《机器学习与模式识别》上机实验指导书

实验3：模型评估与选择（2）

**实验目的：**

1. 熟悉函数说明文档的检索，熟悉主要函数的功能与参数含义。
2. 具备合理修改算法参数和基本的改写或编写简单代码的能力。
3. 基本掌握带交叉验证的网格搜索函数的使用。
4. 基本掌握带交叉验证的随机搜索函数的使用。
5. 基本掌握混淆矩阵、ROC曲线的画法，ROC曲线面积的计算。
6. 进一步熟悉Python编程与绘图方法。

**实验内容：**

学习相关材料，熟悉主要工具函数的使用。使用 Anaconda JupyterLab，运行、调试实验用基础代码，通过修改适当参数，增加必要代码，完成以下实验。

（1）熟悉带交叉验证的最优超参数搜索函数的使用。

简要说明**带交叉验证的网格搜索函数**（GridSearchCV）主要参数的含义、取值类型及其缺省值，函数返回值的意义与数据类型。结合代码参数修改，比如：修改cv值，修改scoring（在'accuracy', 'recall', 'precision', ‘f1’等之间选择）等，观察最优模型的参数，以及最优模型的训练性能（具体指标有scoring指定）。对实验结果进行简要说明。

简要说明**带交叉验证的随机搜索函数**（RandomizedSearchCV）主要参数的含义、取值类型及其缺省值，函数返回值的意义与数据类型。比如：修改cv值，修改逻辑回归的逆正则参数C的取值范围（调整均匀随机数uniform的上限值scale），修改scoring（说明为什么不能使用 'recall', 'precision', ‘f1’，而要改为'recall\_micro', 'precision\_micro', ‘f1\_micro’，但是'accuracy'不用加\_micro）。对实验结果进行简要说明。

class sklearn.model\_selection.GridSearchCV(estimator, param\_grid, \*, scoring=None, n\_jobs=None, iid='deprecated', refit=True, cv=None, verbose=0, pre\_dispatch='2\*n\_jobs', error\_score=nan, return\_train\_score=False)

class sklearn.model\_selection.RandomizedSearchCV(estimator, param\_distributions, \*, n\_iter=10, scoring=None, n\_jobs=None, iid='deprecated', refit=True, cv=None, verbose=0, pre\_dispatch='2\*n\_jobs', random\_state=None, error\_score=nan, return\_train\_score=False)

（参数表见手册或在线文档）

（2）熟悉嵌套交叉验证的最优参数搜索方法。

结合代码，熟悉使用嵌套交叉验证的足有超参数搜索方法的函数调用序列：内部交叉验证(使用GridSearchCV或RandomizedSearchCV)和外部交叉验证（使用cross\_val\_score，注意cross\_val\_score缺省cv=3）。在基础代码上，适当修改内部、外部cv折数，观察、记录并简要分析实验结果。对嵌套交叉验证过程进行简要说明。

（3）熟悉其他性能评估指标的使用。

结合代码，掌握**分类器混淆矩阵**的画法、**分类器ROC曲线**的画法、**分类器ROC曲线面积**的计算方法。

sklearn.metrics.confusion\_matrix(y\_true, y\_pred, \*, labels=None, sample\_weight=None, normalize=None)

sklearn.metrics.roc\_curve(y\_true, y\_score, \*, pos\_label=None, sample\_weight=None, drop\_intermediate=True)

sklearn.metrics.roc\_auc\_score(y\_true, y\_score, \*, average='macro', sample\_weight=None, max\_fpr=None, multi\_class='raise', labels=None)[source]

**数据说明：**

wdbc.datda已经下载。

**参考资料：**

**VanderPlas，第5.3节。**

**scikit-learn user Guide, Release 0.23.1（May 19, 2020）.pdf**

**带交叉验证的网格搜索：sklearn.model\_selection.GridSearchCV**

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html?highlight=gridsearchcv#sklearn.model_selection.GridSearchCV>

**带交叉验证的网格搜索：sklearn.model\_selection.RandomizedSearchCV**

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html#sklearn.model_selection.RandomizedSearchCV>

**混淆矩阵：sklearn.metrics.confusion\_matrix**

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html#sklearn.metrics.confusion_matrix>

**ROC曲线：sklearn.metrics.roc\_curve**

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_curve.html?highlight=roc_curve#sklearn.metrics.roc_curve>

**ROC曲线面积：sklearn.metrics.roc\_auc\_score**

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.roc_auc_score.html#sklearn.metrics.roc_auc_score>

**多类分类器的性能指标为什么不能直接用f1, recall, precision等：**

<https://scikit-learn.org/stable/modules/model_evaluation.html#scoring-parameter>

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html#sklearn.metrics.f1_score>

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html#sklearn.metrics.precision_score>

**分层K折：sklearn.model\_selection.StratifiedKFold**

<https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html?highlight=stratifiedkfold#sklearn.model_selection.StratifiedKFold>