集成学习

本节代码参考: **黄海广-机器学习** https://github.com/fengdu78/WZU-machine-learning-course 推荐自学

1 集成模型比对

1.1 生成数据

生成Hastie等人使用的二进制分类数据。

十个特征是标准独立的高斯。

In [1]:
```
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
from sklearn.model_selection import train_test_split
```

In [ ]:
```
!python -m pip install lightgbm xgboost
```

In [5]:
```
from sklearn.datasets import make_hastie_10_2

data, target = make_hastie_10_2()
X_train, X_test, y_train, y_test = train_test_split(data, target, random_state=2023)
```

In [12]:
```
X_train.shape, X_test.shape
```

Out[12]:
```
((9000, 10), (3000, 10))
```

1.2 模型对比

对比六大模型，都使用默认参数。

使用XGB训练中，出现处错误：Invalid classes inferred from unique values of `y`. Expected: [0 1], got ['0.0' '1.0']

参考博客。

In [10]:

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
y_train = le.fit_transform(y_train)
```

In [11]:

```
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import AdaBoostClassifier
from sklearn.ensemble import GradientBoostingClassifier
from xgboost import XGBClassifier
from lightgbm import LGBMClassifier
from sklearn.model_selection import cross_val_score
import time

clf1 = LogisticRegression()
clf2 = RandomForestClassifier()
clf3 = AdaBoostClassifier()
clf4 = GradientBoostingClassifier()
clf5 = XGBClassifier()
clf6 = LGBMClassifier()

for clf, label in zip([clf1, clf2, clf3, clf4, clf5, clf6], [
        'Logistic Regression', 'Random Forest', 'AdaBoost', 'GBDT', 'XGBoost',
        'LightGBM'
]):
    start = time.time()
    scores = cross_val_score(clf, X_train, y_train, scoring='accuracy', cv=5)
    end = time.time()
    running_time = end - start
    print("Accuracy: %0.8f (+/- %0.2f),耗时%0.2f秒。模型名称[%s]" %
          (scores.mean(), scores.std(), running_time, label))
```

```
Accuracy: 0.47177778 (+/- 0.01),耗时0.04秒。模型名称[Logistic Regression]
Accuracy: 0.88800000 (+/- 0.01),耗时17.34秒。模型名称[Random Forest]
Accuracy: 0.87077778 (+/- 0.00),耗时3.21秒。模型名称[AdaBoost]
Accuracy: 0.91144444 (+/- 0.00),耗时14.74秒。模型名称[GBDT]
Accuracy: 0.92455556 (+/- 0.00),耗时2.13秒。模型名称[XGBoost]
Accuracy: 0.92900000 (+/- 0.00),耗时0.63秒。模型名称[LightGBM]
```

2 集成学习-鸢尾花数据集分类

In [17]:

```
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier

iris = load_iris()  # 加载鸢尾花数据集
X = iris.data  # 样本特征
y = iris.target  # 样本标签
X = X[:,:2]  # 选择前两个特征
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

In [18]:

```
clf1 = LogisticRegression()
clf2 = RandomForestClassifier()
clf3 = AdaBoostClassifier()
clf4 = GradientBoostingClassifier()
clf5 = XGBClassifier()
clf6 = LGBMClassifier()

for clf, label in zip([clf1, clf2, clf3, clf4, clf5, clf6], [
        'Logistic Regression', 'Random Forest', 'AdaBoost', 'GBDT', 'XGBoost',
        'LightGBM'
]):
    start = time.time()
    scores = cross_val_score(clf, X_train, y_train, scoring='accuracy', cv=5)
    end = time.time()
    running_time = end - start
    print("Accuracy: %0.8f (+/- %0.2f),耗时%0.2f秒。模型名称[%s]" %
          (scores.mean(), scores.std(), running_time, label))
```

```
Accuracy: 0.78095238 (+/- 0.07),耗时0.04秒。模型名称[Logistic Regression]
Accuracy: 0.72380952 (+/- 0.04),耗时0.51秒。模型名称[Random Forest]
Accuracy: 0.44761905 (+/- 0.05),耗时0.30秒。模型名称[AdaBoost]
Accuracy: 0.69523810 (+/- 0.05),耗时0.78秒。模型名称[GBDT]
Accuracy: 0.68571429 (+/- 0.04),耗时0.36秒。模型名称[XGBoost]
Accuracy: 0.74285714 (+/- 0.04),耗时0.22秒。模型名称[LightGBM]
```