基于模型预测材料的形成能

杨醒乾 231017000174（23 春大数据）

2023 年 12 月 3 日

1，应用背景：

对材料的研究,通过大量实验观察数据,凭借直觉提出假设然后验证假设的传统材料研究方法,已经很明显不适用于当今高速发展的数字智能时代。利用第一性原理,能够不断完善热电理论和进行新型热电材料设计,但是研发周期依旧过长。因此,探寻新的研究方法来辅助加快新型材料的研发具有重要意义。一种基于机器学习进行材料性能的预测方法,重点是通过机器学习方法对材料的形成能进行预测,帮助后续材料研究人员快速筛选出具有理想材料。

2. 思路：

构建数据集特征：材料分子中原子数，原子周期表序列，材料化学式等。预测目标为材料各原子的形成能；

数据集划分，将数据集划分为训练集与测试集比例为 9:1；

使用 Python 的 Slearn 库自带的机器学习算法训练模型，并对材料的形成能进行预测。评估指标：

均方误差（Mean Squared Error，MSE）：回归模型中，衡量模型预测结果与实际结果之间的距离。

平均绝对误差（Mean Absolute Error，MAE）：回归模型中，衡量模型预测结果与实际结果之间的距离，其值不受离群值的影响。

R2（R-Squared）：回归模型中常见的评估指标，反映模型拟合数据的好坏，其值越接近表示模型的解释能力越强。

3. 使用模型：

Kernel Ridge Regression (KRR)：基于核岭回归的非线性回归方法。

Support Vector Regression (SVR) ：基于 Support Vector Machines (SVM) 的回归方法

Gradient boosting regression：基于决策树的集成学习算法，适用于回归问题。

4. 模型预测关键代码：

GBR 关键代码：

```
gbr = GridSearchCV(GradientBoostingRegressor(),{
    'n_estimators': [2000], 'max_depth': [2], 'min_samples_split': [2], 'learning_rate': [0.1],
    'loss': ['ls'], 'random_state':[72]}, cv=5)

X_train1.drop(columns=['formula'],axis=1,inplace=True)
X_test1.drop(columns=['formula'],axis=1,inplace=True)

gbr.fit(X_train1, y_train1)
y_predicted1 = gbr.predict(X_test1)
gbr_score = gbr.score(X_train1,y_train1)
gbr_score1 = gbr.score(X_test1,y_test1)
plot = plot_pred_act(y_test1, y_predicted1, 'GBR Model', reg_line=True, label='$ (eV/atom)')
```

KRR 关键代码：

```
krr = GridSearchCV(KernelRidge(),{'alpha':[0.001],'kernel':['linear']}, cv=5)
X_train2.drop(columns=['formula'],axis=1,inplace=True)
X_test2.drop(columns=['formula'],axis=1,inplace=True)
krr.fit(X_train2, y_train2)
y_predicted2 = krr.predict(X_test2)
krr_score = krr.score(X_train2,y_train2)
krr_score1 = krr.score(X_test2,y_test2)
plot = plot_pred_act(y_test2, y_predicted2, 'KRR Model', reg_line=True, label='$ (eV/atom)')
```

SVM 关键代码：

```
steps = [('scaler', StandardScaler()), ('SVM', SVR())]
pipeline = Pipeline(steps)
grid = GridSearchCV(pipeline, param_grid={'SVM__C':[100], 'SVM__gamma':['auto'], 'SVM__kernel': ['rbf'],
                                          'SVM__epsilon':[0.001]}, cv=5)

X_train3.drop(columns=['formula'],axis=1,inplace=True)
X_test3.drop(columns=['formula'],axis=1,inplace=True)
grid.fit(X_train3, y_train3)
svr_score = grid.score(X_train3,y_train3)
svr_score1 = grid.score(X_test3,y_test3)
y_predicted3 = grid.predict(X_test3)

plot = plot_pred_act(y_test3, y_predicted3, 'SVR Model', reg_line=True, label='$ (eV/atom)')
```

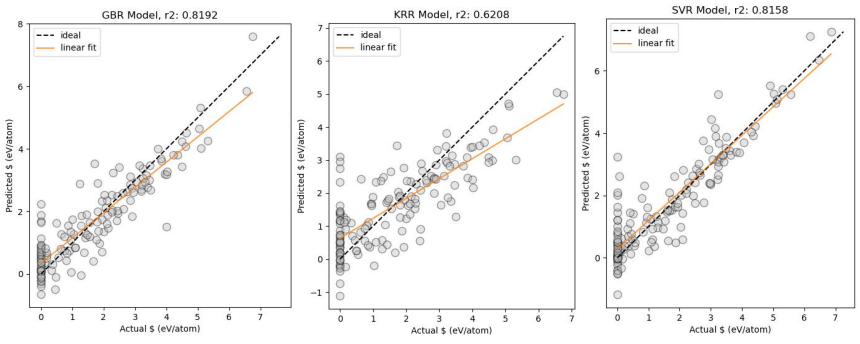
5. 结果：

如图一所示，GBR 与 SVR 的预测效果相较于 KRR 更好，其测试集的 R2 值分别为 0.8182 与 0.8158，这证明 GBR 与 SVR 可以较好的预测材料的形成能。这也表明所建立的特征集可以较好的表征出材料的性能。

```
GBR Model| R2 sq on train set: 0.9928
GBR Model| R2 sq on test set: 0.8192
GBR Model| MSE on test set: 0.4676
GBR Model| MAE on test set: 0.5191
---------------------------------
KRR Model| R2 sq on train set: 0.5659
KRR Model| R2 sq on test set: 0.6208
KRR Model| MSE on test set: 0.9810
KRR Model| MAE on test set: 0.7731
---------------------------------
SVR Model| R2 sq on train set: 0.9288
SVR Model| R2 sq on test set: 0.8158
SVR Model| MSE on test set: 0.5139
SVR Model| MAE on test set: 0.4927
```

图一：不同模型预测材料形成能性能评估

为了更直观的表现出各个模型对材料形成能的预测性能，一下通过图片的方式来显示模型估算值与实际值的偏差。如图二所示，其中黑色虚线为不同材料形成能实际值的线性趋势，而黄色实线为预测值的线性趋势。虚线与实现的夹角越大证明预测性能越差。KRR 的预测效果偏差较大。



图二：不同模型预测结果

总结：

1）由于个人能力有限，模型的预测效果暂不能达到最佳

2）使用 Sklearn 库模型，并未对其进行精确调整。

3）通过该学习以充分理解数据挖掘原理与实现方式。