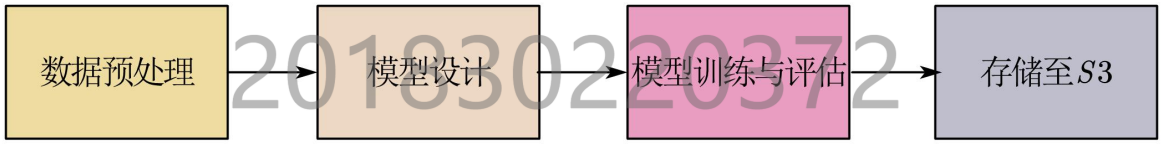
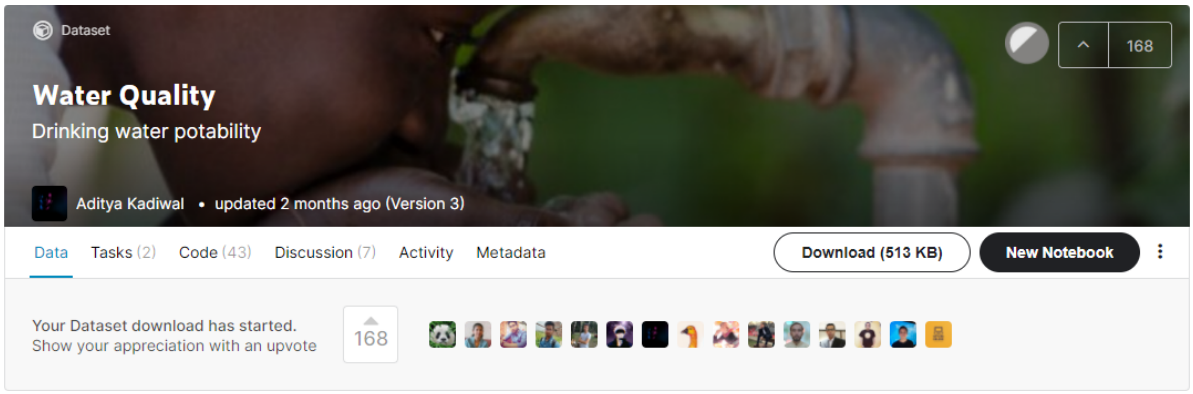
设计说明书

流程图



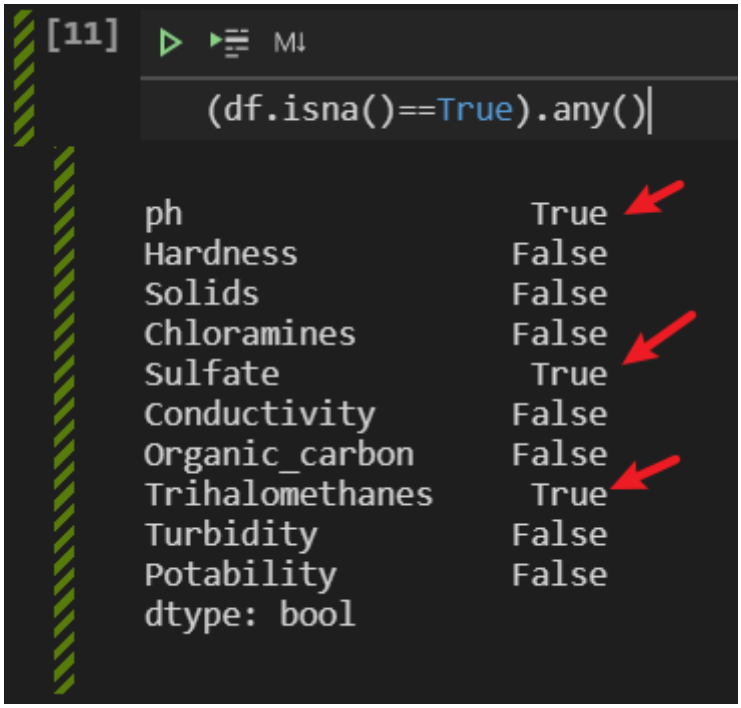
数据集与预处理

在本次实验中，我们选择了一份二分类的数据集，任务是根据水源的八种特征，来预测水源是否适合人类引用。数据集可以在Water Quality | Kaggle下载得到。



在初步的数据可视化中，我们发现了数据存在一些空值，我们采用了平均填充的方法来修复。

```
(df.isna()==True).any()
# df = df.fillna(df.mean())
```

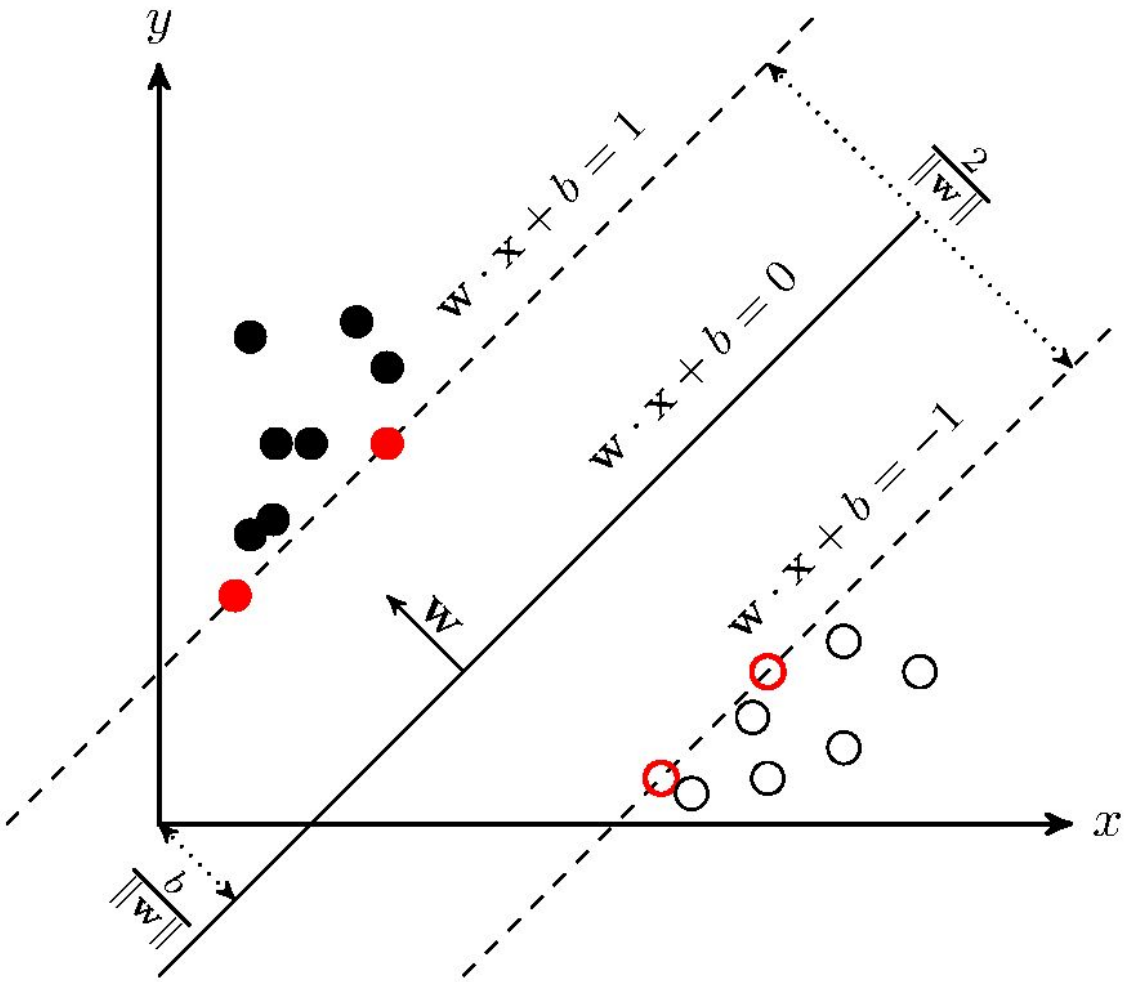


在清洗完数据后，我们将数据进行`3:1`的比例进行训练集和测试集的划分。

- 数据集划分：

```
Y = df['Potability']
X = df.drop(columns=['Potability'], axis=1)
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.3,
random_state=33)
```

模型设计



在本次分类任务中，我们采用SVM模型，该模型的核心思想是在特征空间学习一个最优超平面将样本划分开，学习的目标可以形式化以下的凸二次优化问题，该式可以结合拉格朗日乘子法和KKT条件进行求解。在实验中，我们采用更加方便的`sklearn`库，该库已经将拟合的过程进行了高度的封装，使得我们能够很快地完成模型的训练和预测。

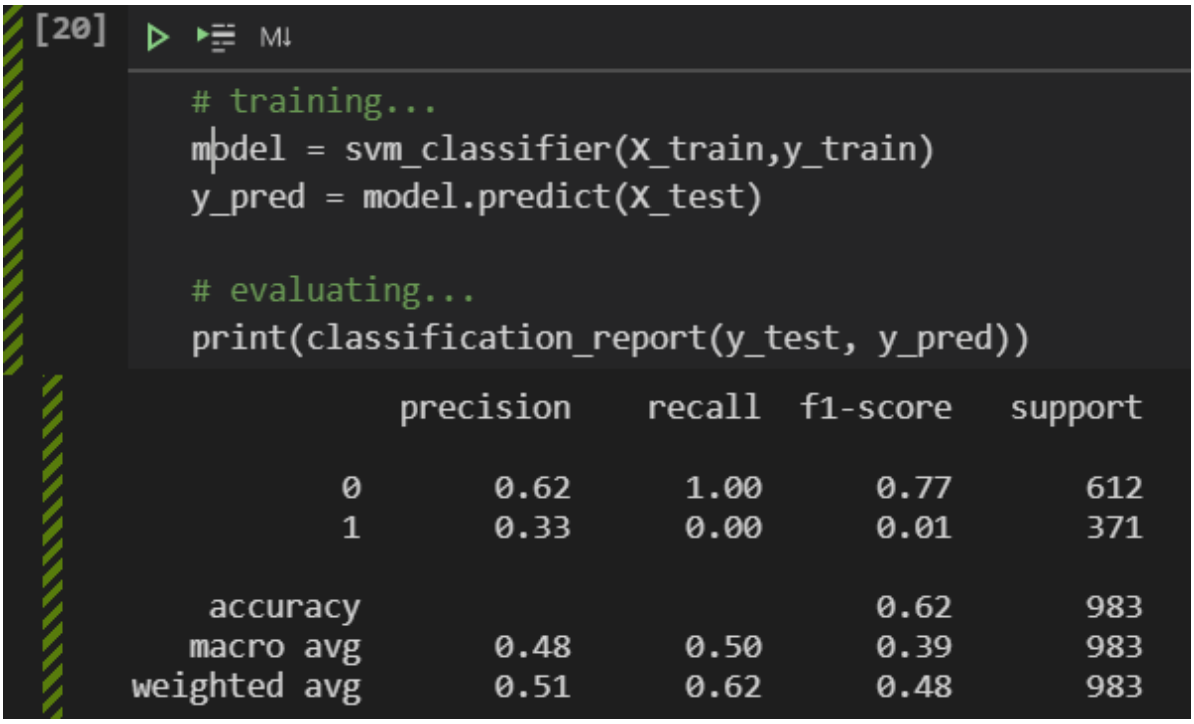
$$\min_{w,b} \frac{1}{2}||w||^2$$
$$s.t.\, y_i(w^T x_i + b) \geqslant 1, i = 1, 2, \ldots, m.$$

- 模型代码：

```python
# SVM Classifier
def svm_classifier(train_x, train_y):
    from sklearn.svm import SVC
    model = SVC(kernel='rbf', probability=True)
    model.fit(train_x, train_y)
    return model
```

模型训练与评估

在构建完模型和划分好数据后，我们直接调用模型即可完成参数拟合。借助`sklearn.metrics`的`classification_report`函数，我们可以从`precision`,`recall`,`f1-score`,`support`四个方面对模型拟合效果进行评估。结果如下

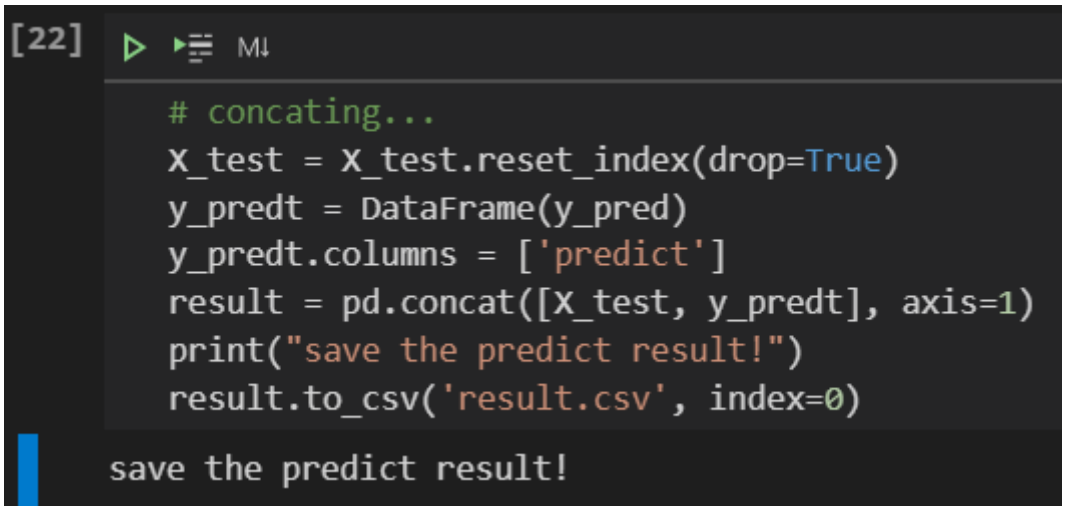


```
[20]
# training...
model = svm_classifier(X_train,y_train)
y_pred = model.predict(X_test)

# evaluating...
print(classification_report(y_test, y_pred))
```

	precision	recall	f1-score	support
0	0.62	1.00	0.77	612
1	0.33	0.00	0.01	371
accuracy			0.62	983
macro avg	0.48	0.50	0.39	983
weighted avg	0.51	0.62	0.48	983

进行完预测后，我们将预测结果保存下来



```
[22]
# concating...
X_test = X_test.reset_index(drop=True)
y_predt = DataFrame(y_pred)
y_predt.columns = ['predict']
result = pd.concat([X_test, y_predt], axis=1)
print("save the predict result!")
result.to_csv('result.csv', index=0)

save the predict result!
```

存储至S3

在最后，我们调用S3的接口，将结果存储到上面。

```python
# uploading...
aws_bucket_name = 'liuyixin'
s3 = boto3.client('s3',
                  aws_access_key_id='E441571B4C6B777EB1F8',

 aws_secret_access_key='WzY1QTg4N0JBOTBGQUJFOUNDQTZERTlFMDRBMkI1',
                  endpoint_url='http://scut.depts.bingosoft.net:29997')
s3.upload_file('result.csv', 'liuyixin', 'result.csv')
print('upload succeed!')
```