

华东师范大学数据学院人工智能机器学习报告

课程名称： 人工智能	年级： 2017
指导教师： 罗轶凤	姓名： 熊双宇

一、实验目的

1. 回归 (Air quality dataset) : 逻辑回归
2. 分类 (BLE&RSSI dataset) : SVM、决策树、随机森林
3. 聚类 (BLE&RSSI dataset) : DBScan、kmeans、GMM、层次聚类算法

二、实验思路

1. 回归和分类中的算法 (监督学习)
 - 划分训练集 (training set) 和测试集 (test set) ;
 - 训练集用于训练模型;
 - 测试集用于评估模型。
2. 聚类中的算法为 (无监督学习)
 - 所有数据用于训练模型;
 - sklearn.metrics.silhouette_score用于评估模型。

三、实验过程

(一)回归

1. 数据预处理
 - drop(): 除去feature为-200 (无效值) 的samples;
 - fit_transform(): 先拟合数据, 再标准化数据, 保证每个维度的特征数据方差为1, 均值为0, 防止某些维度过大的特征值主导预测结果;
 - train_test_split(): 将数据随机划分为训练集和测试集:
 - 训练集X_train包含: Time PT08.S1(CO) NMHC(GT) C6H6(GT) PT08.S2(NMHC) NOx(GT) PT08.S3(NOx) NO2(GT) PT08.S4(NO2) PT08.S5(O3) T RH AH 的samples
 - y_train包含: CO(GT)的samples
 - 测试集X_test同训练集X_train, y_test同y_train
2. 实现回归算法
 - rgs=LinearRegression(): 使用scikit-learn的线性回归算法 (最小二乘法实现该算法) ;
 - rgs.fit(X_train, y_train): 用训练集训练模型;
 - rgs.predict(X_test): 预测测试集labels;

3. 模型评估

- `cross_val_score()`: 交叉验证模型准确度 (cv=10)

```
array([0.95865698, 0.97753118, 0.97882282, 0.95293484, 0.9774073 ,
       0.97068455, 0.9543947 , 0.94802814, 0.95844381, 0.96888202])
Accuracy: 0.96 (+/- 0.02)
```

(二)分类

1. 数据预处理

- `LabelEncoder()`: 分别将date和location由string对应成number;

2. 实现分类算法

- `train_test_split()`: 将数据随机划分为训练集和测试集;
 - 训练集X_train包含: date b3001 b3002 b3003 b3004 b3005 b3006 b3007 b3008 b3009 b3010 b3011 b3012 b3013的samples
y_train包含: location 的samples
 - 测试集X_test同训练集X_train, y_test同y_train
- 用训练集数据训练模型, 三种算法:
 - `svm.SVC().fit()`;
 - `tree.DecisionTreeClassifier().fit()`
 - `RandomForestClassifier().fit()`;
- `predict()`: 预测labels
- `inverse_transform()`: 将分类结果由number转化为对应的string (即location)

3. 评估模型

- `cross_val_score()`: 交叉验证得出准确度, 三种算法效果分别为: (cv=5)

```
■ array([0.34055728, 0.46278317, 0.42657343, 0.56226415, 0.5021097 ])
Accuracy: 0.46 (+/- 0.15)
```

```
■ array([0.43343653, 0.47572816, 0.43706294, 0.5245283 , 0.38818565])
Accuracy: 0.45 (+/- 0.09)
```

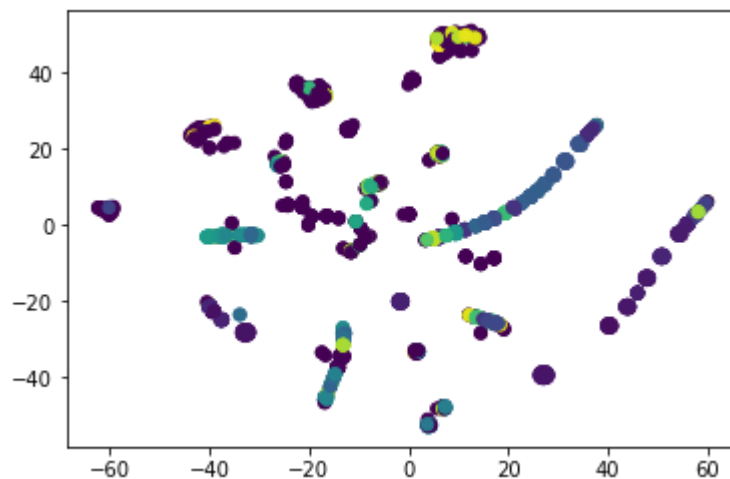
```
■ array([0.38080495, 0.33980583, 0.33916084, 0.39245283, 0.35443038])
Accuracy: 0.36 (+/- 0.04)
```

(三)聚类

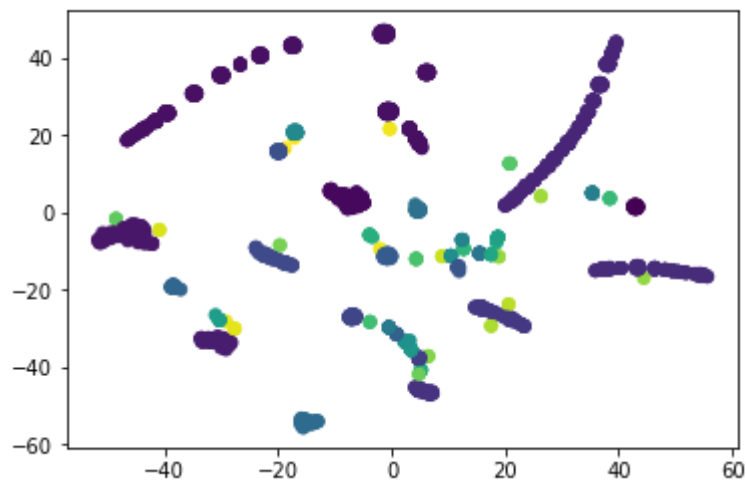
1. 实现聚类算法:

- 选择features: date b3001 b3002 b3003 b3004 b3005 b3006 b3007 b3008 b3009 b3010 b3011 b3012 b3013 对应的samples
- 四种算法进行聚类:
 - `DBSCAN().fit()`
 - `KMeans().fit()`

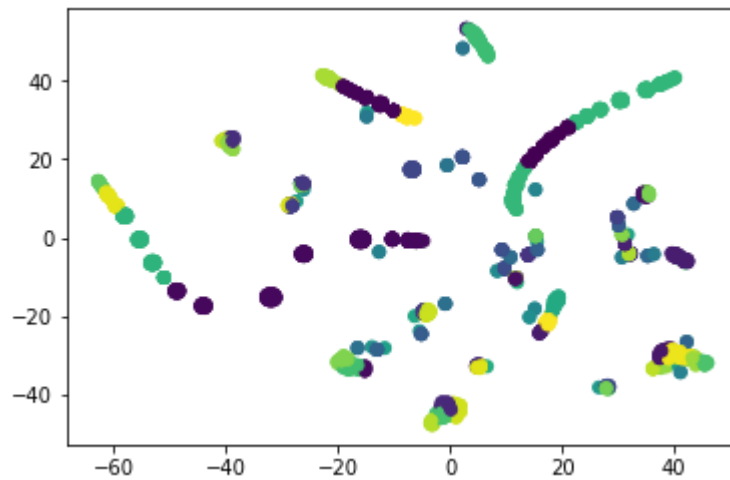
- GaussianMixture().fit()
- AgglomerativeClustering().fit()
- 对于每一种聚类算法的结果, 进行评估: `metrics.silhouette_score()`, 四种算法聚类效果分别为:
 - DBScan: 0.4128485921118394
 - KMeans: 0.9275119837253877
 - GMM: 0.8719717201085537
 - 层次聚类法: 0.2729861875905149
- 对于每一种聚类算法的结果, 使用tSNE进行将13维降低至2维: `TSNE(n_components=2).fit_transform()`
- 可视化:
 - DBScan:



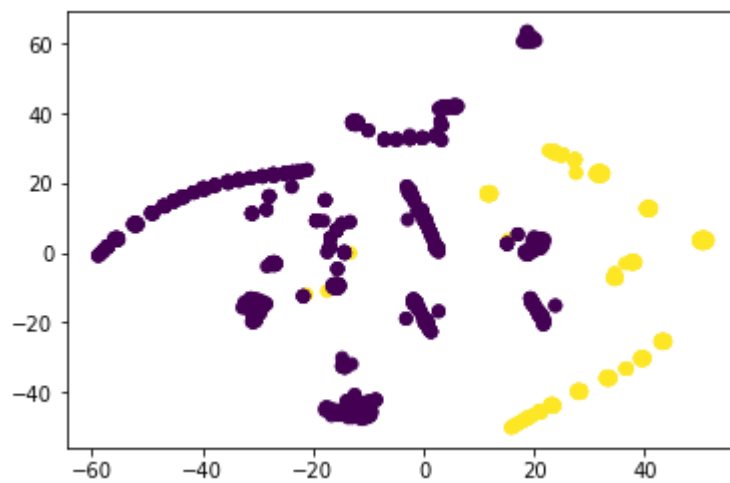
- KMeans:



- GMM:



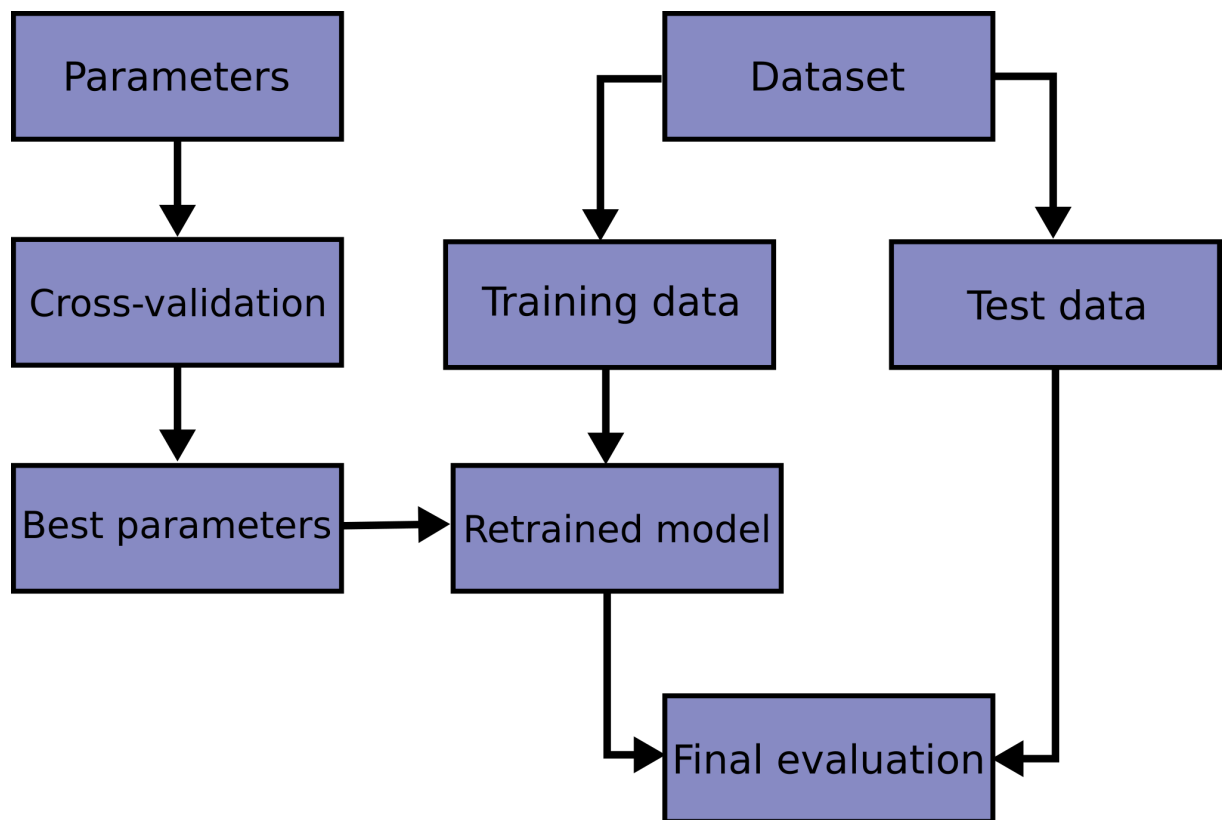
■ 层次聚类法：



四、实验反思

1. 交叉验证评估模型，防止过拟合：

○ 流程图：



○ 原理：

在给定的建模样本中，拿出大部分样本进行建模型，留小部分样本用刚建立的模型进行预报，并求这小部分样本的预报误差，记录它们的平方加和。

交叉验证的一个round包含：

- partitioning a sample of data into complementary subsets;
- performing the analysis on one subset (called the *training set*);
- validating the analysis on the other subset (called the *validation set* or *testing set*)

交叉验证往往是在multiple rounds中使用不同的partitions，并且组合validation的结果（比如用平均数）给出模型预测的效果。

○ 判断过拟合：

▪ 过拟合概念：

- Overfitting is a modeling error which occurs when a function is too closely fit to a limited set of data points.
- The model has extracted some of the residual variation (i.e. the noise) ;
- The model contains more parameters than can be justified by the data;
- The model shows **high variance and low bias**.

▪ 欠拟合概念：

- The model can't adequately capture the underlying structure of the data.
- Underfitting occurs if the model or algorithm shows **low variance but high bias**;

2. 轮廓系数对无监督学习进行评估: 结合内聚度和分离度