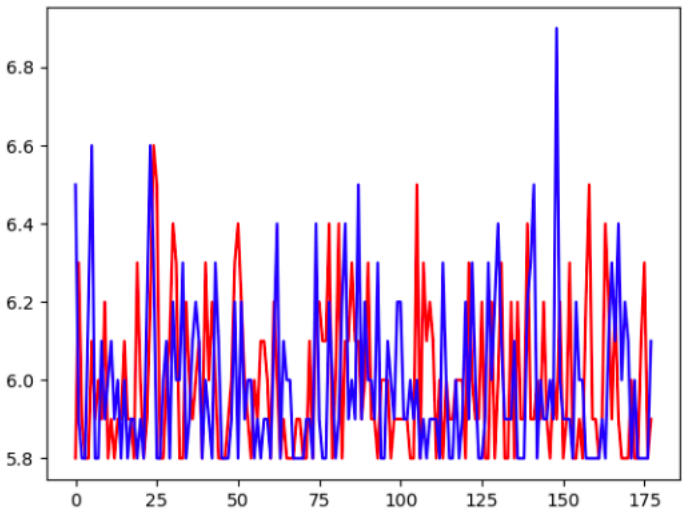

```
/usr/local/anaconda3/envs/ai-project-1/bin/python /Users/kelo/kelo/python/ai-project-1/main.py
mse:0.04135627585056525
rmse:0.2033624248738327
```

- 神经网络多层感知器回归对比图以及评估




```
mse:0.5796946098508011
rmse:0.7613767857314807
```

- 决策树回归对比图以及评估

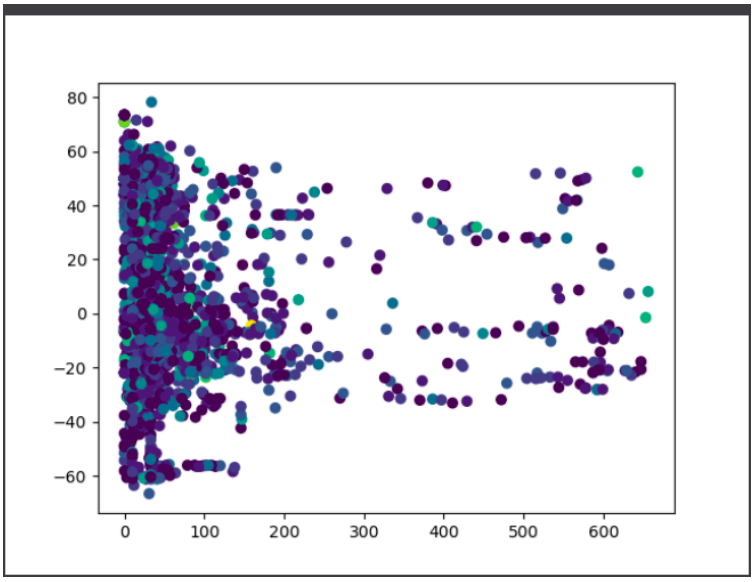


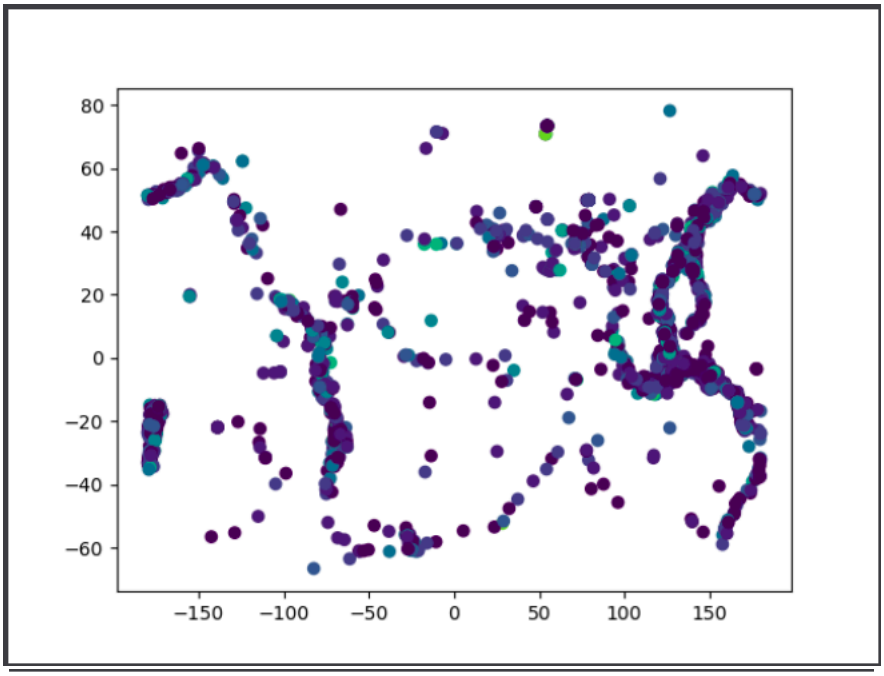

```
/usr/local/anaconda3/envs/ai-project-1/bin/python /Users/kelo/kelo/python/ai-project-1/main.py
mse:0.07719101123595505
rmse:0.2778327036832688

Process finished with exit code 0
```

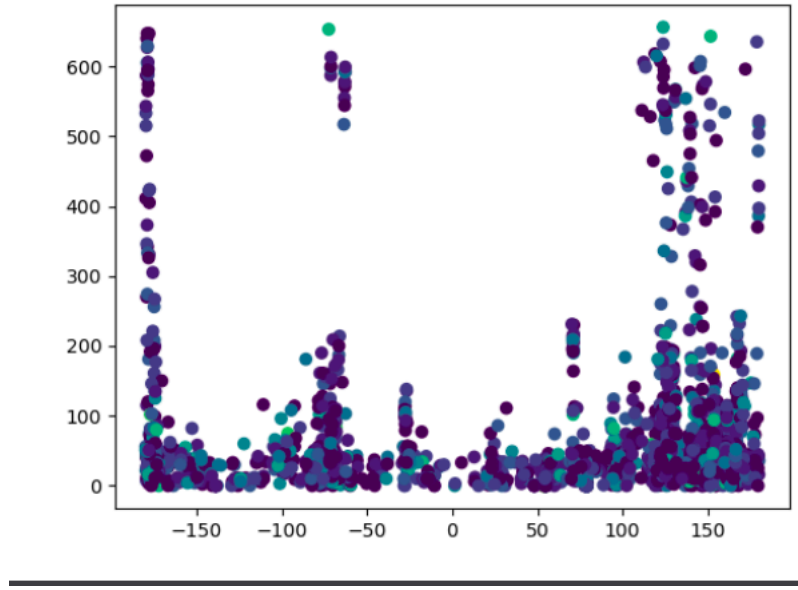
- 不同算法的比较见第九（总结）部分

九、总结及心得体会：

- 这次实验完成了对与地震的回归分析，采用了线性回归，支持向量机回归，决策树回归以及多层感知器模型的神经网络回归，完成了实验要求，达到了实验目的。
- 在结果中可以发现，神经网络的回归效果最佳，而其余算法则都处在一个非常不理想的水平。为了找到原因，我可视化了数据分布。数据分布如下：







- 以上分别是 x0 与 x1，x1 与 x2，x0 与 x2 的特征结果分布图，我们可以发现，大量不同类别（不同颜色）的点揉和在相同的区域，基本上没有任何特征状况，更不用说有什么线性特征，所以当我们使用线性分类，例如线性回归的时候，效果就非常不理想。而使用学习得到的神经网络多层感知器模型的时候，才会有相比于其他算法较高的指标数据。
- 线性分类算法，例如线性回归，建模速度快，不需要很复杂的计算，在数据量大的情况下依然运行速度很快。它也可以根据系数给出每个变量的理解和解释。但是，最关键的一点就是，他在拟合这种非线性数据的时候表现特别

差，所以说我们在使用线性回归的时候要判断数据是否有大致的线性关系，因此，在运用于这个数据上的时候，表现很不理想。

- 支持向量机 SVM，则在解决高维特征的分类问题和回归问题很有效,在特征维度大于样本数时依然有很好的效果。而且，它仅仅使用一部分支持向量来做超平面的决策，无需依赖全部数据。因此，可以在数据量较小的情况下得到较好的结果。最重要的是，相比于上述的线性回归它有大量的核函数可以使用，从而可以很灵活的来解决各种非线性的分类回归问题，这也导致了在训练时，它的效果会好于线性回归，但是，如果特征维度远远大于样本数，则 SVM 表现一般，最重要的是，非线性问题的核函数的选择没有通用标准，难以选择一个合适的核函数，所以我们所选的 rbf kernel 在这次训练中，也没有得到很好的表现，导致效果不理想。
- 决策树相比于 SVM，易于理解和解释，可以可视化分析，容易提取出规则。此外，在运行速度上，它在相对短的时间内能够对大型数据源做出可行且效果良好的结果。也能够处理不相关的特征。然而，相比于其他回归算法，它容易发生过拟合，导致回归结果不理想。同时，容易忽略数据集中属性的相互关联，这也是此算法在本次回归时，造成结果不理想的主要原因之一：单个特征完全不能找到结果的规律，要将各个特征联系起来，才能达到较好的拟合效果。
- 神经网络多层感知器算法相比于其他算法，脱颖而出，这主要是因为，它有非常强的自适应、自学习功能并且具有联想记忆功能，同时也有良好的容错性。此外，在处理非线性数据时，高度的非线性全局作用让他相比于其他算法有了很大的优势。然而，这次学习中，神经网络算法也没有得到非常满意的结果，主要是因为，网络的隐含节点个数选取非常难，同时，他还会容易陷入局部极值，还有更重要的一点是，本次数据样本较小，会有学习不充分的问题。另一方面，该算法停止阈值、学习率、动量常数需要采用”trial-and-error”法，极其耗时，相比于其他算法学习速度慢，可以说是牺牲了时间来保证效果。

十、对本实验过程及方法、手段的改进建议：

- 这次实验选择的数据选择较为杂乱，对于初学者来说不能很好的分析出数据

的特征以及选择合适的回归算法，建议在今后的实验中，给对数据的分析做出指导，推荐不同的有针对性的算法进行比较，突出不同算法在处理不同数据方面的特性，给学生更加直观的感受。

报告评分：

指导教师签字：