《大数据分析方法》课程实验报告

学号：2017202105

姓名：林子豪

专业：计算机科学与技术

班级：3

# 实验二：分类回归算法的实验基础

## 一 实验目的

在学习线性回归，逻辑回归，knn，决策树，贝叶斯分类等算法的原理的基础上，掌握已有的数据分类算法软件包中数据分类回归算法的使用

## 二 实验内容

学习scikit-learn包中的Classification

## 三 实验要求

至少实验2种算法的使用

## 四 实验过程

1.导入需要的包

from sklearn.datasets import load\_iris

import numpy as np

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn import linear\_model

from sklearn.tree import DecisionTreeClassifier

import math

from sklearn.naive\_bayes import MultinomialNB

%matplotlib inline

2.定义acc函数，求预测准确率

def acc(predic,real):

ans=0

for i in range(len(predic)):

if predic[i]==real[i]:

ans+=1

return ans/len(predic)

3.从数据集中读取数据

iris\_dataset = load\_iris()

X\_train,X\_test,y\_train,y\_test = train\_test\_split(iris\_dataset['data'],iris\_dataset['target'],random\_state=0)

iris\_dataframe = pd.DataFrame(X\_train,columns=iris\_dataset.feature\_names)

iris\_dataframe.head()

1. 线性模型

model=linear\_model.LinearRegression()

model.fit(X\_train,y\_train)

ans=model.predict(X\_test)

for index,a in enumerate(ans):

k=np.zeros(3)

for i in range(3):

k[i]=math.fabs(i-a)

ans[index]=np.argmin(k)

print(acc(ans.astype(np.int32),y\_test))

准确率0.9736842105263158

5.逻辑回归

model=linear\_model.LogisticRegression()

model.fit(X\_train,y\_train)

ans=model.predict(X\_test)

print(acc(ans.astype(np.int32),y\_test))

准确率0.868421052631579

1. 决策树

model=DecisionTreeClassifier()

model.fit(X\_train,y\_train)

ans=model.predict(X\_test)

print(acc(ans.astype(np.int32),y\_test))

准确率0.9736842105263158

1. 朴素贝叶斯

model=MultinomialNB()

model.fit(X\_train,y\_train)

ans=model.predict(X\_test)

print(acc(ans.astype(np.int32),y\_test))

准确率0.5789473684210527